Outsourcing the news? An empirical assessment of the role of sources and news agencies in the contemporary news landscape

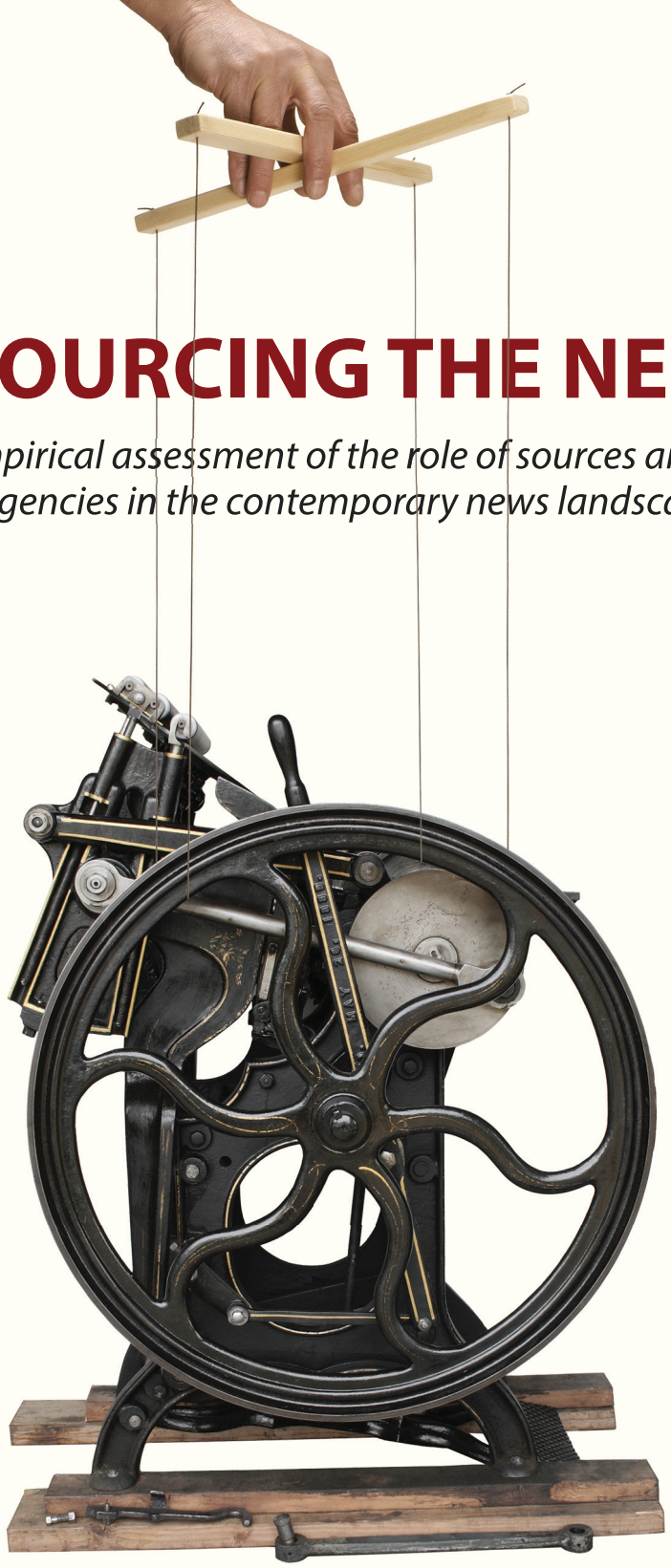
Boumans, J.W.

Publication date
2016

Document Version
Final published version

OUTSOURCING THE NEWS?

An empirical assessment of the role of sources and news agencies in the contemporary news landscape



Jelle W. Boumans

OUTSOURCING THE NEWS?

An empirical assessment of the role of sources and news agencies in the contemporary news landscape

Jelle W. Boumans

University of Amsterdam

Outsourcing the news?
An empirical assessment of the role of sources and news agencies in the contemporary news landscape

ISBN: 978-94-6328-043-3

The research presented in this dissertation was conducted at the Amsterdam School of Communication Research (ASCoR). The project was funded by the Netherlands Organisation for Scientific Research (NWO Graduate Program).

Layout by Martijn van der Meer
Cover design by Jelle Boumans and Martijn van der Meer
Printed by CPI – Koninklijke Wöhrmann

Amsterdam School of Communication Research (ASCoR)
Department of Communication, University of Amsterdam
PO Box 15793
1001 NG Amsterdam
The Netherlands

Contact: J.W.Boumans@uva.nl

Outsourcing the News?

An empirical assessment of the role of sources and news agencies in the contemporary news landscape

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 25 mei 2016, te 11.00 uur

door Jelle Wiebe Boumans

geboren te Amstelveen

Promotiecommissie:

Promotores:

Prof. dr. R. Vliegenthart, University of Amsterdam

Prof. dr. H. G. Boomgaarden, University of Vienna

Overige leden:

Prof. dr. P. van Aelst, University of Antwerp

Prof. dr. J. W. J. Beentjes, University of Amsterdam

Prof. dr. K. Raeymaeckers, University of Gent

Dr. P. Verhoeven, University of Amsterdam

Prof. dr. C. H. de Vreese, University of Amsterdam

Faculteit der Maatschappij en Gedragswetenschappen

TABLE OF CONTENTS

INTRODUCTION

DISSERTATION OVERVIEW

'Is the newspaper sector in a crisis?' Ask cult icon Jeff Lebowski – better known as *the Dude* – and he would probably ask the question 'Does the pope shit in the woods?' in return (Bevan, Coen, & Coen, 1998). Ask a random newspaper journalist or media scholar and they would most likely confirm as well, albeit perhaps more eloquently formulated. To ask the question is almost a platitude: The multiple challenges that the newspaper sector, the "stronghold of journalism" (Bakker, 2008, p. 427), faces have been widely documented for years now. Newspaper circulation rates and readership levels are at an all-time low (Bakker, 2008; Davis, 2000; Witschge, Fenton, & Freedman, 2010). The most recent numbers indicate that the circulation rate of daily print newspapers in Europe has decreased with 21.3% between 2010 and 2015, while the rates in the US (-8.7%) and Australia (-22.3%) also show a clear drop in that same period (World Association of Newspapers and News Publishers [WAN-IFRA], 2015). Meanwhile, ad revenues – the newspaper's life line – are steadily going downhill: from 47.4 billion in 2005 to 16.4 billion ten years later in the US (Pew Research Center, 2015). Worldwide, newspaper print advertising revenues have gone down with 17.5% between 2010 and 2015; In Europe, revenues have decreased by nearly a quarter in that period (23.1%) (WAN-IFRA, 2015). The steady but modest increase of online advertising revenues over that same period cannot nearly make up for the losses (Pew Research Center, 2015). Meanwhile, competition from other media is rampant (Van der Wurff, 2012), online news aggregators and rivaling organizations repurpose original news content without paying a cent for it (Freedman, 2010), and a successful online survival strategy for the newspapers has not yet been found (Jones & Salter, 2012; Pew Research Center, 2015; Van der Wurff, 2012).

The crisis is partly being managed by closing papers or shedding staff. Data in the US show that the journalistic workforce has decreased with 39% between 2005 and 2015 ("Employment in daily newspapers", 2015). Industry estimations on the workforce in the UK between 2000 and 2010 suggest a similar decrease (Nel, 2010), although it has also been suggested that the percentage of job losses in the UK is considerably lower than in the US (Franklin, 2012). At the same time however, editorial

content in the UK has increased substantially in absolute terms. Consequently, a journalist in 2006 was expected to produce three times as much output as his counterpart two decades earlier (Lewis, Williams, & Franklin, 2008a). This dissertation focuses on the developments in the Dutch context, which are largely in line with the global trends sketched above. The figures on circulation rates, advertising revenues, and journalism employment rates over the period 1998-2008 show a similar downward spiral (Tijdelijke Commissie Innovatie en Toekomst Pers [TCITP], 2009). In 1990, newspapers received 56.4 % of the total advertising revenues of the media market: by 2009, that share was marginalized to 5.2 % (Niewold et al., 2010). The economic crisis has further worsened the state of affairs on the advertising market: More than half of the net print media expenditures has vaporized between 2008 and 2013 ("Netto mediabestedingen opnieuw gedaald", 2014). Data on circulation rates over the most recent years show that the decrease continues ("Papieren oplage daalt verder", 2015). It is clear that the problems on the newspaper market are more urgent than ever, and future reorganizations and cutbacks seem inevitable if Dutch newspapers are to survive (TCITP, 2009).

CONSEQUENCES OF THE NEWSPAPER CRISIS

The consequences of journalism's crisis are not only felt in the professional sector, but have a far wider impact: News media are often viewed as the lifeblood of a democracy (Witschge et al., 2011). In the current mediatized society, the healthy functioning of journalism is of greater significance than ever before (Mazzoleni & Schulz, 1999). For democracies to function, it is essential that news media fulfil certain responsibilities. While there is a range of diverse and often conflicting ideas about what these democratic responsibilities of media entail (for a discussion see Berger, 2000), at the bare minimum news media are expected to 1) provide full, fair and relevant information to the public on which citizens can make informed choices, 2) facilitate an arena for public debate in which a wide range of voices are present, and 3) enable the participation of citizens in social and political life (Curran, 1996; McQuail, 2000). Furthermore, most perspectives on the role of the news include the responsibility being a guardian of democracy by holding those in power accountable (Fenton, 2011). This function is sometimes associated with the role conception of participatory journalism, in which journalists actively take sides and become advocates for a certain cause. Yet social grievances and political misconduct can well be reported on from a reasonably neutral and objective perspective – as journalists in fact also do (Hanitzsch, Hanusch, & Lauerer, 2016).

The economic hardship that news organizations face is believed to place considerable constraints on newspapers' ability to fulfill their democratic duties: expensive editorial commitments like foreign news reporting for instance has been "cut to the bone and beyond" (Davies, 2008, p. 99), with foreign bureaus being closed and staff replaced for subscriptions to the global news agencies (Freedman, 2010; Hafez, 2007). As a consequence, international news that is being consumed worldwide is largely supplied by three global agencies (Reuters, AP and AFP), who treat news foremost as "a saleable commodity produced in bulk" (MacGregor, 2013, p. 36). It is not surprising that this dominance raises concerns on news diversity and increasing homogenization (Boyd-Barrett & Rantanen, 2000; Hafez, 2007; MacGregor, 2013; Paterson, 2006). Another consequence of the pressures that media suffer is the decay of traditional investigative journalism, a key instrument to hold power and authority accountable in any meaningful way (Fenton, 2011; Marsh, 2013). The declining quality of the media is also said to be visible in an increased dependency on subsidized content provided by sources and news agencies (Davies, 2008; Lewis, Williams, & Franklin, 2008b). Lacking time and resources, journalists are pressured to incorporate public relations material and news agency copy into their news output (Davis, 2000; Davies, 2008; Lewis et al., 2008b; Phillips, 2010a). It is this dependency that is the focus of this dissertation. If journalists increasingly rely on subsidized contents, their ability to report accurately and independently is endangered. This dissertation asks to what extent public relations material affects the content of the national news agency and print and online news. Furthermore, it provides the first large scale assessment of the impact of news agency copy on print and online news.

MEDIA'S DEPENDENCY ON SUBSIDIZED CONTENT

The underlying mechanism explaining journalists' overreliance on subsidized content is evident: Fewer journalists are forced to produce more output, at a faster pace. To be able to cope with these demands, they become less proactive in searching for news and instead rely on subsidized content, provided by news agencies and sources that aim for coverage that advances their interests (Davis, 2000). This form of journalism has been coined *churnalism* (Davies, 2008; Harcup, 2015) and is believed to lead to less accurate and trustworthy news. The uncritical recycling of subsidized material – frequently without validating or cross-checking (Erjavec, 2005; Lewis et al., 2008a) – is believed to compromise journalism's autonomy, reducing journalists to "passive processors of whatever material comes their way, churning out stories, whether real events or PR artifice, important or trivial, true or false" (Davies, 2008, p. 59).

Thirty years ago, Sigal described news coverage as *"a sampling of sources' portrayals of reality, mediated by news organizations"* (Sigal, 1986, pp. 27-8). The key concern nowadays is that news coverage involves increasingly less mediation and more sampling. Studies repeatedly have found that journalists have a tendency to rely on elite sources, whose trustworthiness is generally accepted and who are often in a position that grants them access to accurate and specialized information (for an overview see Carlson, 2009). The social prominence of a source is a key filtering principle: The more prominent the politician, public figure, expert or organization, the more likely their statements are disseminated by the agency (Boyer, 2011). This tendency is allegedly even stronger for agency journalists compared to regular journalists due to the specific context in which they operate; the limited time to evaluate the overload of information combined with the agency's primary goal of delivering factual, trustworthy information (Boyer, 2011; Livingston & Bennett, 2003). One aim of this dissertation is to assess whether indeed certain source categories systematically enjoy better access to the agency's and newspapers' agenda. Unlike corporations, NGOs are traditionally shared under the category 'alternative' or non-mainstream sources that have to struggle to gain access to the news (Allan, 2004; Gans, 1979, Manning, 2001). Yet more recent research suggests that (resource-rich) NGOs have increasingly adapted to respond to the media's need for content, and that the current times provide more opportunities for NGOs to access the news – if they subscribe to journalistic criteria of expertise and professionalism (Castells, 2008; Fenton, 2010; Van Leuven & Joye, 2014). By comparing the impact of corporations and NGOs on the media agenda, the final three chapters investigate to what extent Dutch NGOs are succeeding in their attempts.

Simultaneously with the decline of journalism, the PR-industry has grown exponentially. Similar to reported trends in many other countries, there has been a doubling of PR-employees in the Netherlands between 2000 and 2010 (Prenger, Van der Valk, Van Vree, & Van der Wal, 2011). Consequently, the information that is being fed to the media has substantially professionalized over the last two decades: Most organizations and institutions – whether commercial, governmental or public – now have a public relations department whose main concern is managing media relations (Davies, 2008). Sources constitute a vital role in professional journalism: There is no story without a source. Sources are indispensable to confirm and validate the information that underlies news stories (Broersma, Den Herder, & Schohaus, 2013). The combination of a rapid decline of editorial resources and the expansion of professional public relations has created an unequal playing field in which sources dominate (Davis, 2000; Davies, 2008; Manning, 2001; Moloney, Jackson, & McQueen, 2013).

FOCUS AND RELEVANCE OF THIS DISSERTATION

A review of the literature reveals that churnalism has been subject to a plethora of normative considerations, but received considerably less empirical attention. A striking observation is that the largest provider of input, the news agency, is structurally overlooked. Given the centrality of the agency in the media landscape, the general scarcity of detailed academic analyses of the performance of news agencies and their exact impact on the news agenda and content is remarkable (Johnston & Forde, 2011; Paterson, 2006). Filling this void, the *first* contribution of this dissertation is an extensive empirical assessment on the news agencies' reliance on source content, as well as the impact of the agency copy on the news landscape. A possible explanation for the neglect of the role of the agency is a practical one: that reproduced news agency content is not always accessible. In the United States for instance, agency content that is verbatim replicated in newspapers is excluded from digital news archives like LexisNexis due to copyright issues (Weaver & Bimber, 2008). In the case of Dutch newspapers archived by LexisNexis such restrictions are absent, providing a unique window of opportunity to assess the impact of agency copy on the news on a large scale. The analyses deliver valuable new insights on what role the agency exactly plays in the news construction process, and to what extent this differs across different newspaper categories and media.

The majority of the literature on the relation between journalism's crisis and its impact on reliance on subsidized content stems from the United States (e.g. McChesney, 2003, 2008) or the United Kingdom (Davies, 2008; Franklin, 2012; Lewis et al., 2008b). A second contribution of this dissertation is that it studies the relation between sources, agencies and news media in a different context than typically has been done. Conducting research in other geographical and journalistic contexts is a valuable way to gain a better understanding of the factors that explain sourcing practices. Comparative research on journalism practices consistently finds that national contexts explain substantive variation in "the fortune of journalism" (Franklin, 2012: p. 665; Hanitszch et al., 2016; Tiffen et al., 2014), and the Dutch context differs in many respect from the UK and the US (Brüggemann, Engesser, Büchel, Humprecht, & Castro, 2014). With respect to the current Dutch context, the consensus is that newspapers' dependency on subsidized content is worrisome, but not (yet) as problematic as in other countries (Prenger et al., 2011; Hijmans, Schafraad, Buijs & d'Haenens, 2011; Kroon & Schafraad, 2013).

The third contribution stems from the unique datasets that have been gathered. Thus far the alleged increased reliance of journalists on both sources as well as agency copy has been subject of much debate, but considerably less empirical scrutiny. While the context – decreasing revenues and newspaper circulation rates, a decline of journalistic capacities, a steady increase of output – has been convincingly demonstrated, it remains an empirical question what the consequences of these developments are for news content. To our knowledge, no research on churnalism has drawn upon extensive longitudinal datasets that would put to test claims of a trend. Exceptions like the empirical studies of Lewis et al. (2006) and Van Leuven, Deprez, & Raeymaeckers (2013) aside, the majority of doom and gloom predictions are based on plausibility rather than empirical proof. The scarce attempts to trace churnalism practices and the dominance of news agencies predominantly stem from qualitative studies or small scale empirical analyses. While their findings are important and inform the debate, it remains the question to what extent the results reported in these studies are representative for journalism across the board. It is argued that systematic large-scale empirical analyses are essential for an informed debate on the state of journalism. An important contribution of this dissertation therefore lies in the creation of large-scale collections of texts that can be used to put claims on churnalism and intensifying source and agency reliance to the test. Two of the three datasets span a period of a decade, including a variety of sources and newspaper types.

A fourth contribution is the development of automated tools of analysis to optimally build and analyze the gathered data. This not only offers researchers a reliable and objective measure to study source reliance and churnalism, it also aspires to stimulate journalism researchers' awareness of the potential of automated tools. While journalism theory has beyond doubt been advanced over the past decennia, this is considerably less the case for the methodologies use in journalism research (Karlsson & Sjovaag, 2016). Automated content analysis (ACA) methods for instance are typically not part of the journalism researcher's methodological arsenal (Boumans & Trilling, 2016). Yet, digital developments have profoundly changed both the nature of the data under study as well as the sheer amounts of it, and so the application of new methods is much-needed. A methodological aim of this project was therefore to develop and apply an automated content analysis tool that enables the media researcher to study large scale datasets. This is relevant to the academic community for two reasons. First, it creates opportunities for the systematic analysis of large-scale text collections without massive funding support. Second, automated content analyses can identify patterns in data that traditional analysis cannot, or only with great effort (Flaounas et al., 2013). The approach proposed in this

dissertation illustrates this point: It introduces three measures that allow for a very specific assessment of the degree of overlap between texts, the advantages of which will become clear in the next section.

OUTLINE OF THE DISSERTATION

This dissertation explores the relationship between organizational press releases, agency content and news content, and is organized as follows. Chapter One empirically investigates to what extent the news agency steers national print and online news, accounting for both differences within and between these two categories. The chapter introduces two measures: The intermedia agenda-building ratio indicates what percentage of news articles is initiated by the news agency, while the churnalism index indicates to what extent a news article consists of replicated agency material. Chapter Two expands the scope of the analysis by considering the impact of subsidized content in the shape of organizational press releases on both news agency and print content. By tracing this impact for a considerable number of organizations over a period of ten years, I am able to investigate whether indeed media have become more reliant on subsidized content over the recent past. A related concern is that some sources are structurally more successful in accessing the news agenda than others. By comparing the influence of press releases from NGOs and corporations, this point is addressed in both Chapter Two and Chapter Three. The focus of both studies is different however: while the second chapter – like the first chapter – investigates to what extent reproduce literal content, the third chapter is concerned with the construction of meaning by news media. It assesses to what extent the themes in the media are a reflection of the themes promoted by sources in the context of a highly contested issue: nuclear energy. Chapter Three furthermore adds the role of the regional newspapers as well as the second largest news agency to the equation. Finally, Chapter Four expands the applicability of the proposed approach of automated text comparison to the field of framing research. The following sections provide an outline of each chapter.

Chapter One: The Agency Makes the (Online) News World Go Round

Chapter One starts with exposing the impact of the news agency on the news landscape. This study addresses the question *to what extent the news agency content influences the agenda and content of print and online news.* The dataset consists of all publicly available news that has been published in 2014 in the print and online

versions of three main Dutch national newspapers as well as the largest online news provider, nu.nl. Furthermore, all the articles published in that same period by the Dutch' current only national news agency, *ANP*, are integrated in the dataset. In total, the study compared 119,452 news agency articles with 75,434 print news articles and 171,727 online news articles. An innovative automated tool determines the level of overlap between an agency text and a news article. This information is used to indicate what percentage of the total news articles has been initiated by agency copy, and how strong the content overlap is. In a final step, the study also investigates to what extent news organizations are transparent in their reliance on the agency by measuring agency attribution in the articles that are based on agency copy. The study presented in *Chapter One* extends previous research in two important ways. First, reliance on agency copy has typically been measured on the basis of manifest attributions to the agency (Hijmans et al., 2011; Van Leuven, Deprez, & Raeymaeckers, 2014; Powers & Benson, 2014; Sjovaag, 2014). This is problematic because research has shown that often news organizations veil their reliance on agency copy, and do not attribute the agency (Reich, 2010). Studies that do take the actual agency copy into account generally rely on case studies, which limit their generalizability. In contrast, the approach presented in this study compares the news content with agency copy in a systematic and automated fashion. This implies not only that the methodological shortcoming of previous studies is bypassed, but also enables tracing agency copy in news content with an unprecedented accuracy, on an unmatched scale. A richer overview of the impact of the agency on the news landscape is thereby provided than previously has been possible. A second contribution is that the automated method produces very precise measures, facilitating a systematic comparison. This is all the more relevant since the study makes an explicit comparison between print and online news, and also distinguishes between quality, popular, and free titles. Comparing print and online news providers is important because literature has suggested that online news is increasingly the primary source of information, particularly for young consumers (Mitchelstein & Boczkowski 2010; Trilling & Schoenbach, 2015). While potentially, more space equals more news, it has been demonstrated that online news is basically an alternative market for agency copy and extensively recycled material (Doyle, 2015; Fenton, 2010; Johnston, 2009). To date reliance on agency content has been studied either in the print context (Lewis et al., 2008), or online (Johnston, 2009), but never in a comparative fashion. *Chapter One* presents the first attempt to investigate both print as well as online news reliance on agency copy in one study.

Results show that online news is highly dependent on the agency's information supplies. The agency is responsible for the majority (66 percent) of the online agenda; this is even up to 75 percent in the case of the largest online news provider, nu.nl. The agendas of the print titles are statistically significantly less strongly dependent on the agency's input: overall, 23 percent of the print articles are initiated by agency copy. There is large variation between the quality, popular and free titles however, with percentages ranging from 12% (popular) and 16% (quality) to 48% for the free outlet. The churnalism measure revealed that many of the online articles are more or less verbatim agency copy. In other words, the news article consists practically entirely of agency copy and contains no additional information. For print, the churnalism measure is significantly lower – with again the free outlet scoring substantially higher than the quality and popular newspaper. With regard to the degree of attribution, results demonstrate that online news providers attribute the agency in an overall consistent way, and perform better in this respect than their print counterpart. One newspaper title, *De Telegraaf*, does not attribute the agency at all (except for visual content), neither in print nor online.

Chapter Two: Subsidizing the News?

The study described in *Chapter Two* focuses on a step earlier in the news production process: The relation between source material and news content. It is one of the few studies that investigate the triangular relationship between sources, news agencies and newspapers (see also Forde & Johnston, 2013). The study relies on the same two impact indicators as the first study: one that determines whether an article is *initiated* by a press release, and one that determines to what extent subsidized *content is literally reproduced* in the news article. While the relation between PR-material and newspaper content has extensively been studied, only seldom has this been done for news agency content. Yet we know that much like any other player on the news market, news agencies too face many pressures and are struggling to maintain profitable (Fenton, 2011; TCIFP, 2009; Vermaas & Jansen, 2009). Furthermore, time pressures – often seen as central explanatory factor of churnalism – are indefinitely higher in the newsroom of an agency compared to that of a newspaper, making agencies especially vulnerable to churnalism practices. An important contribution of this chapter therefore is that it provides unique insights in the relationship between press releases and news agency copy by comparing 4,455 organizational press releases to 6,142 agency articles that refer to these organizations. A second contribution of the study is that it looks at patterns over time. To investigate the seldom verified claim of an emerging trend of source reliance, the study analyzes the impact the organizational press releases over a period of ten years.

It furthermore clarifies on the role of two important contextual factors: the type of source (NGO or corporation) and the type of newspaper (quality, popular and free).

Findings indicate that overall, one in every ten news articles included in this study is initiated by a press release; for the agency this is slightly higher (16 percent). These ratios remain stable over time; there are thus no indications of a trend of increased agenda building capacities of sources. News on NGOs is relatively more often initiated by a press release than news on corporations. NGOs thus appear more successful than corporations in setting the media agenda, particularly with respect to the agendas of the news agency and the free newspapers. Results on the churnalism index score indicate that overall, differences between the content of a news articles and the content of a press release it is based on, are fairly large. Literal copy-paste practices have not been found. When looking at the different media and source categories, a similar pattern emerges from the data as found for the agenda building ratio: first, the free newspapers and the news agency show the largest reliance on source content; and second, the content of NGOs is to a greater extent used in news articles than corporate content is.

Chapter Three: Nuclear Voices in the News

In *Chapter Three,* the debate on nuclear energy serves as illustration of the extent to which organizational press releases have an influence on the content of the agencies and newspapers. The study also compares the reliance on the selected sources of the two national news agencies during that period. The position of the traditional news agency *ANP*, active since 1934, was threatened in this period by a new rival: *Novum*, whose main strength lies in the fact that they are cheaper, but whose quality is questioned by its clients (*NRC*.nl, 2008). By comparing the content of the two agencies, we can assess to what extent the two agencies rely on subsidized content for their coverage of the nuclear energy issue. A second important contribution of the study is that apart from agencies and national newspapers, it includes regional news as well. Holding regional powers to account, these newspapers have an important function for the democratic functioning at the regional level. Yet, it has been suggested that the extent of newsroom cuts in local and regional news makes this sector particularly vulnerable to PR material (Jackson & Moloney, 2015; O'Neill and O'Connor, 2008; Prenger et al., 2011). The longitudinal approach enables assessing whether indeed regional news has increasingly became reliant on subsidized content. The dataset consists of 393 organizational press releases (two corporations and one NGO), 947 articles from the two national news agencies, and 2,132 newspaper articles from three national newspapers and two regional

newspapers. To assess the degree of overlap between the press releases and news content, I rely on a measure called cosine similarity. This measure has proven its value across various disciplines and is also at the core of the measures developed for study one and study two.

Consistent with the second study, the findings show that the overlap of media content is largest with the press releases of NGOs. In other words, news agencies and newspapers typically refer to the issue of nuclear energy in terms that match the themes of Greenpeace (i.e. danger, waste and disasters), rather than in terms of safety, benefits and progress, like the corporate advocates. A notable exception however is one of the two regional newspapers, whose themes strongly resemble that of the regionally active corporation while Greenpeace's themes are absent. Although it is a very particular context and thus hard top draw strong inferences from, this finding does seem to support concerns raised in literature on an unhealthy dependency of regional newspapers on pre-packaged news from resource-rich organizations (Jackson & Moloney, 2015; O'Neill and O'Connor, 2008). Again in line with study two, data suggest no trend of an increasing similarity of newspaper content with the content of either the sources or the news agencies. Regarding the comparison between the two agencies, results indicate that the mean cosine similarity score of the second largest national agency *Novum* is considerably higher than is the case for *ANP* with respect to all three organizations. It implies that *Novum*'s copy is to a larger extent shaped by the organizations' content and may explain why *Novum*, which has been taken over in 2015 by *ANP*, was able to offer their services at lower rates than *ANP* (Haakman, 2008).

Chapter Four: Introducing cosine similarity to assess frame overlap

Chapter Four foremost has a methodological aim and presents an attempt to advance research on implicit framing, an automated framing approach. Implicit framing analysis is a powerful technique to extract meaning from large collections of texts. As the object and volume of data changes in the current digital world, techniques like this are becoming more and more relevant (Karlsson & Sjovaag, 2016). At the basis of the implicit framing approach is the idea that meaning is constructed through the selection and co-occurrence of words. Using clustering techniques like factor analysis, the researcher is able to make sense of patterns of co-occurring words, or 'implicit frames' (Hellsten, Dawston, & Leydesdorff, 2010). The chapter first presents an overview of recent implicit framing studies and then points to limitations of the studies relating to the interpretation and articulation of the results. Most notably, interpreting the results

is often done subjectively on the basis of the word visualizations or the dominant terms in the factors, which undermines a key strength of the approach – the value-free determination of patterns. Furthermore, interpreting the data this way only provides limited insight in the extent to which frames from different domains overlap. The implicit framing approach would greatly benefit from a standardized measure a less subjective technique to assess content overlap. This article proposes that the cosine similarity score, introduced in Chapter 3, is a suitable measure for this purpose. Since implicit frame extraction and cosine scores share the same basis – weighted word frequencies –the two approaches can be considered complementary to each other: The implicit frames provide a qualitative understanding of the meaning of the text collections, while the cosine similarity scores provide quantitative insight in the degree to which the different text collections share frames.

Relying on the same dataset used for the previous study, *Chapter Four* demonstrates the added value of cosine similarity as a means to interpret frame overlap. Results show that the higher the cosine similarity between two corpora, the more likely it is that there are corresponding frames. Likewise, two text collections that score low on the similarity indicator have little or no mutual frames. As such the cosine similarity measure provides a precise and objective tool to assess the degree to which the frames of different text collections overlap. Consistent with the findings in Chapter Three, the implicit framing analysis demonstrated that the frames of the NGO are overall most well reflected – again, with the exception of the regional newspaper. These results are also consistent with the findings of a manual analysis of a subset of the data (Boumans & Vliegenthart, 2014).

KEY FINDINGS

This dissertation provides the empirical data that is critical for an informed debate on the role of the news agencies and sources in the news production process. The studies described in this dissertation have addressed various characteristics of this triangular relationship, and found several important results that contribute to our insight in these dynamics. The key findings are summarized below.

1. Dutch (print) newspapers are no copy factories

Concerns on the impact of increased economic pressures on journalism are widespread. Part of the concern is that the shrinking journalistic capacities lead to an overreliance on 'news subsidiaries': content provided by sources. Newsroom studies suggest that PR-sources are a main source of information (Erjavec, 2005) and it has been demonstrated that journalists publish public relations material almost or completely unchanged (Sissons, 2012; Lewis et al., 2008b), mainly due to a lack of time (Erjavec, 2005; Davies, 2008). Both academics as well as professionals fear that newspapers are turning into copy factories (Lewis et al., 2008b: p. 47). In this climate of growing concern about "rampant PR and weakening journalism" (Jackson & Moloney, 2015: p. 2), the results presented here for the Dutch context are comforting. The large-scale comparison of press releases with news articles described in the second chapter shows that about one in ten news articles are initiated by a press release. If there already is a "Niagara of propaganda" (Moloney, 2006:p . 2) as reported in other Western democracies such as the United Kingdom and the United States, Dutch journalists appear relatively well capable to withstand its force. Furthermore, the churnalism index measure indicates that in the case that a press release is being followed on, the news content differs substantially from the press release. In other words, journalists do not exclusively rely on the provided information, but add new information as well. It is telling that no more than three of the 4,455 press releases included in the study were literally replicated in the newspapers. The findings of the case study on nuclear energy, described in the final two chapters, also reject the idea that news would be a replication of source content. It is worth noting in this respect, however, that the similarity score of the regional newspaper and the regionally active nuclear organization's content is relatively high. This is consistent with previous findings on regional journalism's relative vulnerability to PR (O'Neill & O'Connor, 2008; Jackson & Moloney, 2015) and gives reason for concern as it regarded indicative of the deplorable state of the regional news (Fenton, 2011). The regional newspapers fulfill an

important function of keeping people informed on their regional community and holding regional powers accountable. Furthermore, they have a signaling function for national newspapers. When regional newspapers collapse, parts of the map will no longer be covered (Broersma, 2009). The results thus invite further research for the relationship between sources and regional newspapers.

2. No signs of increased newspaper reliance on agency copy

As a result of the increasing marketization of news and growing economic pressures for news organizations, literature suggests that news media's reliance on the cost-efficient services of agencies has grown over time. The diminishing journalistic staff is required to produced increased levels of output, a combination that is effectively turning journalists into "churnalists" that repackage existing content – predominantly agency copy - rather than create unique content (Davies, 2008; Fenton, 2011; Lewis et al., 2008b). Since no study has presented longitudinal data on this matter, it remained an empirical question whether indeed we can speak of a trend. This dissertation aimed to address that lacuna. The results of the studies described in Chapter Two and Chapter Three are unambiguous: No signs of an increased reliance on agency copy since the beginning of this millennium have been found across the Dutch print news landscape. In light of the literature this is an important finding, since it contradicts the dominant belief a causal relation between the economic crisis and churnalism practices. This observation strongly calls for longitudinal analysis in other contexts. Furthermore, it would be very insightful to investigate this relation over a longer time period. Judged by circulation rates, the decline in the Dutch newspaper sector already started in the mid-1990s (Bakker, 2009). It may well be that a longitudinal analysis that includes data from the early 1990s onwards does reveal a trend of increasing reliance on agency copy.

3. News agencies do redistribute press releases relatively more than newspapers, but rarely verbatim

The concerns raised above have also been associated with news agencies, namely that they extensively rely on subsidized content and lack the time for proper fact-checking, background research and adding original material (Johnston & Forde, 2011). In the words of a PR-professional: "PA [Press Association] does not do as much of the probing and difficult questions. They are journalists but to some extent they are an information service" (Davies, 2008: 91). Newsroom observation at a national German agency reports a "dizzying spectacle of circulation and flow" of information: A single agency

journalist typically receives up to 5,000 pieces of information per day (Boyer, 2011). Critics argue that as a consequence of this information overload and limited journalistic capacity, agency articles that appear to be the work of journalists may actually well be PR artifice (Johnston & Forde, 2011). Lewis and colleagues describe the news production process as a "a clear linear process in which PR material is reproduced by agency journalists whose copy is, in turn, reproduced in the news media" (2008b: 15). Agencies thus form an important mediator between sources and journalists. To date, however, only anecdotal evidence of the agencies' dependence on PR material has been presented. Chapter Two and Chapter Two present the first empirical assessments of this claim, and the picture that arises does not correspond with the doom and gloom scenario sketched above. Illustrative is the percentage of literal replications of press releases by the agency found in the Chapter Two: .02 percent (one article on a total of 6,142 articles analyzed).

4. The news agency dominates the agenda and content of free newspapers and online news

As already suggested by Paterson fifteen years ago on the basis of an exploratory study, online news content is heavily dependent on the national news agency (Paterson, 2001). The large scale and recent dataset that Chapter 1 draws upon provides convincing evidence that in the current Dutch news landscape, this is indeed the case. While the results may not be surprising in the context of free newspapers, the extent to which mainstream online news drives on agency copy is disconcerting. The findings unequivocally show that by and large, both the online news agenda as well as the literal content of the news that is not behind a paywall is based on one source only: the national news agency. Although, as mentioned above, the agency's content is decidedly more than merely reproduced PR-material, it does give the agency considerable influence on what the daily talk at the coffee machine is about. This raises concerns on the level of homogeneity and potential lack of diverse viewpoints in online news, concerns that have been voiced by other scholars earlier (Fenton, 2010; Doyle, 2015). The findings are even more alarming when the agency's economic hardship is taken into consideration: According to the agency's president, the increasing pressures are compromising the agency's capability of providing the desired quality to their clients (TNO, 2011, p. 25). This concern was expressed a year before another substantial round of layoffs further reduced the journalistic workforce at the agency with 15 % ("Persbureau *ANP* gaat flink bezuinigen", 2011). Judged by the most recent available figures, this measure has not stopped the bleeding: In 2014,

the agency suffered an operational loss of nearly one million euro (Brandenburg – Van de Ven, 2015). The pressures – caused by declining incomes from newspapers and broadcasters and illegal spread of agency content on the internet – are threatening the agency's continuity (TCITP, 2009; TNO, 2011). As becomes clear from Chapter One, the implications for the online news landscape of a collapsing agency are colossal: effectively, the online news would be depleted of content.

5. NGOs are more successful in accessing the news agenda

The systematic comparison between the content of NGOs and corporations in Chapters 2 and 3 lead to a conclusion that contrasts the view of corporations as privileged sources that are better able to influence the media's agenda (Lewis et al., 2008). Instead, results suggest that NGOs are more successful than corporations in their efforts to spread their message. This could be explained by the fact that the majority of the NGOs included in the studies are established, international organizations with considerable budgets. This is in line with the view that the dissemination of professional public relations has brought NGOs many opportunities (Davis, 2000) and the observation that NGOs are increasingly presenting themselves as experts. As such, over the recent past these organizations may well have gained promotion in the normative order of 'credible' sources (Allan, 2004, p. 63).

METHODOLOGICAL REFLECTION

As mentioned, a specific aim of this dissertation has been to develop and apply a tool for automated content analysis. Overall, the tools lived up to its expectation: It made possible the analysis of large-scale datasets and produced precise and well-interpretable output. Each having their specific qualities, the proposed measures transcend the theoretic context in which they are applied here: They can equally well be applied to measure concepts like news diversity or content homogenization, strands of research that also study content overlap. The operationalization of content overlap in these studies is typically limited to the question whether two articles are about the same topic or event (Bozckowski and de Santos, 2007; Mujica & Bachmann, 2015). It is debatable however whether one can speak of homogenization only on the basis of a shared topic; it may for instance be that both articles articulate radically different or opposing perspectives on that topic. The proposed automated approach in this dissertation is a valuable instrument for a more refined assessment of content homogenization.

When comparing the automated approach to a traditional manual analysis, ACA's ability to handle large amounts of data is an important advantage. This can for instance facilitate longitudinal analysis, as demonstrated in this dissertation. The ability to analyze more data over longer period of time creates more concise results, and is helpful in determining (the lack of) patterns. The capacity to analyze large scale datasets also enables researchers to move beyond context-dependent research and towards cross-national research. There is a growing awareness of the importance of different levels of influence that shape journalism and news (Curran, Iyengar, Brink Lund, & Salovaara-Moring, 2009; Hanitzsch et al., 2010; Hanitzsch, Hanusch, & Lauerer, 2016; Hanitszch & Mellado, 2011; Humprecht & Büchel, 2013; Tiffen et al., 2014). Given the considerable volume of data that such studies typically require, automated approaches like the one proposed here demonstrate its value in various stages of the research, including data collection, processing and analyses. I will come back to this point in the final discussion.

LIMITATIONS

Building on large-scale datasets and advanced content analyses techniques, this dissertation has provided unique empirical insights to a field that is characterized by normative theorizing and assumptions: news media's reliance on subsidized content provided by sources and news agencies. With the specific interest in news agencies, it has put the spotlight on a partner of news media that typically remains in the shadows and contributes to a better understanding of the influence of content from sources and agencies on the news. However, valuable as these insights may be, many aspects of the media's reliance on sources and agency copy and the functioning of agencies remain unaddressed.

An important implication of our content-analytical research design is that it allows for the systematic observation and comparison of the characteristics of the input and output of media, but we cannot draw any inferences on the journalistic process that has led to the final media output. Content analysis does not give insight in the practice of journalism, and I would like to stress the limitations that this brings along. First, an iron rule in journalism is the confirmation and validation of potential news stories. It matters whether an information subsidy that has been picked up and disseminated by the news agency, is thoroughly checked by that agency. Equally, this applies to the news agency copy that forms the basis of a news article: Research suggests that agency copy is typically

not checked (Davies, 2008; Lewis et al., 2008a). In the worst case, then, there is little to no checking involved by either the agency or the news organization, resulting in a news article that consists of unverified content from the PR-desk of an organization or institution (Lewis et al., 2008b). It is thus clear that the validation practices of journalists are an important facet of news media's reliance on sources which our quantitative input-output analysis cannot address. What we can tell on the basis of the proposed measures however is the extent to which information has been added by the journalist. Thus, when the churnalism index indicates that substantial editorial efforts have been made, it is safe to assume that alternative sources have been consulted. While this does not imply that the original content has been checked, it does increase the likelihood that inaccurate information is detected before it reaches the actual news columns.

A second limitation of the research design relates to our choice to focus our study on the impact of organizational press releases in the final three chapters. While press releases constitute an important part of an organizations' external communication (Forde and Johnston, 2013), public relations comprise far more activities – many of which leave few traces (Jackson & Moloney, 2015; Reich, 2010). News media's reliance on organizational press conferences, video material, and direct mail or phone contact for instance, is not accounted for in this dissertation. Interview studies with PR-professionals and journalists suggest that the PR industry is relying less on traditional monolithic press releases, instead employing increasingly sophisticated strategies to achieve positive news coverage (Davis, 2008; Jackson & Moloney, 2015; Reich, 2010) – it could be that particularly corporations apply these strategies and hence dedicate relatively less attention to press releases. I do not claim to have sketched a complete picture of the impact of PR on the news, and research on other interactions between sources and journalists are required to complete the picture. Having said this, recent research has indicated that press releases and other classical news gathering channels are still central communication instruments for organizations (Forde and Johnston, 2013; Van Leuven et al., 2014) and continue to be a legitimate object of academic interest.

A word of caution is also in place with respect to the selection of the specific organizations for Chapter Two: While all efforts were made to construct a representative sample of press releases from the corporate and nongovernmental domains, it is not unlikely that the composition of the sample has specific effects on the outcomes of the study. Large within-group variations in the values of the similarity measures within the groups suggest that factors that are not accounted for explain a considerable part of the organization's media access. Thus, caution is warranted when extrapolating the results from the sample to a broader population.

While pioneering this methodological approach, a number of limitations arose that are worth acknowledging. I will mention two of the most prominent limits I encountered. First, the standardized measures of churnalism and cosine similarity lead to values that enable assessing statistical differences and have intuitively natural bounds of 0-1. Yet relating the values to the substantive reality can be challenging, as the score on itself is still a rather abstract value. It is evident that on a scale from 0 to 1, the difference between a cosine similarity score of .34 and a score of .89 is large, but how this difference is manifested in the actual texts only becomes clear after returning to the data and studying examples. Coming to grips with the implications of the measures' values thus requires some effort. Given practical limitations of space, the researcher will typically be able to provide one or two examples in an article. It is thus a challenge to make the results tangible, a challenge that can only become less when the familiarity of the reader with the measure increases. The second limit concerns the attempt to combine the cosine similarity score with the implicit framing approach described in Chapter Four. The technique behind the implicit framing approach has proven very helpful in extracting the dominant topics present in a collection of texts. It remains debatable however whether frames in the traditional sense of the word can be deduced from the word clusters that form the technique's output. Proponents of the technique argue that it can extract 'latent meanings' from texts which are hard to trace by a manual approach (Leydesdorff and Hellsten, 2005). After having applied the technique, I am not fully convinced the technique does justice to the depth and complexity that is inherent in language, and which framing research typically seeks to lay bare. Related to this, one could question the similarity measure's ability to assess the degree of frame overlap between different text domains. It can be argued that it rather demonstrates to what extent the same topics are covered. Since it does not for instance take the valence of the texts into account, the possibility remains that two text domains that focus on the same topic(s) yet take opposing stand points in their evaluation of it, do score high on the similarity measure. This is a limitation of the current technique which, while not a

light task, could be overcome by enhancing the analytical tool with a valence measure. Since multiple frames can occur within one text, ideally the tool is able to determine the valence on the frame level rather than the article level. In short, these considerations are illustrative of both the potential as well as the current limitations of automated content analysis approaches.

MAIN CONCLUSIONS AND IMPLICATIONS

The overall picture of Dutch journalism that arises from this dissertation is not as dark as one would expect on basis of the literature. On the basis of the data the existence of a strong journalistic routine of copying and pasting subsidized content can be dismissed – both for news agencies as well as newspapers. What has been confirmed is that the national news agency is an important partner of news organizations. For online news organizations, they are more than that: Both the agenda as well as the content is a direct reflection of the agency output to a large extent. The dominance of the agency in the online news landscape has a number of important consequences. First, the fact that the majority of online news stems from one and the same source is highly undesirable in light of news diversity. The lack of added journalistic value (as illustrated by the churnalism index) implies that not only the news agenda but also the very way in which issues on that agenda are framed comes from the agency. Knowingly or not, in a society where news is increasingly consumed online, the agency strongly determines the boundaries of public debate. This monopoly not only leads to a relatively narrow range of perspectives, according to literature, the type of information that passes the agency's gates may also be structurally biased. In light of this discussion on unequal access of sources to the news agenda and the news agency in particular, the results of Chapter Two and Chapter Three on the news access of NGOs are worth highlighting. The relative success of NGOs to access the news that this dissertation has found, supports the call for a more nuanced consideration of traditional notions on unequal access. For instance, the specific causes that NGOs stand for should be taken into account (Deacon, 2003), as well as the multiple and often changing relations between NGOs and journalists (Waisbord, 2011). Apart from these contingencies, large resources to develop sophisticated media strategies are the key explanatory factor whether an NGO is successful in accessing the news (Manning, 2001; Waisbord, 2011). The overwhelming majority of non-profit institutions cannot rely on these resources and are likely to remain 'outsiders' both in the political as well as the media arena (Phillips, 2010).

A second concern about the central role of the news agencies is related to their production routines. Summarized, the main responsibility of a news agency is to filter and structure the large quantities of circulating information, present it to their clients, and help them manage the information by suggesting priority orders (Czarniawska, 2011). Thus, rather than proactively acquiring information, agencies strongly rely on incoming information. This implies that a potential news story or a public issue has little chance to get coverage if it hasn't been presented by a source. Yet apart from generating positive publicity, a large part of PR-work consists of keeping stories out of the media to limit bad publicity (Curtin, 1999). According to Aeron Davis, who studies the relation between PR and the media, 'for every story fed to the media, there is one being carefully kept out' (2002, p. 27). News agencies are not designed to trace these stories, and will typically not reveal news that does not want to be revealed: Investigative journalism has never been the responsibility of the agencies (Phillips, 2010). It can thus be reasoned that the larger the role of agencies in the news production process, the smaller the likelihood that journalism is able to subject political, governmental and corporate power to rigorous examination. In this context, the finding that generally nine out of ten articles in the paid newspapers is initiated through other channels than the news agency is comforting. Furthermore, in light of the value of investigative journalism, the rise of alternative forms of journalism is a positive development. The success of collaborative journalism as well as a range of new organizations aimed at in-depth reporting for instance, demonstrates that this type of journalism does not need to come only from traditional journalistic outlets (Carlson, 2011). Similarly, the number of crowdfunded journalism projects is witnessing exponential growth (Vogt & Mitchell, 2016). In the Dutch context, the online journalism platform De Correspondent is a prime example. These initiatives represent a new segment of nontraditional journalism that is driven in large part by public interest and largely takes place online.

FUTURE RESEARCH

The key findings of this dissertation provide clear directions for future research. Foremost, more insight is needed in how subsidized content is treated by journalists from news agencies, print and online news. Particularly for the first category there is little research to draw upon. Is information consistently crosschecked? Or is that unfeasible given the current time pressures, as suggested (Erjavec, 2005; Lewis et al., 2006; Van Leuven, Deprez, & Raeymaeckers, 2015), and do journalists instead use

quotes to divert responsibility (Diekerhof & Bakker, 2009)? And to what extent does the checking of information depend on other factors, like journalistic intuition or the authority of a source? A related question is what strategies journalists apply when following up on agency copy. Is the follow-up a matter of contacting the sources already mentioned in agency content to get an additional quote, or has the journalist critically interrogated the story and explored counter perspectives? It is these types of question that a design like the one presented here obviously cannot address, and where the advantages of for instance ethnographic approaches (Boyer, 2011; Czarniawska, 2011; Erjavec, 2005) or reconstruction interviews (Reich, 2010) are evident.

The second, arguably more complex research agenda would aim to build a comprehensive model that can account for the diverging findings on churnalism in different contexts. The conclusions of this dissertation underline "the first lesson" drawn from a recent comparative study on sourcing routines: "news practices – and hence, in important ways, news content – are far from uniform" (Tiffen et al., 2014, p. 387). As mentioned earlier, Anglo-American research is characterized by disconcerting observations on a fast declining journalistic quality and autonomy and comparable observations have been made in other countries, among which Australia (Johnston & Forde, 2011; Forde & Johnston, 2013), New Zealand (Sissons, 2012) and Slovenia (Erjavec, 2005). However, the general picture that emerges from this dissertation and other research in the Dutch (Hijmans et al., 2011, Scholten & Ruigrok, 2009), but also for instance the Belgian (Van Leuven et al., 2013, 2015) context, is considerably less grave.

Shrinking revenues are often seen as underlying cause for media's declining workforce and consequential increased dependency on subsidized content (Davis, 2000; Jukes, 2013). Just as observed by Van Leuven and colleagues (2013), the longitudinal findings of this dissertation suggest that there is more to the story than that. The Dutch and Belgian newspaper sectors have experienced comparable decreases in circulation rates, revenues and employment rates in the past decades as the United States and the United Kingdom, yet neither a journalistic routine of copy and paste practice nor a trend of increased reliance on subsidized content has been found from either organizations or the news agency. This observation is consistent with the findings of a large-scale comparative survey study that found that journalists perceive organization, professional, and procedural influences as more powerful constraints than economic influences (Hanitzsch et al., 2010). It is clear that the Anglo-American findings cannot simply be extrapolated to other contexts, and that the various structures in which journalists work need to be taken into account to explain variance in source reliance.

These levels of influence can be conceptualized in different ways: Preston (2009) for instance, distinguishes between the domains of individual and organizational forces, media routines and norms, political-economic factors, and cultural and ideological forces. Other conceptualizations can be found in Hallin & Mancini (2004), Hanitzsch et al. (2010) and Brüggemann et al. (2014). On the level of media routines, the power relations between journalists and sources are identified as a primary factor, where a relationship is suggested between the number of journalists and the number of PR-professionals (Carlson, 2009; Moloney, Jackson, & McQueen, 2013; Prenger et al., 2011) as well as the volume of produced content per journalist (Lewis et al., 2008b). On the organizational level, ownership has been established as a key factor that shapes news production (Hanitzsch et al., 2016; McChesney, 2008). In the case of the Netherlands, recent experiences with "robber barons" APAX and Mecom have demonstrated the disturbing impact that ownership can have on media organizations and journalism (Broersma, 2009, p. 27; Bakker, 2009). Similarly, a relation between the diversity of the news and the concentration of ownership has been established (Baker, 2009), with more monopolistic media markets such as the US containing less diversity (Powers & Benson, 2014). On the political level, variations in for instance legislation, press subsidies and the strength of a public service system account for much of the cross-national variation in journalistic cultures (Hallin & Mancini, 2004; Hanitzsch et al., 2016; Tiffen et al., 2014). It is evident that variations exist in many of the above-mentioned factors between the countries where high levels of churnalism and source reliance are reported (including the US and the UK) and countries where low levels of churnalism are found (like the Netherlands and Belgium). A cross-national comparative research design could explicate to what extent the demonstrated variations in source reliance can be explained by variations on these various levels, and identify possible interaction effects between levels: Certain factors at the individual and organizational level may matter more in particular political or economic contexts. The academic community has only just begun to unravel the dynamics between the various factors that shape news (Hanitzsch et al., 2010), and there is a need for comparative integrative research to further theoretical insights (Hanitzsch, Hanusch, & Lauerer, 2016). By empowering researchers to thoroughly examine large-scale datasets, the proposed automated approach could offer a valuable contribution to the thriving comparative research venue.

CHAPTER 1

THE AGENCY MAKES THE (ONLINE) NEWS WORLD GO ROUND:

The Impact of News Agency Content on Print and Online News

Abstract

While agenda setting research has a varied and rich history, a key player in the news production process is systematically overlooked: the news agency. Given the equally central as well as precarious position of the news agency in the construction of news, investigating the influence of agencies on news providers is of vital importance. This study addressed the question to what extent the news agency content influences the agenda and content of print and online news. Results show that particularly online news is highly dependent on the agency's information supplies, with the agency being responsible for up to 75 percent of the online agenda. Furthermore, many of the online articles are more or less verbatim agency copy. The results provide a strong rationale for intermedia agenda setting researchers to place news agencies prominently on their academic agenda. Moreover, the findings on the agencies' online news domination should be alarming to anyone who is concerned with news diversity.

Contributing authors:
J.W. Boumans, D. Trilling, R. Vliegenthart, & H.G. Boomgaarden

News agencies have a long history in the world of the news. Developed as a joint initiative by newspapers in the 1830s and 1840s, the traditional purpose of the agency was to lower the costs and expand the scope of foreign correspondence. In the past decennia however, their reach has increased drastically: today's agencies can be described as "wholesale news providers" (Paterson, 2006), covering everything from politics to sports and from financial to entertainment news. They have a reputation of delivering maximally accurate, maximally fast, and maximally factual information (Boyer, 2011). As described by the editor of the British national agency Press Association (PA), they are "the central heart of the media industry" (Davies, 2008, p. 74). News media increasingly rely upon news agencies both to cope with the 24/7 news cycle and the need to be "first with the news", as well as to reduce their production costs (Boyer, 2011; Forde & Johnston, 2013). This is even more so the case for online journalism (Johnston 2009; Paterson, 2006). Unlike most of the other material that a journalist draws upon to produce a story, news from agencies has an authority so absolute that the news media generally do not question its content or factual accuracy (Davies, 2008; Johnston & Forde, 2011).

Given the centrality of the agency in the media landscape, the scarcity of detailed academic analyses of the performance of news agencies and their impact on the news agenda is remarkable. A probable explanation is the fact that in the United States – where the majority of agenda setting research takes place – agency content that is verbatim replicated in newspapers is systematically excluded from digital news archives like LexisNexis due to copyright issues (Weaver & Bimber, 2008). Consequently, to our knowledge no recent studies have systematically compared agency content with news content. We argue that, in order to fully understand how media agendas are shaped, it is essential to compare actual agency content with news content. To this end, the current study investigates the influence of these "silent partners of the news" (Johnston & Forde, 2011) in the Dutch context. Not being hindered by any copyright issues, we present the first large-scale investigation into the impact of the agency on the news by comparing agency and news data over the period of an entire year (2014). The automated approach that is specifically developed for this purpose enables us to assess the impact of the agency on the news with an unprecedented accuracy. It explicates the influence of the principal national news agency on two related levels of the news production process: First, on the media agenda level (to what extent is the news agenda initiated by the agency) and second, on the content level (to what extent does the news content consist of agency copy?).

Given the scarcity of empirical data on these matters, our general research question is descriptive in nature: *To what extent are the national news agenda and news content based on agency copy?* In theoretical terms, the first level concerns the gathering of information and is embedded in the intermedia agenda setting framework. Intermedia agenda setting is concerned with the question how news media affect each other's attention for issues. It is the mechanism that creates a common definition of what news is and what not (Vliegenthart & Walgrave, 2008). The second level focuses on the processing of the information and will be related to the notion of churnalism, or the practice of recycling content provided by sources or news agencies (Davies, 2008). By merging these two theoretical frameworks in one research design, this study aims to advance a detailed explication of the influence of the news agency on the national news. A second contribution is that the study assesses the influence of the news agency on print as well as online news in one design. Pointing to different professional routines and scarcer journalistic resources, scholars argue that online journalism is relatively more susceptible to rely on agency copy (Doudaki & Spiridou, 2013; Paterson, 2006). Yet to date, reliance on agency content has been studied either in the print context (Lewis et al., 2008), or online (Johnston, 2009). This study presents the first exploration of agency reliance in a comparative fashion.

Apart from the aim to inform the debate on agency reliance with new evidence, the study makes a methodological contribution by introducing an innovative automated tool that enables tracing agenda setting and churnalism patterns in large-scale datasets. At best, reliance on news agency copy is measured by counting the references to the agency in the credit line (Hijmans et al., 2011; Powers & Benson, 2014; Sjovaag, 2014; Van Leuven et al., 2014) – a notoriously inaccurate measurement given news organizations' inconsistent attributing practices (Quandt, 2008, Reich, 2010). Studies that do take the actual agency copy into account generally rely on case studies (Johnston & Forde, 2011), which limit their generalizability. Notable exceptions aside (Lewis et al., 2006), the same applies to research on churnalism. Relying on a large scale dataset that includes agency copy as well newspaper content and applying automated content analysis, the measures used in this study provide unique insights in the impact of the agency on the print and online news landscape.

The few empirical studies that have traced agency copy in newspaper content found that the reliance of news organizations on agency copy is rarely attributed (Hijmans et al., 2011; Johnston, 2009; Lewis et al., 2006; Scholten & Ruigrok, 2009). In a climate of declining public trust in journalism, transparency is increasingly seen as an important norm to maintain journalism's authority as information provider and uphold the public's trust (Karlsson, 2011; Phillips, 2010b; Plaisance, 2007). Source attribution is a central instrument to promote transparency. In this light, the lack of attribution and the purposely veiling of reliance on agency copy (Reich, 2010) is not contributing to public trust (Quandt, 2008; Reich & Godler, 2014). The automated analysis presented here is able to detect agency reliance as well as (lack of) attribution, allowing for a highly accurate observation of the extent to which Dutch print and online news media perform on this indicator of transparency.

INTERMEDIA AGENDA SETTING

Traditional agenda setting research is concerned with the effect of the media agenda on the public agenda (McCombs & Shaw, 1972; for an overview see Weaver, McCombs & Shaw, 2004). The specific strand of agenda setting research that focuses on the degree to which media influence each other is known as intermedia agenda setting (Lopez-Escobar et al., 1998). The vast majority of research highlights the influence of national newspapers on the agenda of national television news (Golan, 2006), while attention has also been directed towards the relationship between newspapers and internet news (Lee,Lancendorfer, & Lee, 2005), national newspapers and the local news (McCombs & Funk, 2011), and the influence of entertainment-oriented media on the news (Soroka, 2000). In sum these studies show that intermedia influences are an important explanatory factor of the news agenda of media (for an overview see Vliegenthart & Walgrave, 2008).

In spite of the wide variety of intermedia relationships that have been studied, there is very little recent research that explicitly addresses the influence of the news agency. This has not always been the case: Many of the early studies started their investigations with arguably the most obvious influencer to consider, reporting substantial agenda-setting effects (Gold & Simmons, 1965; Whitney & Becker, 1982; White, 1950). Yet as a result of increased copyright protection in many countries, including the U.S., disseminated wire stories by newspapers are not retrievable in the largest news archive used for academic research, LexisNexis (Weaver & Bimber, 2008). As an unfortunate consequence, the news agency as actor has largely disappeared from the academic

radar, particularly so in (intermedia) agenda setting research. A rare opportunity provided by Google's News service enabled Weaver and Bimber (2008) to assess the role of the wire services in more recent years (2006-2007). They conclude that "the importance of the wire services to news is not, at least in this case, confined to local coverage of national issues but extends to coverage in major newspapers." (p. 524). In line with this observation, the modest amount of recent research on agencies – mostly situated in a non-US context - continues to underline the central role of the agency in the shaping of daily news. It is therefore not unlikely that the high level of homogeneity between the news agendas of newspapers, television news and other outlets – often explained by pointing to the dominant role of elite media or the professional norms that news media share (Protess & McCombs, 1991; McCombs & Funk, 2011) – can in fact to a considerable extent be attributed to the agencies' influence. The next section will elaborate on the role of the agencies in the news landscape.

News agencies are a central hub of information and fulfill various functions for news media. Most of the scholarly attention that agencies have generated stems from the decade 1970-80, when the dominance of the major western-based agencies on the production and dissemination of international news was examined (Boyd-Barrett, 1980; Golding & Elliott, 1979; Phillips Davison, 1974). By determining which global issues are worth covering, the news agencies hold considerable power to set the international agenda for public debate. The coverage of global issues has been criticized for being biased towards Western interests (Joye, 2000). Apart from their core task of gathering and disseminating information, agencies have an important signaling function: Many newsworthy events only become news after the agencies report on it (Boyd-Barrett, 1980). The wider implications of the capacity to set the news agenda are that agencies have a large say in what voices and ideas make it into news discourse and, consequently, are part of political discourse and debate. When a government department, political party, civil organization, or major corporation wants to reach the public with a statement, they send it to the agencies. From the perspective of the sender, it is logistically far more preferable to disseminate information via an agency instead of contacting individual journalists. An additional advantage is that mainly due to the high workload of an agency journalist, they typically do not cross-examine the information they receive as thoroughly as a regular journalist (Davies, 2008; Johnston& Forde, 2011).

Only more recently, a number of concerns about the role and functioning of agencies on the national level have been raised. These concerns vary from agencies' inability to cover the entire spectrum of potential news events (Davies, 2008) to the detrimental impact of the agencies' dominant position for news diversity (Johnston, 2009; Paterson, 2006; McChesney, 2003, 2008).

Furthermore, it has been pointed out that investigative journalism has never been a task of agencies. The more the news landscape is shaped by agency content, the less critical it may become (Phillips, 2010a). The core task of an agency journalist is to accurately report what people say, rather than verify the truthfulness of it. In other words, they report the news rather than make it (Davies, 2008, p. 83). A final concern is related to the speed of the agency's news cycle combined with the reported journalists' high workloads (Davies, 2008). At the German bureau of Associated Press, a news agency journalist typically receives 4,000 to 5,000 potential news stories per day (Boyer, 2011, p. 10). Through ethnographic observations, an average of 97 different discrete activities per hour was measured, implying that the journalist changes her focus every 37 seconds. These numbers illustrate the speed at which journalists need to cope with information. Agency journalists need to decide in a split second what stories will be sent out and how they will be covered.

A series of developments have resulted in challenging times for news agencies. First, the internet has made news – once published – omnipresent and accessible instantly. The agencies have thus lost their traditional monopoly on 'raw' news and find themselves faced with multiple parties that operate on their market (Boyer, 2011). In the digital age it is increasingly difficult to secure exclusivity of the content: As soon as a medium publishes it online, other media are able to profit for free (TCITP, 2009, p. 49). Second, as a consequence of the media organization's cost orientation, the agencies are under increased pressure to deliver more services at lower prices (Phillips, 2010a). They respond by cutting substantively in personnel and the pioneering of new markets (Manning, 2008), including offering their expertise to organizations in writing successful press releases (TNO, 2011). Yet, the continuity of the agency is unwarranted (TCITP, 2009) and the recent developments have led the president of the Dutch national news agency *ANP* to express his concern that the agency has reached a "critical borderline of providing the desired quality" (TNO, 2011, p. 25). The critical position of the agency notwithstanding, newspapers continue to rely on agency copy ("*ANP* bepaalt in grote mate de nieuwsagenda", 2015).

CHURNALISM

This increased reliance of news organizations on agency copy is often explained by referring to journalists increased workload (Boyer, 2011; Johnston & Forde, 2011) and the need to cope with the "news cyclone" (Klinenberg, 2005). According to Davies (2008), many journalists are in fact information brokers who mainly recycle existing content: A practice that is dubbed *churnalism*. While different terms have also been used (see for an overview Van Leuven et al., 2014), the critique is consistent: Journalists habitually incorporate subsidized content from agencies, other media, or PR-professionals into their news output (Davis, 2000; Jackson & Moloney, 2015; Lewis et al., 2006; Reich, 2010). In a sense, churnalism has more far reaching consequences than agenda-setting: When journalists are less well able to conduct the crucial practices of fact-checking and independent investigation, the news may become structurally biased by the specific interests of those gaining access to the media (Jackson & Moloney, 2015). This is particularly discomforting when the massive resource advantages of official and corporate actors are taken in account (Davies, 2008; Lewis et al., 2006). Critics are concerned that the reactive and desk-bound practice of today's journalism leads to a decline of the diversity of viewpoints in the news, as less resourced voices are unable to make themselves heard in the "Niagara of PR propaganda" that floods the journalistic system (Moloney, 2006, p. 2). The growing dominance of news agencies in the news production process is feared to only increase this unequal access, since the tendency to rely on established sources is allegedly even stronger for agency journalists due to the specific context in which they operate: the limited time to evaluate the overload of information combined with the agency's primary goal of delivering factual, trustworthy information (Boyer, 2011; Livingston & Bennett, 2003). The widespread concerns notwithstanding, few quantitative analyses of agency and news content have been presented that indeed substantiate the claims that journalists routinely resort to churnalism of agency copy (but see Lewis et al., 2006, 2008). This study aims to advance research on churnalism in two ways: First, by offering a detailed analysis of the degree to which different newspaper titles and categories rely on agency copy on the basis of a substantial data set, and second, by introducing a measure to standardize the degree of churnalism. This measure will be explained in the method section.

ONLINE VERSUS OFFLINE NEWS

Online news is increasingly the primary source of information, particularly for young consumers (e.g. Mitchelstein & Boczkowski, 2010; Trilling & Schoenbach, 2015). Web journalists have different work routines and role conceptions compared to print journalists. Immediacy, for instance, is of greater value to an online news journalist: online journalism has led to the rise of 'high speed news' and a constant publication pressure. This sense of immediacy leads to a situation where online news journalists have generally less time than traditional journalists to create news content. Recent studies have linked greater time pressures to less original journalistic output and reliance on fewer news sources (Reich & Godler, 2014). The lack of time not only results in less research and original writing, but also less cross-checking (Erjavec, 2005; Lewis et al., 2008a; Quandt, 2008). Apart from immediacy, there are a number of other factors that explain why online and offline media should be seen as distinct "manufacturing houses of news" (Reich, 2015, p. 14). Technological and market restrictions as well as different resources, tools and working routines all contribute to online media reporting being generally less meticulous compared to print media. Often lacking working business models (Humprecht & Buchel, 2013), it could be assumed that online media in general rely strongly on the most efficient way of producing news: disseminating agency copy. Indeed, as a consequence of particularly time and resources constraints, online journalists are characterized as distributors rather than creators of content (Doudaki & Spyridou 2013, p. 916).

While the web is sometimes viewed as an alternative news source, others claim it is an alternative market for agency copy (Doyle, 2015; Fenton, 2010; Johnston, 2009). According to Paterson (2006), who analyzed the reliance on news agencies in the context of international news, the Internet presents "the dangerous illusion of multiple perspectives which actually emanate from very few sources" (p. 20). A recent cross-national study on print and online news homogeneity found that the prevalence of news agencies is up to four times higher in online news as compared to print in all three studied countries (the US, France and Denmark) (Powers & Benson, 2014). Yet, this observation is made on the basis of counting explicit references to the agency in the byline – a method that has proven notoriously unreliable (Quandt, 2008). Another study focusing on the news flow of a limited amount of press releases found that the majority of 'breaking' online news content is wholly or predominantly taken directly from agency copy, with an additional 20 per cent containing verbatim agency copy (Forde & Johnston, 2013). While the study applies a novel approach and delivers informative results in the process from PR-message to newspaper content, its findings cannot

merely be generalized to the broader population of online news. Tracing the flow of a limited number of PR-releases (n = 65), the conclusions are drawn on the basis of the analysis of only 37 agency stories and 62 news articles. Furthermore, these articles are based on the output of a specific source category: governmental and political institutions. The innovative design of the current study is not hindered by these two constraints: It compares ***all*** agency copy with ***all*** the news of the largest national news titles over the period of an entire year (2014), which allows us to make an informed conclusion on the agency's influence on the print and online news. Guided by the previous findings we tentatively formulate the following hypotheses:

H1: Print news will have a relatively lower percentage of articles that is initiated by agency copy than online news.

H2: The replication of agency copy will be lower for print news than for online news.

ATTRIBUTION TO THE NEWS AGENCY

Research has suggested that the recent developments in particularly the online news production results in reliance on less sources and less cross-checking (Quandt, 2008; Reich, 2010). These practices may compromise the accuracy of the news. Transparency is increasingly seen as an important norm to maintain journalism's authority as information provider and uphold the public's trust (Karlsson, 2011). Enabling the public to inspect the working process of journalists not only indicates the journalists' accountability, it also promotes the legitimacy of the news production towards the public (Rennen, 2000). Attributing to sources – such as the news agency – is a prime example of the journalistic code of transparency. Yet while the role of the agencies seems to be growing, it has been argued that dependency on agency journalism is rarely openly recognized by the agencies' clients (Boyer, 2011; Paterson, 2006). Content is often subtly rewritten and either not attributed or repurposed under a journalist's byline. This reluctance to be transparent about reliance on agency copy may be explained by the desire to present original news products (Boyer, 2011).The practice of poor or incorrect attribution to agencies has been empirically demonstrated in various countries (Quandt, 2008). For example, a lack of source attribution is found to be one of the key characteristics of online news texts in Greece. Print articles perform slightly better, but still attribute less than half the time (Doudaki & Spiridou, 2013). Similarly, content analyses of two Australian online newspapers revealed that a large proportion of their news content that is attributed to a newspaper journalist, is in fact primarily wire copy (Johnston, 2009). A study of Dutch print newspapers demonstrated

consistent lack of agency attribution in newspapers as well: Half of the national paid-for newspapers include articles that are demonstrably based on agency copy, yet do not mention the agency as a source (Scholten & Ruigrok, 2009). This study extends the last study by explicitly comparing print with online news. Our third hypothesis is based on the research already conducted on the matter:

H3: There is more attribution to the agency in print news than in online news.

METHOD

DATA COLLECTION

The impact of the agency on the news can be assessed on several levels of the production process. This article introduces an innovative automated tool to investigate two related levels: first, on the level of the media agenda (to what extent is the news agenda initiated by agency releases) and second, on the content level (how similar is the news content to the agency content?). To assess these questions, an extensive dataset has been collected: It includes all releases of the Dutch' largest news agency, *ANP*, of the year 2014, as well as all news articles that appeared in both the print as well as online edition of the national newspapers *de Volkskrant*[1], *De Telegraaf*[1], and *Metro*. These newspapers are respectively the largest quality, popular and free newspapers of the Netherlands. Agency and print data has been retrieved from the LexisNexis archives. The online data were collected in real time by running a script on a server that queried the RSS-feeds of the websites once an hour and downloaded the full text of the article. Analysis of the data has confirmed that in the case of the Dutch LexisNexis database, there is no "wire service blind-spot" as reported by Weaver and Bimber (2008): multiple cases have been found where newspaper content is identical to agency content.

[1] *Regarding the online news of de Volkskrant and De Telegraaf, it must be noted that only freely available content has been collected. While the majority of the articles are available, part of their content is behind a paywall. Our findings on the online editions of de Volkskrant and De Telegraaf are thus limited to their freely available content.*

Finally, we have also included news site Nu.nl, the largest 'pure player' on the Dutch online news market with an average of 2.5 million unique visitors per day ("bereik nu.nl", *n*.d; Trilling & Schoenbach, 2015). Given our interest in the impact of the agency on current affairs, a series of sections has been filtered out. This filtering is based on a close examination of the news sections as labeled by the titles. Filtered sections include opinion pieces, columns, editorials, cartoons, traveling, showbiz, lifestyle, weather, and reviews of books, films and music. Since our focus is on (inter)national news, regional sections have also been excluded. The final dataset consists of 119,452 agency articles (equalling 327 articles per day) and 247,161 news articles. Table 1 provides an overview.

Table 1. Total Number of Articles in the Dataset

Agency			Print			Online		
	n articles	average length	Title	*n* articles	average length	Title	*n* articles	average length
ANP	119452	169	*Volkskrant*	25191	501	*Volkskrant*.nl	49471	246
			Telegraaf	38625	299	*Telegraaf*.nl	72378	148
			Metro	11618	186	Metronieuws.nl	34503	146
						nu.nl	15375	235
			total *n* news articles	75434			171727	
			mean news article length		329			194

From Table 1 we can infer that for all three newspapers, the total number of online news articles is substantially higher than their print edition. The popular *De Telegraaf* produced by far the most news articles, both in their print as well as online edition: a daily average of respectively 124 and 198 articles, based on six print publication days and seven online publication days per week. The online-only medium nu.nl has published the least articles of the selected media: n = 15375 articles, equaling 42 articles per day. It is worth to keep in mind that these numbers reflect the net amount of articles on current events, not the entire content of the newspapers and news sites. In terms of article length, the average length of the online articles is substantially lower than that of the print articles.

ANALYSIS

We developed a software tool to automatically compare the contents of print and online news with agency content. In a first step, it parses and preprocesses the text (stopword and punctuation removal) and creates a tf-idf representation of each text (see for an elaborate explanation the method section of Chapter Three or Boumans & Trilling, 2016). In a second step, it calculates two measures we propose: the agenda setting ratio and the churnalism index. The calculation of both is described below. The tool is developed with open source software: the Python libraries NumPy (version 1.10.4; Van Der Walt, Colbert, & Varoqaux, 2011) and SciPy (version 0.16.1; Van der Walt et al., 2011) and scikit-learn (version 0.17.0 ; Pedregosa et al., 2011).

Intermedia agenda setting ratio. We operationalize the intermediate agenda setting ratio as the percentage of news articles that is initiated by agency copy. The determination of whether an article is initiated by an agency release is based on a measure called cosine distance. This measure indicates how similar two documents are likely to be in terms of their subject matter (Tan, Steinbach & Kumar, 2006), or put more simply: to what extent they share the same words. A cosine score can take any value between [0,1], depending on the degree of similar terms in the two texts. A score of zero implies that two documents do not share any terms, while a score of 1 implies that the words in the two documents are identical (disregarding the word order). A systematic manual analysis of a subset of the data showed that a similarity score of .65 and above indicates a content overlap that is too high to be reasonably explained by chance: The majority of the content is identical. This value therefore serves as threshold to determine whether or not there is a link between a media article and a press release. When the cosine score is .65 or higher, the software reports the existence of a link. With this threshold as criterion, it can thus be determined for every individual news article if it is initiated by (based on) an agency release. As an additional condition, we specified that the agency copy must be published on the same day, the day before, or two days before the news article. The formula of the ratio is as follows:

$$\textit{Intermedia agenda setting ratio} = n\_ia \,/\, n\_ta \qquad (1)$$

where *n*_ia is the number of media articles that is initiated by an agency release and *n*_ta is the total number of media articles on that day. The following hypothetical example illustrates the agenda setting ratio. Let us assume that on a certain day, the newspaper *De Telegraaf* publishes 40 articles (*n*_ta = 40). The software tool finds that 10 of those newspaper articles are based on agency releases that have been published in the three

days before the newspaper article (*n*_ia = 10). The agenda setting ratio is thus 10/40 = .25, indicating that 25% of the *De Telegraaf*'s coverage on that day is initiated by (based on) the agency.

Churnalism Index. The next step is to establish the extent to which the news article's literal content is shaped by the agency copy. The Churnalism Index is our measure to assess this. At the core of this index is the Levenshtein distance (lev), a well-established measure in computer science and information theory that is among others often used in plagiarism detection tools. Lev is the minimum number of edits that is required to change one string of words into the other. An edit can be the insertion, deletion or substitution of a word. The more two texts differ from each other, the more edits are required to make the two texts identical and thus the higher the lev score. When two texts are identical, the lev score is zero. For the purpose of this study we are interested in the degree to which a news article is made up of agency copy. When a news article is a shortened replication of agency copy, the lev score equals the number of excess words of the agency text. Yet, the act of merely cutting does not constitute a journalistic contribution since there is no information added: The end product still entirely consists of agency copy and can be regarded as a prime example of the copy and paste practice that is at the core of the churnalism notion. Therefore the measure controls for length difference. Furthermore, we want to analyze the ***relative*** effort a journalist has put into a media text. Twenty alterations of a source text that consists of only 30 words are not the same as twenty alterations based on a 500 words text. We therefore also control for the length of the journalistic text. Formally then, the formula for the Churnalism Index is the following:

$$Churnalism\ Index = 1 - (lev_{a,b} - (LI_a - LI_b))/LI_b \qquad (2)$$

Where $\mathrm{lev}_{a,b}$ is the Levenshtein distance between source text a and media text b, LI_a is the length indicator of the source text and LI_b is the length indicator of the media text. By definition, the value of the Churnalism Index ranges from 0 to 1. The measure is inversed to facilitate interpretation: the higher the score, the higher the overlap (and thus the higher the degree of churnalism). A systematic manual analysis of a subset of the data showed that when the Churnalism Index > .7, the two texts are nearly identical, whereas an Index score of 0 indicates that while there is some resemblance in terms of topic and word usage, the media text differs substantially from the source text.

Results

Our overall expectation is that online news media rely more on agency content than print news media. The first dimension of this reliance considers to what the news agenda is initiated by the agency (*H1*), the second considers to what extent the actual content is based on agency copy (*H2*). Finally, we shall assess whether indeed online titles attribute the agency less than the print titles, as hypothesized (*H3*). We shall first discuss some key descriptives that can offer a first answer to our explorative research question on the extent to which the national news agenda and content are based on agency copy. Table 2 presents the number of articles that is based on agency copy per title and section.

Table 2. *Number of articles based on agency copy per section per newspaper*

	Title							
Agency	*Volkskrant*	*Volkskrant*.nl	*Telegraaf*	*Telegraaf*.nl	*Metro*	*Metro*.nl	nu.nl	Total
Section								
Domestic	987	9713	586	13462	2104	6328	4639	37819
	(20.7)	(26.6)	(13.3)	(33.4)	(35.9)	(37.2)	(43.1)	(31.6)
Foreign	660	9620	559	9719	3912	5832	3912	31039
	(13.8)	(26.4)	(12.7)	(24.1)	(36.4)	(31.7)	(36.4)	(26.0)
Economic	1167	1963	1756	5341	823	142	823	11454
	(24.4)	(5.4)	(39.9)	(13.3)	(7.7)	(.8)	(7.7)	(9.6)
Sports	1554	10702	1487	11758	485	4904	485	32278
	(32.5)	(29.3)	(23.7)	(29.2)	(4.5)	(28.8)	(4.5)	(27.0)
Undefined	410	4483	8		923	247	894	6965
	(8.6)	(12.3)	(.2)		(15.7)	(1.5)	(8.3)	(5.8)
Total	4778	36481	4396	40280	5864	17003	10753	119555

Note. data from January – December 2014. Percentages scores per column are in parentheses. Information on sections is provided by the agency.

In total, 119,555 news articles that are predominantly based on agency copy have been published in 2014. This implies that overall nearly half of the news content is based on agency copy (48.4%). Before specifying this ratio for the different titles, we first look at the distribution of agency copy per section. Overall, the domestic news section has the largest share of intermedia agenda setting content: 32 % of the articles based on agency copy falls within this section. However, there are variations in the extent to which online and offline titles rely on agency copy for the different sections. The quality *de Volkskrant* for instance uses agency copy predominantly for their sports coverage (32.5 % for print, 29.3 for online title), while for the print popular *De Telegraaf* it is economic news (39.9 %). For the online titles, domestic and foreign news are relatively most strongly based on agency copy.

The extent to which the agency exerts influence on the news can also be approached from the angle of the output of the agency. Thus, we could investigate what proportion of the agency's total output is followed up on by the news media. It can be argued that higher proportions are indicative of a greater structural dependency on the agency. On the other hand, lower proportions of followed-up agency copy would indicate that media deliberately cherry-pick from the large volume of agency supply to optimize their content. The combination of this proportion and the intermedia agenda setting ratio are an indication of the convergence of the agency and news agenda.

As depicted in Table 1, the total number of agency articles in 2014 amounts to 119,452. Results indicate that more than two of every five articles published by the agency are followed up on at least once (48,884 reproduced unique *ANP*-articles / 119,452 total *ANP*-articles = 41 %). On average, a followed-up article provides input for 1.76 news articles (*SD* = 1.08). When distinguishing between online and offline, results indicate that the online media are considerably more relying on the agency's agenda: 58 % of the agency's content is picked up by at least one online news site, whereas only 17 % of the agency's content is used by the print media. There thus appears to be a high demand for agency copy on the online news market: over half of the agency's proposed items on the agenda are taken over by at least one online news provider. In the next section we look into the intermedia agenda setting ratio to determine to what extent this reproduced copy shapes the print and online news agendas.

INTERMEDIA AGENDA SETTING RATIO

Figure 1 depicts the intermedia agenda setting ratio per title.

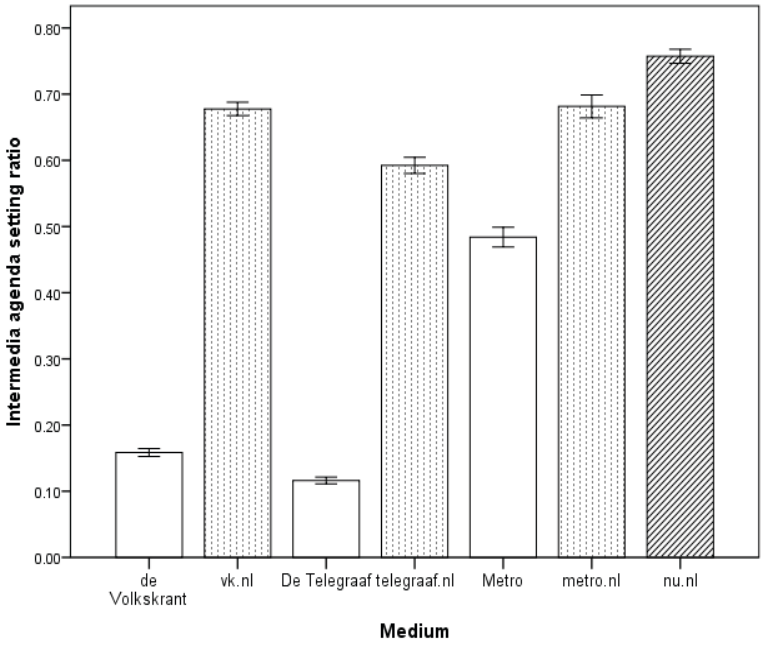


Figure 1. *Intermedia agenda setting ratio per title. Online media are patterned.*
Note. *Error bars represent standard errors.*

From Figure 1 it becomes clear that for all three newspapers, the online edition is to a greater extent initiated by agency content. The ratio is highest for the online-only title nu.nl. The ratio indicates that 75% of the news articles that appear on the website are based on agency copy. The news agenda of nu.nl is thus predominantly driven by agency input. To compare, the ratio of the print titles is 16% for *de Volkskrant,* 12% for *De Telegraaf,* and 48% for *Metro*. An independent-samples t-test was run to determine if there are differences in the agenda setting ratio between print and online news. Levene's test for equality of variances indicated heterogeneity of variances ($p = .00$). Analysis confirms that online media rely statistically significantly more strongly on agency content than print media do (*M online* = .68, *SD* = .13; *M print* = .23, *SD* = .18, *Mdiff* = .44, $t(1537) = 64.82$, $p = .00$).

This finding is thus in line with our expectation that the agency has relatively the most impact on the online news agenda. Relating these outcomes to the findings that the online news providers reproduce 58 % of the agency's total output, it becomes clear that the agency's agenda to a large degree ***is*** the online news agenda. *H1* is thus confirmed: The online news agenda is to a greater extent initiated by the news agency.

Thus far the focus has been on the extent to which the agency shapes the news agenda. The second part of the analysis sketches a detailed insight in the extent to which the agency influences the content of the news.

CHURNALISM INDEX

The second hypothesis stated that the content of online news media is more similar to news agency copy than the content of print newspapers. Figure 2 presents the average churnalism index score for all the news articles that have demonstrable overlap with agency copy.

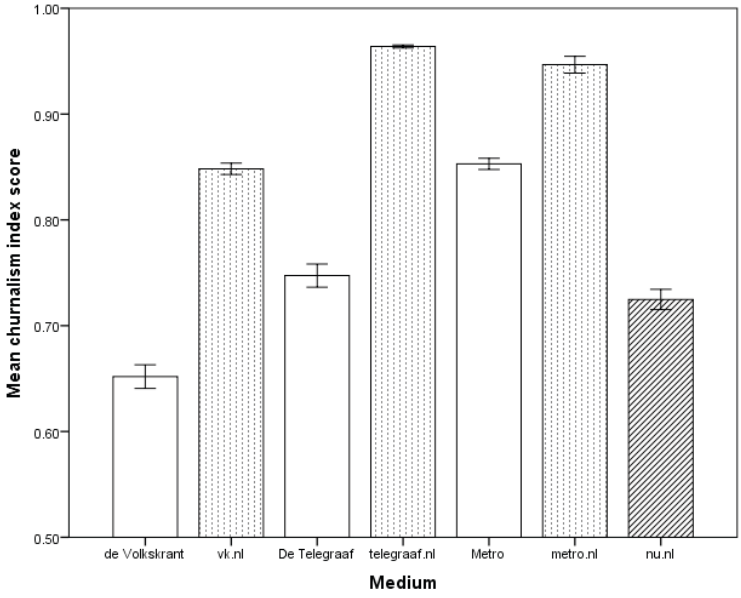
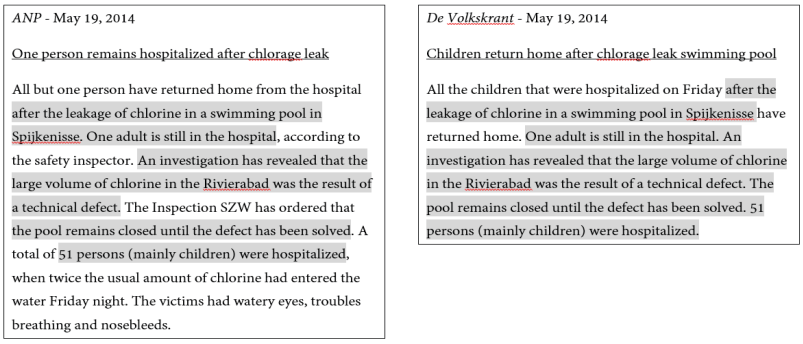


Figure 2. Mean Churnalism Index Scores per medium. Online media are patterned.
Note. Error bars represent standard errors.

Figure 2 indeed indicates that articles in the online edition generally score higher on the churnalism index than the print edition. This is true for all three newspaper titles. In other words, print articles that are based on agency copy have been edited to a larger extent than online news articles. To assess whether these differences are statistically significant, an independent sample t-test has been performed. There was homogeneity of variances, as assessed by Levene's test for equality of variances (p = .801). Results confirm that the churnalism index score is higher for online (*M* = .87, *SD* = .11) than for print, *M* = .74, *SD* = .12, a statistically significant difference, Mdiff = .13, SE = .00, t(2667) = 29.82, p = .00. The expectation that the content of online news media is more similar to news agency copy than the content of print newspapers (*H2*) is thus confirmed. However, it is worth noting the performance of online-only news provider Nu.nl: With a mean churnalism index score of .72 (*SD* = .09, it is slightly lower than the average print score, and in fact the lowest score after the print quality newspaper *de Volkskrant* (*M* = .65, *SD* = .10). Thus, while the far majority of their texts comes from the agency (ratio of .75), they rewrite and edit this input relatively more than the other media. In other words, it appears that the journalistic routine at nu.nl can best be described as re-purposing agency copy. A final result worth highlighting is the difference within print newspapers: The pattern that emerges is in line with the reasoning that quality newspapers, due to larger journalistic capacities, are least inclined to verbatim reproduce agency copy, followed by the popular and free newspapers.

To illustrate what the various scores practically mean in terms of content overlap, we will briefly discuss the mean scores. A churnalism index score of .87 (mean score of the online titles) indicates that the agency copy is verbatim taken over by the news outlet: typically, the agency text is shortened and at most one or two words are rephrased. An index score of .74 (mean score of the print titles) indicates that the agency text has been edited to a larger extent, yet still rule out that the texts differ substantially from each other: the majority of the news article still consists of agency content. Figure 3 demonstrates this in an example of an agency text and a newspaper article with an index score of .74.



ANP - May 19, 2014

One person remains hospitalized after chlorage leak

All but one person have returned home from the hospital after the leakage of chlorine in a swimming pool in Spijkenisse. One adult is still in the hospital, according to the safety inspector. An investigation has revealed that the large volume of chlorine in the Rivierabad was the result of a technical defect. The Inspection SZW has ordered that the pool remains closed until the defect has been solved. A total of 51 persons (mainly children) were hospitalized, when twice the usual amount of chlorine had entered the water Friday night. The victims had watery eyes, troubles breathing and nosebleeds.

De Volkskrant - May 19, 2014

Children return home after chlorage leak swimming pool

All the children that were hospitalized on Friday after the leakage of chlorine in a swimming pool in Spijkenisse have returned home. One adult is still in the hospital. An investigation has revealed that the large volume of chlorine in the Rivierabad was the result of a technical defect. The pool remains closed until the defect has been solved. 51 persons (mainly children) were hospitalized.

Figure 3. Example of two texts with a churnalism index score of .74. Identical phrases are highlighted.

Figure 3 shows how the newspaper article (depicted on the right) has copied the majority of the agency's text, yet chose a slightly different perspective by focusing on the children, as indicated by the headline and first sentence. The article contains no other information than the info provided by the agency.

ATTRIBUTION

Our final research question concerns the extent to which titles attribute the agency as a source. Figure 4 provides an overview of the agency attribution ratio per title.

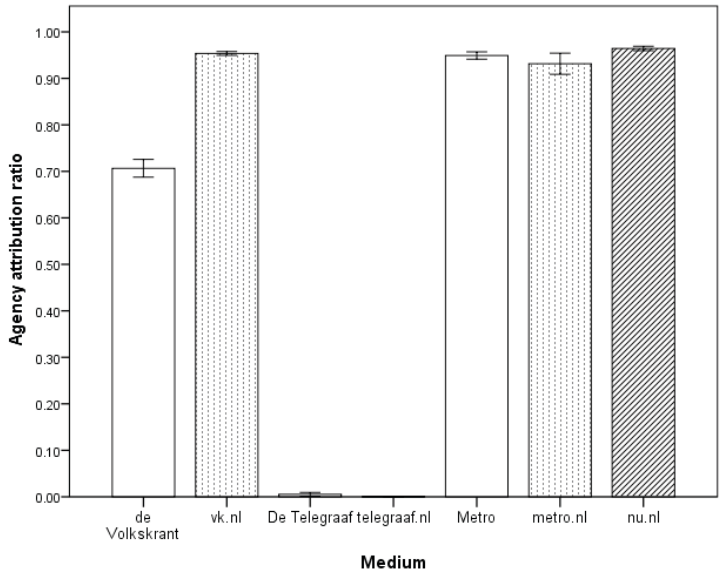


Figure 4. Agency attribution ratio. Online media are patterned.
Note. Error bars represent standard errors.

A glance at Figure 4 quickly learns that the popular *De Telegraaf* attributes the agency neither in print nor online. Thus, readers of *De Telegraaf* are not informed that the information they consume comes from an agency rather than a journalist of *De Telegraaf*. It was our expectation that online titles would attribute less often than print titles (*H3*). For *de Volkskrant*, the data suggest a reverse pattern: whereas the print version attributes the agency in 70 percent, the online version acknowledges the agency in 95% of the agency-based content. The print and online editions of the free daily *Metro* score comparably high (respectively .94 and .93). Of all titles, nu.nl scores highest on the attribution ratio: .96. Hypothesis 3, stating that print media attribute the agency more than online media, must thus be rejected. In fact, results from an independent samples t-test confirm that the opposite is the case: Overall, the online titles perform better on the attribution ratio than the print titles (M diff = .19, SD = .02, t(2184) = 10.32, p = .00 (equal variance assumed, Levene's test statistic p = .20).

Conclusion

Informed by a unique dataset, this study has provided insight into the impact of the news agency on the shaping of the daily news. Analyses demonstrate that the agency provides input for almost a quarter of print news. This percentage is strongly influenced by the high reliance on agency copy of the free outlet however: For the quality and popular newspapers, the ratio is respectively a factor two and three lower as the ratio of the free outlet. It indicates that the overwhelming majority of news in the paid newspapers does not contain (substantial) agency copy. The churnalism score of the print newspaper follows the expected pattern: lowest for the quality, highest for the free outlet. Regarding the reliance of online news on agency copy, we see a very different picture than for the print sector: 68 % of the online news comes from the desk of the agency. In the case of the primary online news provider, nu.nl, this is even higher: three out of every four articles is derived from agency copy. Keeping in mind that almost three out of every five agency articles ends up as online news, it can be argued that the agency's agenda de facto ***is*** the online news agenda. Furthermore, the online content shows very high overlap with the agency copy, as indicated by the churnalism index: the mean scores indicate that in general, no value in the sense of new information (for instance from an alternative source) is added. While the distinction between online and print news can be somewhat diffuse – some news organizations have one newsroom to fill both print and online content – the findings presented here do suggest that the two domains in general rely on different routines to create their product: online news is strongly dependent on the agency. Earlier expressed fears that "copy & paste becomes the basic principle" for online news (Quandt, 2008: 729) thus seem reality. While some minor editing may have taken place, the height of the churnalism-index indicates a general absence of original reporting. The results furthermore suggest a distinct news routine of the online-only news provider nu.nl: it is largely a serving-hatch of agency content, yet invests relatively large effort into the editing of this content. Journalists thus appear to perform a function that is decidedly different from a traditional journalist.

Regarding attributing practices, it is worth noting that the online news providers are not hiding that they are on the agency's drip: apart from the popular, all online outlets attribute consistently. For the print titles, the free newspaper *Metro* performs better than the quality title *de Volkskrant.* The entire absence of attribution by the country's largest popular, *De Telegraaf*, on the other hand, is disconcerting for those who value what has been labeled "a new norm within journalism": transparency (Karlsson, 2011, p. 279). While transparency can be achieved in various ways (Karlsson, 2010), an obvious and little demanding step for a newspaper would be to be transparent on agency reliance.

The novel measurements presented here warrant some additional consideration. One point to take into account is the threshold chosen to determine whether a link exists between an agency text and a news article. The threshold is carefully decided upon after manual analysis of hundreds of links. To avoid including spurious links – which could occur when an agency text and a newspaper article both use the same paragraph or quote from a press release or statement – it is deliberately set at a moderately high value. While this gives the assurance that all detected links are indeed indicative of newspapers' usage of agency copy rather than coincidence, a consequence is that articles that are to a lesser degree relying on agency copy are not included in the agenda setting ratio. The ratio thus gives insight into the proportion of news articles that are *largely* based on agency copy, rather than the full extent to which newspapers rely on agency copy. In other words, the numbers presented here are conservative estimates: the actual impact of the agency is even higher if we take weaker links into account as well. With this mind, the already high ratios for the online titles are even more striking. The churnalism index scores on the other hand are not to be misunderstood as representing the media's *overall* churnalism score: Just as is the case for the agenda setting ratio, it are the scores of the links as determined by the threshold. Concerning the Levenshtein measure that forms the basis of our churnalism index score, it is also worth noting that it informs on the differences between two texts, not on the nature of that difference: it can point to a reordering, rewriting or shortening of the original information. As is the case with all automated content analysis methods then, close reading of the actual data remains of great importance. Overall, the measures have proved useful indicators and facilitate systematic comparative research on both agenda setting and churnalism. The ability to analyze large volumes of data allows for cross national as well as longitudinal research. Furthermore, the proposed churnalism measure allows for a systematic assessment of the degree to which subsidized content from different suppliers makes it into the news.

Taken together, the findings confirm earlier claims of the agencies' determining impact on the news and make a strong case for a systematic scrutiny of their role. The approach presented here can foster such research. While the unavailability of the print data in digital archives may pose a problem for research in certain countries, retrieving news in real time is well possible, both online as well as print. To convert print to digital data would admittedly require some effort, but can hardly be considered a legitimate reason to continue ignoring the possibly most dominant intermedia agenda setter. Not only researchers interested in agenda setting dynamics should be alerted: The fact that the most important online news providers largely present news that comes from

one and the same pool is problematic for anyone concerned with news diversity. That the output of news agencies is typically not the result of investigative journalism, but rather reproduced input from particularly routine sources (Boyer, 2011; Davies, 2008; Phillips, 2010), should be all the more reason for thorough follow-up research. This research could advance in at least two directions: a) towards a better insight in the news production process of the agencies, including the quality of their output; and b) towards a better insight in the journalistic routines with respect to agency copy. With respect to the latter, different approaches could assess how selection processes take place and to what extent journalists fact check agency information. Concerning the first, sourcing and verification practices are a particularly important area of interest. The picture sketched by the limited literature on agencies – a continuous information overload, fast working pace and high output volumes – gives rise to concerns on an uncritical stance and an overreliance on elite sources, among others. A conclusion that be drawn from this study is that it is due time for a rigorous assessment of these concerns.

ACKNOWLEDGEMENT

This work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

CHAPTER TWO

SUBSIDIZING THE NEWS?

Organizational Press Releases' Influence on the Agenda and Content of News Media

Abstract

While the relation between PR-material and newspaper content has generated considerable attention, longitudinal evidence to substantiate claims of increasing reliance is scarce. Furthermore, the reliance of the news agency – a dominant player in the news production process - on PR-material is remarkably understudied. Applying an automated approach, this study assesses the impact of 4,455 organizational press releases over a period of ten years. It furthermore clarifies on the role of two important contextual factors: the type of source (NGO or corporation) and the type of newspaper (quality, popular and free). Two indications of source reliance are distinguished: the extent to which news articles are initiated by a press release, and the extent to which the literal press release content is reproduced in the news article. Findings indicate that one in every ten news articles included in this study is initiated by a press release; for the agency this is slightly higher (16 %). A routine of 'churnalism' – literal copy-pasting of press releases – has been found for neither the agency nor the newspapers. The content of NGOs is more often followed up on and to a greater extent reproduced in news articles than corporate content is. Overall, the findings of a moderate reliance on press releases and the fact that this reliance remains stable over time call for a more nuanced perspective on journalists' dependency on press releases.

Contributing author: J.W. Boumans

Of vital importance for democracies is a healthy and uncontrolled flow of information and ideas upon which the public can make informed choices. News media play an important part in the circulation of such information. While different perspectives on the societal responsibility of the press do exist, there is a general consensus that at the very least, news journalists have the task of actively gathering information from various sources, packaging this information into news and communicating it to the public (Manning, 2001). Recently, concerns have been raised that this news production process is drastically changing: rather than producing information, journalists are increasingly consuming and processing public relations (PR) material provided by sources (Jackson & Moloney, 2015). The fear that news media content is becoming dominated by sources that do not necessarily serve the public interest, is shared by both professional as well academic observers (Davis, 2000; Lewis et al., 2008; Prenger et al., 2011). More reliance on PR generally implies less journalistic independence, less initiative and less rigorous journalistic efforts (Reich, 2010). Not all sources are equally successful in accessing the news: Particularly governmental institutions and large corporations have the resources and expertise to provide attractive content for news media. As such, they have relatively more access to the news than sources that lack those means (Manning, 2001). The ultimate consequence of journalists routinely accepting PR material is that the powerful or wealthy at least to some degree can dominate the flow of information which is so vital for a healthy democracy (Moloney, 2006; Sissons, 2012).

Increased influence of sources on news media is often linked to developments in the media market: Decreasing newsroom capacity, faster (online) news cycles, high levels of competition, declining readership and falling advertising revenues are just some of the challenges that newspapers face these days (Manning, 2001; Lewis et al., 2008). The role of investors and the capital market in general has put further pressure on the need for media to be profitable (McChesney, 2008). These pressures are believed to lead to an unhealthy strong dependency on content subsidiaries. The professionalization of public relations accelerates this trend. Sources are increasingly sophisticated in meeting the media's demand for information, for instance by preformulating a press release according to journalistic conventions (Jacobs, 1999). It has led to the emergence of "public relations news": news that contains basically unchanged PR information, does not cite the source, and serves particular actor interests (Erjavec, 2005, p. 156).

Several studies from the field of pragmatics demonstrate the substantial influence of press releases on news coverage (Sissons, 2012; Ervajec, 2005; Maat, 2007, 2008). An inevitable drawback of these thorough, in-depth analyses is that they typically rely on either case studies or very small sample sizes, which makes it hard to draw generalizable conclusions. Journalism scholars deliver additional evidence, generally by means of a content-analytical (Lewis et al. 2008; Hijmans et al., 2011) or interview approach (MacManara, 2014; Reich, 2006, 2010). Yet, an empirical measure to systematically determine the presence and influence of subsidized content in the news on a large scale and over time is still lacking. Moreover, an important institute in the news production process is structurally overlooked: the news agency. The scarce research that is dedicated to agencies demonstrates they have a large impact on the news (Scholten & Ruigrok, 2009; Hijmans et al., 2011, Lewis et al., 2008). Concerning their sourcing practices however, only anecdotal evidence has been published (Boyer, 2011). This article therefore not only investigates newspapers' source reliance, but also that of the news agencies. Our general research question is *to what extent do organizational press releases influence the agenda and content of news agencies and newspapers?*

Media's reliance on source material will be explained from a political economy framework. The impact of sources can be assessed on several levels of the news production process. This article introduces an innovative automated tool to investigate two related levels: first, at the level of accessing the media (***agenda building***) and second, on the content level (***churnalism***). Agenda building refers to sources' attempts to influence the media agenda (and in turn public opinion) by obtaining media placement of their subsidies (Curtin, 1999). Churnalism refers to the criticized journalistic practice of recycling existent (PR) content (Davies, 2008). Over a research period of ten years, we investigate the influence of press releases on these two stages of news production and whether this influence has increased. The study thereby adds an important longitudinal perspective to the extant research on agenda building and churnalism.

The article furthermore clarifies on the role of two important contextual factors of the media-source relationship: the type of source and the type of newspaper. Aside from official (governmental) actors, a distinction often made in literature is between corporate and nongovernmental actors (NGOs) (Manning, 2001). While traditionally, corporate actors have been found more successful in accessing the media agenda, recent studies suggest that NGOs are increasingly able to gain access as well (Fenton,

2010; Van Leuven & Joye, 2014). Just like distinguishing between types of source offers a more refined insight, so too does distinguishing between types of newspaper. In general one can expect a relation between newsroom capacity and reliance on subsidized content (Davies, 2008). Both contextual factors will be discussed in detail below. Again, this study is the first to systematically compare the effects of agenda building efforts from different source types on different newspaper types over a longer time period.

NEWS MEDIA'S RELIANCE ON SUBSIDIZED CONTENT FROM A POLITICAL ECONOMY PERSPECTIVE

From a political economy perspective, the changing relation between sources and journalists cannot be explained without looking at the wider context in which media organizations operate. According to political economists, institutional and (market) structural factors are the main determinant of journalistic quality (Manning, 2001). This context includes the strong pursuit of profit, the size of the media organization, the amount and nature of competition on the media market, the influence of advertising and the specific interests of media owners and managers (McChesney, 2003). Central in McManus' (1994) 'market-driven journalism' model is the observation that economic rationalism is replacing social responsibility as the reasoning underlying the news production process. Consequently, the 'commercial underpinnings' of the news industry affect the content the media produce (Johnston & Forde, 2011) and some scholars believe they are driving the current crisis of journalistic quality (Bergman, 2014; McChesney, 2008). In a cynical response to the ongoing cutting of newsrooms and investigative journalism, McChesney (2008) observes that "doing journalism is bad for the bottom line" (p. 124). In short, business norms are believed to prevail over journalism norms. Because 'passive discovery' of news through news subsidiaries is the most cost-efficient practice, sources are able to gain power over the news production process. Extensive reliance on news subsidiaries has been documented in various countries, among which the US (McManus, 1994; Curtin, 1999; McChesney, 2003), the UK (Jackson & Moloney, 2015; Lewis et al., 2008), Slovenia (Erjavec, 2005) and Israel (Reich, 2010). Exceptions to the trend have also been reported however: a longitudinal content analysis of Belgian newspaper coverage from 1995-2010 did not show any signs of increased source reliance that might be explained by cost-cutting (Van Leuven et al., 2014). An explanation for the countries' differences may be found in the level of competition on the news media market, which is less fierce in Belgium. Additionally, the Belgian media market has a strong public service ideology in journalism.

This study is situated in the Dutch context, which has a similar media system as Belgium in terms of the level of competition and of a journalistic ethos that carries the print marks of public service broadcasting values (Brants & Van Praag, 2006). Yet, as is the case in many countries, the Dutch news media market – and in particular the print sector - faces challenging times. An advisory commission on the future of journalism in the Netherlands (TCITP, 2009) reports steadily declining readership numbers and advertising revenues since the early 2000s. These developments, combined with cutbacks in the news organizations to promote efficiency, have "inflicted deep wounds on the print sector" (p. 5). Consequently, the commission calls upon the government to safeguard the quality of the journalistic infrastructure (p. 8). The position of the news agency is in many respects comparable to that of the newspapers (TNO, 2011). While the number of employees is structurally reducing the past years, the output of the national Dutch agency *ANP* steadily increases (TCITP, 2009). If indeed reliance on subsidized content is partly explained by journalistic resources, an increased reliance may be expected for newspapers as well as the agency. In the next two sections we will describe two levels at which this source reliance manifests itself: agenda building and churnalism.

AGENDA BUILDING

News coverage is the outcome of an ongoing negotiation between sources and journalists. Agenda-building examines how outside forces influence media coverage (Curtin, 1999). Sources subsidize the media with material in an attempt to promote an organization's image and reputation. Several studies have demonstrated that these attempts are generally successful (Sissons, 2012; Ervajec, 2005; Pander Maat, 2008). Reich (2010) shows that only 40 percent of the Israeli news items involve no direct PR-input. Similarly, a content analysis of the UK quality press demonstrates that nearly one in five newspaper stories are verifiably derived ***mainly*** or ***wholly*** from PR material (Lewis et al., 2008, p. 7). Reported numbers for the Netherlands as well as Belgium are considerably lower: around one in ten of the newspaper stories are traceable to subsidized content (Hijmans et al., 2011; Van Leuven et al., 2014). These latter findings should be treated with caution, however, since the studies only measured explicit references to subsidized content in the news output. Given the fact that journalists generally prefer to veil their contacts (Reich, 2010), the actual number is likely to be much higher. By measuring the actual degree of similarity between press releases and media articles rather than relying on explicit source references, the "smokescreen of anonymity" (Reich, 2010, p. 811) does not pose an obstacle for this study.

While the focal point can be various types of objects, most of the research on agenda-building processes has focused on either political actors (Kiousis et al., 2014; Ragas & Kiousis, 2010; Roberts & McCombs, 1994) or corporate actors (Kim, Kiousis, & Xiang, 2015; Kiousis, Popescu, & Mitrook, 2007). Backed by substantial resource advantages, these established sources — particularly political actors, government institutions, and large companies – generally enjoy privileged access to the media compared to, for instance, nongovernmental organizations (NGOs) (Carlson, 2009; Cottle, 2000; Manning, 2001; McChesney, 2008). In the UK news, for instance, corporations are nearly four times as likely as NGOs or pressure groups to have their material included into news stories (Lewis et al., 2008, p. 12). It has been argued however that the changing news ecology provides opportunities for NGOs to gain ground in the struggle over news access (Castells, 2008). The diminishing of foreign news coverage has created a news hole that NGOs fill by positioning themselves as expert news sources, providing background information and reliable eyewitness accounts (Van Leuven & Joye, 2014). Particularly the well-resourced international NGOs, for whom the media have become "the battleground for their campaigns" (Castells, 2008: 85), are increasingly sophisticated in supplying the media with this type of information (Manning, 2001). Recent research suggests that NGOs are utilizing the same type of PR practices as corporations (Greenberg, Knight, & Westersund, 2011). Compared to corporations and political parties, (international) NGOs furthermore have the advantage of enjoying greater popularity and legitimacy among both journalists as well as the public (Castells, 2008), which makes them more likely to be viewed as a suitable source.

While the agenda-building processes initiated by many different types of sources have been studied, rarely has this been done in a comparative and integrated fashion over a longer time period. The lone exception of an empirical study that explicitly makes a distinction between corporate and NGO sources (Lewis et al., 2008) is based on a sample of news items from two weeks in 2006, which means no inferences can be made regarding a trend over time. This article contributes by making an explicit distinction between corporations' and NGOs' agenda building capacities over a period of ten years. Given the recent mixed findings on which source type is most successful in terms of influencing the media agenda, we refrain from formulating a hypothesis. Instead, potential differences between the two categories are assessed by means of an open research question:

RQ1: To what extent is newspaper coverage initiated by press releases from corporations and NGOs, and (how) has this degree of source-initiated coverage changed over time? (agenda building)

CHURNALISM

The allegedly increasing influence of sources on news production is not limited to accessing the media agenda, but is also visible in the literal news content. Being consistently forced to increase output without a corresponding increase in resources, journalists nowadays rely more and more on information subsidies provided by news agencies and PR practitioners to fill the newshole (Davies, 2008; Jackson & Moloney, 2015; Lewis et al., 2008). Sources are increasingly employing professional communicators to optimally supply the media with these subsidies (Davis, 2000; Prenger et al., 2011). The subsidies typically come in the shape of a media release: 'ready-made' material that is written in a journalistic writing style and meets journalistic standards and practices such as news values, accuracy, and timeliness. The combination of reduced journalistic capacities and PR-practitioners that are increasingly adapted to journalism has led to a situation where news outlets often publish PR material almost or completely unchanged (Erjavec, 2005; Sissons, 2012). Other literature reports that journalists do perform 'neutralizing operations' to reduce or eliminate the bias in press releases, or add information to make the news report more attractive to readers. Yet also in these instances, the dominant source message is often maintained (Pander Maat & de Jong, 2012). The growing impact of churnalism does not only worry media professionals and academics: An interview study with UK PR-practitioners showed that a majority of them are either professionally or personally concerned about their influence on the news (Jackson & Moloney, 2015). Practitioners recognize that 'unfiltered PR is a very powerful manipulation' (p. 11). To assess whether there is reason for concern in the Dutch context, the second research question is:

RQ2: To what extent do newspapers reproduce press releases from corporations and NGOs, and is there a trend between 2004—2013?

DIFFERENT NEWSPAPERS, DIFFERENT PRACTICES?

It would be a simplification to treat the newspaper sector as a homogenous group: inevitably, there will be differences in sourcing practices due to for instance newspaper size and organizational contexts. This study therefore distinguishes three types of newspapers: quality, popular, and free newspapers. Research confirms that newsrooms with greater capacities have more opportunities to create unique content, and are less receptive to subsidized content (Lewis et al., 2008). Free newspapers for instance have considerably less journalistic resources than quality newspapers; the latter typically employ at least ten times as many journalists (Bakker, 2002). Apart

from newsroom capacities, professional values may vary as well between newspaper types. Quality newspapers are often considered to hold the journalistic standards and ethics in the highest regards. For instance, a survey among Danish journalists found that the perceived importance of objectivity is substantially lower for popular (popular) journalists than for journalists working for quality outlets (Skovsgaard, 2014, p. 209). In this context it is likely that extensive reliance on subsidized content is considered more problematic for quality journalists than for popular journalists. In sum, we expect quality newspapers to rely least heavily on subsidized content, followed by popular newspapers and free newspapers. This varying reliance is expected to manifest itself in both the agenda building as well as the churnalism dimension:

H1: The content of free newspapers and popular newspapers is to a greater extent initiated by organizational press releases than the content of quality newspapers (agenda building)

H2: The content of free newspapers and popular newspapers is more similar to organizational press releases than the content of quality newspapers (churnalism)

UNDER THE RADAR: NEWS AGENCIES

Much for the same reasons as the alleged increased source reliance – most notably cost reduction - news organizations also strongly rely on agency copy (Hijmans et al., 2011; Scholten & Ruigrok, 2009). Surprisingly though, there has been little academic attention in news agencies' functioning, earning them the title of 'silent partners' of news organizations (Johnston & Forde, 2011). Traditionally, they occupy a unique position: as no other party, they monitor newsworthy events all over the world and generate news from a wide variety of sources. To illustrate, a journalist at the German department of news agency AP receives about 4,000 to 5,000 press releases that strive for agency coverage per day (Boyer, 2011). News agencies have an important validation function, a message gains in status when agencies report on them (Vermaas & Fransen, 2009). The perceived prominence of a source is a key filtering principle for agencies: the more prominent the politician, public figure, organization or expert is, the more likely their statements would be turned into agency releases (Boyer, 2011; Livingston & Bennett, 2003; Reich, 2011). This tendency to rely on pre-packaged information provided by credible sources has increased recently (Boyer, 2011; Carlson, 2009; Forde & Johnston, 2013).

A series of developments have led to challenging times for news agencies. First, the internet has made news omnipresent and accessible in a split second. The agencies have thus lost their traditional monopoly on 'raw' news and find themselves faced with multiple parties that operate on their market. In the digital age it is increasingly difficult to secure exclusivity of the content: as soon as a medium publishes it online, other media are able to profit for free. Second, as a consequence of the media organization's cost orientation, the agencies are under increased pressure to deliver more services at lower prices. Facing a highly concentrated newspaper market, the position of the Dutch news agencies vis-à-vis their customers is particularly disadvantageous. The news organizations are large enough to consider ***insourcing*** the agencies' services (TNO, 2011, p. 4) and thus making the agency in essence obsolete. The agencies respond to the pressures by cutting substantively in personnel and the pioneering of new markets (Manning, 2008), including offering their expertise to organizations in writing successful press releases, as the leading national agency *ANP* (TNO, 2011). Indeed, the recent developments have led the president of *ANP* to express his concern that the agency has reached a "critical borderline of providing the desired quality" (TNO, 2011, p. 25). Given the central as well as precarious position of the news agency in the construction of news, investigating the agency's reliance on sources is of vital importance. The same type of research questions that have been formulated for the newspapers will thus also be formulated for the news agency:

RQ3: To what extent is news agency coverage initiated by press releases from corporations and NGOs, and is there a trend between 2004—2013?

RQ4: To what extent does the news agency reproduce press releases from corporations and NGOs, and is there a trend between 2004—2013?

Literature on churnalism often describes an increased reliance of newspapers on agency copy, yet longitudinal data that can give support to this claim is typically not provided or at best outdated: Empirical research on the intermedia agenda setting capacity of news agencies mainly dates from the early days of agenda setting research (Gold & Simmons, 1965; Whitney & Becker, 1982; White, 1950). Given the need for more recent empirical insight on the actual impact of agency copy, our final research question will thus be:

RQ5: To what extent is newspaper content based on agency copy, and (how) has this degree of agency-initiated coverage changed over time?

DATASET AND METHODOLOGY

While input from sources can take various shapes and forms (for instance public speeches, press conferences, or background briefings for the press), the press release is still regarded as a key instrument to inform the press (Forde and Johnston, 2013; Van Leuven et al., 2014). Assessment of the influence of press releases on the media's agenda and content is based on automated quantitative analyses. In the following section we first briefly describe the dataset, after which we explain the mechanisms behind the analytical tool.

OVERVIEW DATASET

Sources. The selection of corporations to include was based on three criteria: they are among the largest companies in the Netherlands, their press releases are written in Dutch, and they need to represent different types of industry. Specifically, these are electronica (Apple), finance (insurance company Aegon), energy (Nuon), and consumer goods (Ahold). Considering the additional criterion of availability of press releases, the final selection can be considered a convenience sample. The selection of nongovernmental organizations was based on equal criteria: they are all among the largest NGOs in terms of budget, they need to cover a range of different missions, and their press release archive must be available. The organizations either focus on humanitarian aid (Artsen zonder Grenzen; Unicef; Vluchtelingenwerk), the environment (Greenpeace), wildlife (WWF) or health (KWF). The press releases are downloaded through the websites of the organizations. In two cases the online archive did not cover the entire time period; these data were obtained after personal contact with the organization. Appendix A provides a specific overview of the number of releases per organization. In total, 2518 NGO press releases and 1937 corporate press releases are included.

News media. The media data (newspapers and agency combined) consist of all articles published between 2004 – 2013 in which one of the above organizations is mentioned at least once. We selected the oldest and largest of the two Dutch national news agencies, Algemeen Nederlands Persbureau (*ANP*). The newspaper selection covers the spectrum of the print media landscape: *NRC*, *Volkskrant* and *Financieel Dagblad* are the largest quality newspapers, *De Telegraaf* and *Algemeen Dagblad* (AD) are the largest popular newspapers and *Metro* and *Spits* are the two national free newspapers. The articles are obtained through the LexisNexis database, using a search string that includes all names of the organizations. Editorials, sports news, and letters to the editor

are excluded from the dataset. In total, 6142 agency releases and 22928 newspaper articles are included. Appendix A provides a specific overview of the number of releases per newspaper.

Data preparation and measurement. After obtaining the data, the articles are categorized in separate folders per source, agency and newspaper. Next, a metafile is constructed that includes the following elementary information of every individual file: the name of the article, the date of publication, the type of domain (source/agency/newspaper), and the name of the organization or medium. For every media article, the customized software then evaluates whether (agenda building) and to what degree (churnalism) that media article is based on a press release. To be sure that the media article is initiated by the press release and not the other way around, a criterion is that the press release must precede the media article in date, with a maximum of three days. The following two sections will describe the agenda building and churnalism measures in detail.

Agenda building ratio. This ratio will be used to answer *RQ1*, *H1*, and *RQ*3. In line with Reich (2010) we operationalize the agenda building capacity as the percentage of media items about an organization that is initiated by a press release from that organization. It is thus an indication of the extent to which press releases set the media agenda. Determining whether there is a link between two articles is based on a measure called cosine distance. This measure indicates how similar two documents are likely to be in terms of their subject matter (Tan et al., 2006), or in other words: to what extent they share the same terms. A cosine score can take any value between [0,1], depending on the degree of similar terms in the two texts. A score of zero implies that two documents do not share any terms, while a score of 1 implies that the terms of in the two documents are identical. Appendix B provides an example of three texts and their mutual cosine scores. A more elaborate explanation of the cosine measure is found in the method section of Chapter Three. A systematic manual analysis of a subset of the data showed that a similarity score of .33 and above indicates that two articles discuss the same subject matter. This value therefore serves as threshold to determine whether or not there is a link between a media article and a press release. When the cosine score is .33 or higher, the software reports the existence of a link. With this threshold as criterion, every news article can thus either be initiated by a press release, or not. The formula of the ratio is as follows:

$$\textit{Agenda building ratio} = n\_i / n\_t \quad (1)$$

where *n*_i is the number of media articles that is initiated by a press release from an organization and *n*_t is the total number of media articles related to that organization. The following hypothetical example illustrates the agenda building ratio. Let us presume that in a given period, the newspaper *De Telegraaf* publishes 14 articles about Greenpeace (*n*_t = 14). The software tool finds that seven of those newspaper articles share content with a press release of Greenpeace that have been published in the three days before the newspaper article (*n*_i = 7). The agenda building ratio is thus 7/14 = .50, indicating that 50 percent of the *De Telegraaf*'s coverage on Greenpeace is initiated by (based on) subsidized content from that actor.

Churnalism Index. The Churnalism Index is our measure to answer *RQ2*, *H2* and *RQ4*. The Churnalism Index informs on the degree of similarity between a press release and a media text. At the core of this index is Levenshtein distance (lev), a well-established measure in computer science and information theory that is among others often used in plagiarism detection tools. The Levenshtein distance measures the difference between two sequences. Commonly applied to compare words, lev is the minimum number of edits that are required to change one word into the other. An edit can be an insertion, a deletion or a substitution. In this case, the unit of analysis is the article instead of a word. We are interested in the degree to which media content consists of subsidized content. Therefore, the measure needs to control for difference in length: Deletion is not considered a journalistic effort, as for instance adding information is. Furthermore, we want to analyze the relative effort a journalist has put into a media text. Twenty alterations of a source text that consists of only 30 words is not the same as 20 alterations based on a 500 words text. We therefore also control for the length of the journalistic text. Formally then, the formula for the Churnalism Index is the following:

$$\textit{Churnalism Index} = 1 - (\textit{leva,b} - (\textit{LIa} - \textit{LIb}))/\textit{LIb} \quad (2)$$

where $\mathrm{lev}_{a,b}$ is the Levenshtein distance between source text a and media text b, LI_a is the length indicator of the source text and LI_b is the length indicator of the media text. By definition, the value of the Churnalism Index ranges from 0 to 1. The measure is inversed to facilitate interpretation: the higher the score, the higher the overlap (and thus the higher the degree of churnalism). A systematic manual analysis of a subset of the data showed that when the Churnalism Index > .7, the two texts are nearly identical, whereas an Index score of 0 indicates that while there is some resemblance in terms of topic and word usage, the media text differs substantially from the source text.

Intermedia agenda setting ratio. This final measure is used to determine the share of news articles that is based on agency copy and follows the same logic as the agenda building ratio described above to assess the media's relation with the sources. The only modification is that the threshold is set at a higher level, namely at a cosine score of .65. The reason for this is that the goal of the intermedia agenda setting measure is to indicate whether a news articles consists predominantly of agency copy. The threshold of .65 certifies that the majority of the text is identical to the agency text. This is a different aim than the agenda building measure has, namely to detect whether a press release and an agency or news article are related.

ANALYSIS

The six research questions and hypotheses either focus on differences between source and media categories or trends over time. To assess differences between groups, analysis of variance (ANOVA) is the appropriate technique. To assess trends over time, regression analysis is most suitable. Additionally, the ratios will be plotted visually over time to get a good overview of the trends. Because we are interested in possible differences between source types (*RQ1*) as well as types of newspapers (*H1*) on the agenda building ratio, a factorial ANOVA is applied to determine the effects of both variables. *RQ3* is equivalent to *RQ1* but related to the agenda building ratio of the news agency. To place this ratio in perspective to the newspapers, the ratios of the news agency and of the newspapers will be included in one factorial ANOVA. The second part of *RQ1* deals with a possible trend in the agenda building ratio over time. To assess this, regression analyses are conducted. For this purpose, the number of links has been aggregated to quarters to guarantee sufficiently large cell sizes. Again, the agency scores will be plotted in the same figure as the newspapers.

The research questions and hypothesis on churnalism (*RQ2*, *H1* and *RQ4*) are treated in the same way as the questions on agenda building. Thus, factorial ANOVAs for group differences and regression analyses for trends over time will be applied. Just as is the case for the agenda building ratio, the agency will be included in the same ANOVA and regression plot for the Churnalism Index as well.

RESULTS

AGENDA BUILDING RATIO

The first part of the analysis focuses on the degree to which press releases are being picked up by the media and specifies between type of source (NGO/corporate) as well as type of media agency/popular/quality/free). The first research question we address concerns the extent to which newspaper coverage is initiated by organizational press releases.

The mean agenda building ratio for newspapers and sources is .096 (*SD* = .087), indicating that overall, about 10 percent of the newspaper texts are initiated by a source release. To assess possible differences between source types and media types, Figure 1 gives a graphical insight in the agenda building ratio per type of source and type of medium.

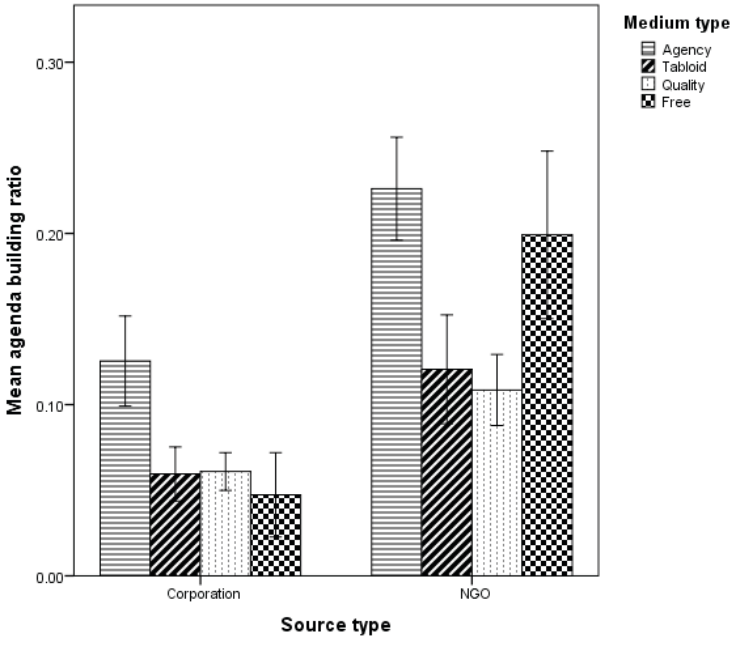


Figure 1. Mean agenda building ratio categorized by source type and medium type.
Note. Error bars depict 95% confidence interval.

We expect to find statistically significant differences in the agenda building ratios between the different newspaper categories (*H1*). Specifically, we expect that the coverage of free newspapers and popular newspapers is more strongly initiated by press releases (and thus score higher on the agenda building ratio). Results of the ANOVA indicate that the interaction between source type and media type is significant, $F(3,252) = 5.898$, $p < .01$, partial $\eta^2 = .066$. This implies that the effect of source type on the agenda building ratio differs per medium type. The assumption of homogeneity of variance is violated (as assessed by Levene's Test of Homogeneity of Variance ($p = .000$)). The Welch *F* test is considered robust for violations of the homogeneity of variance assumption, thus separate one-way ANOVAs for the two source categories in combination with Welch's *F* test are performed.

With respect to the corporate press releases, the agenda building ratios for the media categories are statistically significantly different, Welch's $F(3,131) = 16.734$, $p = .000$. A Games-Howell post-hoc test reveals that the ratio of the agency ($M = .126$, $n = 34$, $SD = .075$) is statistically significantly higher than the ratio of all three newspaper types: quality ($M = .061$, $n = 34$, $SD = .032$), popular ($M = .060$, $n = 34$, $SD = .045$) and free newspapers ($M = .047$, $n = 30$, $SD = .066$). Between the three newspaper types, the mean differences are not statistically significant different. *H1* is thus refuted.

With respect to the NGO press releases, the agenda building ratios for the media categories are statistically significant different as well, Welch's $F(3,63) = 16.483$, $p = .000$. A Games-Howell post-hoc test reveals that the agenda building ratios of the agency ($M = .226$, $n = 34$, $SD = .086$) and of the free newspapers ($M = .200$, $n = 26$, $SD = .121$) are statistically significantly higher than the ratios of the quality ($M = .109$, $n = 34$, $SD = .059$) and the popular newspapers ($M = .121$, $n = 34$, $SD = .091$). Differences between the agency and the free newspapers are not statistically significant, and neither is this the case for the combination quality/popular.

On the basis of these results, we can reject *H1:* In most instances, the content of free newspapers and popular newspapers is not to a higher extent initiated by press releases than the content of quality newspapers. Only in the case of NGO content, the coverage of free newspapers is statistically significantly more initiated by source content than is the case for quality and popular newspapers. We can now also formulate a partial answer on the extent to which news agency content is initiated by subsidized content. Figure 1 shows that respectively 12 percent of the agency's corporate coverage and 22 percent of their NGO coverage is initiated by a press release.

For newspapers, this lays statistically significantly lower: between respectively five to six percent and eleven to twenty percent. The next section will consider the second part of *RQ1* and *RQ3*: how the agenda building ratio evolves over time.

AGENDA BUILDING RATIO OVER TIME

To visually assess whether overall the agenda building ratios of the newspapers and the agency have increased (*RQ1* and *RQ3*), the ratios have been plotted over time (see Figure 2).

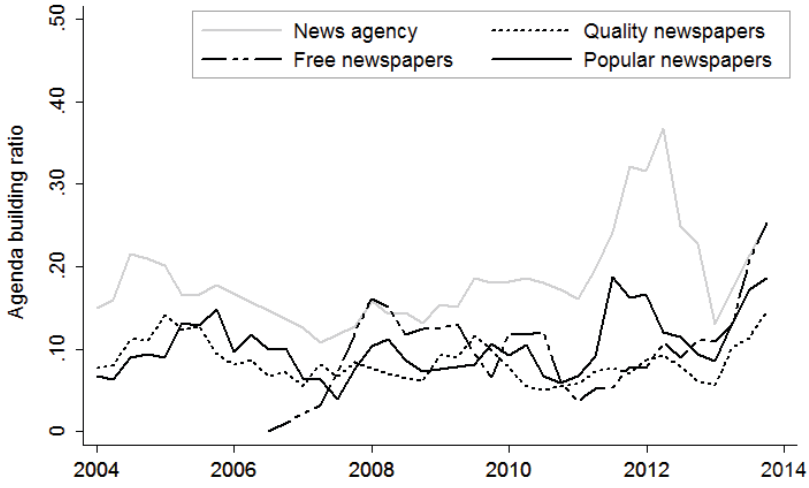


Figure 2. Agenda building ratio 2004 - 2013 per media type.

On the basis of Figure 2 it becomes clear that the ratio of all categories fluctuates over time, but that trends are not immediately evident. To offer a more refined insight and formulate and a statistically-based answer, regression analyses were conducted for each combination of source type and media type separately. Table 2 depicts the results.

Table 2. Agenda building ratio per media / source relationship over time.

Relation	Beta	*t*	p-value	equation
Agency / Corporation	.061	.663	.509	.059 + .000x
Agency / NGO	.063	.962	.337	.168 - .001x
Quality / Corporation	-.009	1.248	.863	.047 - .000x
Quality / NGO	-.086	-2.00	.046**	.157 - .002x
Popular / Corporation	.084	1.248	.213	.020 + .000x
Popular / NGO	.005	.090	.928	.125 + .000x
Free / Corporation	-.099	-1.117	.266	.078 - .001x
Free / NGO	.039	.601	.549	.146 + .002c

Note. Statistics of linear regression model, effect of x (time in quarters of a year) on agenda building ratio. ** significant at the $p < .05$ confidence interval level.

The results indicate that time is a statistically significant predictor for the Agenda building ratio for only one of the combinations, namely quality / NGO. The negative coefficient points to a decrease of the ratio. In other words, the percentage of NGO-related articles in quality newspapers that is based on a press release decreases with -.002 per quarter of a year, or .008 percent per year. For the other combinations, no statistically significant trends over time have been demonstrated. On the basis of this information, ***RQ1*** and ***RQ3*** can now be answered in their entirety: Both news agency as well as newspaper coverage is to a moderate extent initiated by press releases (around sixteen percent for agencies and ten percent for newspapers), the ratio is significantly higher in the case of releases from NGOs, and the ratios remain overall stable over time.

CHURNALISM INDEX

The second research question concerns the degree to which source content is reproduced by newspapers (Churnalism Index). Statistics indicate that the mean Churnalism Index-score of newspaper articles related to press releases is .260 (n = 1887, SD = .171). A score of .260 implies that while parts of the media text overlap with parts of the press release, the media text has a large share of non-overlapping content. Literal copy-pasting of press releases (indicated by a Churnalism Index > .70) is rare: only three of such cases have been found in the dataset (.16 percent). Concerning ***RQ2***, we can thus infer that the differences between source texts and media texts are

generally so high that we cannot speak of mere copy and pasting practices. The score of the agency is slightly higher than for the newspapers (n = 771, M = .280, n = 8874, SD = .30), but also not of a level that indicate routine churnalism practices. The number of literal replications of press releases is equally low: one (.1 percent). Figure 3 gives a visual insight in the means of the Churnalism Index specified by source and media type.

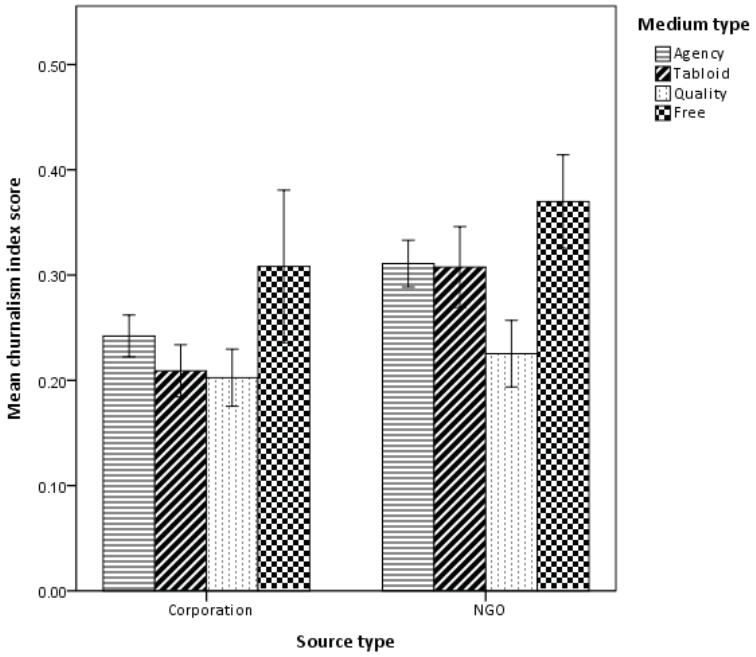


Figure 3. Means of Churnalism Index, categorized by source type and media type.

On the basis of the unequal financial and journalistic resources, we expect the quality newspapers to conduct less churnalism than popular and free newspapers (*H2*). Residual analysis was performed to test for the assumptions of the two-way ANOVA. Outliers were assessed by inspection of a boxplot, normality was assessed using Shapiro-Wilk's normality test and homogeneity of variances was assessed by Levene's test. The residuals of the combination quality newspapers / corporations showed two outliers, for the combination free newspapers / NGOs three outliers are reported. Since these outliers are not extreme and are genuinely unusual data rather than measurement errors, the data points are maintained in the dataset. The residuals were normally distributed (p > .05) for all but the two relationships mentioned above (quality/corporations: p = .045; free/NGO: p = .080) and for the combination agency / NGO (p = .063). Since the ANOVA is considered fairly robust to deviations from normality (Maxwell & Delaney, 2004), this is

not considered problematic. Finally, the Levene's test was violated, indicating there is heterogeneity of variances (p = .000). Therefore, the Welch test is the appropriate test to run.

Results of the 4x2 factorial ANOVA show that there is no statistical significant interaction between source type and media type on the index, $F(3,264) = 2.069$, $p = .105$, partial $\eta^2 = .023$. Yet, considering that the p-value is on the edge of being significant, meaningful differences between the scores per media type and source type may be expected. Therefore we decide to run one-way ANOVA's per compare the Churnalism Index-scores for media type for the corporate and the NGO category separately, rather than combined.

With respect to the press releases of corporations, the Churnalism Index-scores for the media categories are statistically significantly different, Welch's $F(3,59) = 4.353$, $p = .008$. A Games-Howell post-hoc test reveals that the Churnalism Index-score of the free newspapers ($M = .308$, $n = 21$, $SD = .159$) is statistically significantly higher than the ratios of the quality newspapers ($M = .203$, $n = 40$, $SD = .085$) as well as the popular newspapers ($M = .209$, $n = 40$, $SD = .077$). In other words, the content of free newspapers is statistically significantly more similar to corporate press releases than the content of quality and popular newspapers is. Between the categories quality and popular, no statistically significantly differences exist. Popular newspapers thus do not reproduce more subsidized content than their quality counterparts, at least with respect to corporate press releases. The score of the news agency lies between the free newspapers and the popular/popular ($M = .241$, $n = 34$, $SD = .056$), yet does not differ statistically significantly from either of the three newspaper types.

With respect to the press releases of the NGOs, the Index-scores of the media categories are statistically significantly different, Welch's $F(3,68) = 11.264$, $p = .000$. A Games-Howell post-hoc test reveals that the Index-score of the quality newspapers ($M = .225$, $n = 39$, $SD = .100$) is statistically significantly lower than the ratios of the agency ($M = .311$, $n = 34$, $SD = .064$) as well as the two other newspaper types: popular ($M = .311$, $n = 38$, $SD = .117$) and free ($M = .367$, $n = 26$, $SD = .110$). The findings indicate that the reproduction of subsidized content from NGOs is significantly higher for the agency, popular as well as free newspapers than it is for quality newspapers. The agency, popular and free newspapers do not differ statistically significantly from each other. In sum, *H2: The content of free newspapers and popular newspapers is more similar to press releases than the content of quality newspapers* can partly be

confirmed. It is true in the case of free newspapers, yet for the comparison between quality and popular newspapers it only holds in the case of NGO content.

The results above can also address the first part of *RQ4*, regarding the degree to which the agency reproduces subsidized content. Overall, the Churnalism Index-score of the agency does not differ much from the newspapers. While the agency's score is statistically significantly higher than the score of the quality newspapers, it is equal to the popular newspapers and lower than the score of the free newspapers. In other words, just as is the case for the newspapers, the agency dedicates a fair amount of effort into its content. The Index-score indicates that in general, the content of a media text differs substantially from the source text it is initiated by.

CHURNALISM OVER TIME

The second parts of *RQ2* and *RQ4* focus on the degree of churnalism over time. In Figure 4, the Churnalism Index score between 2004 - 2013 is plotted.

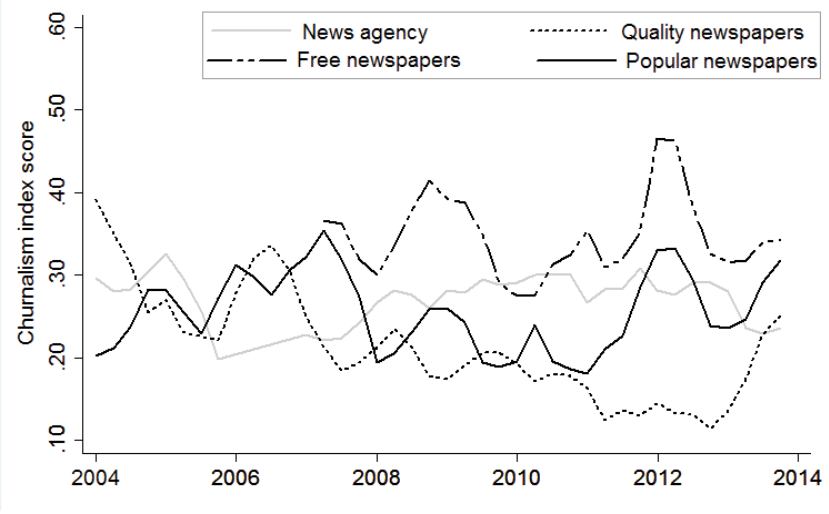


Figure 4. Churnalism Index scores 2004 - 2013 per media type.
Note. Quarterly-level aggregated data. Plotted lines represent the moving average of three time periods.

Figure 4 shows that fluctuations over time exist in the dataset. Of particular interest is the index of the quality newspapers, which shows a decrease between 2004 – 13. This indicates that churnalism has thus actually declined over the years. The trend line of the popular newspapers seems to show a slight upward movement, suggesting that churnalism has increased for this category. Table 4 shows the regression statistics.

Table 4. Degree of subsidized content per media source relationship over time.

Relation	Beta	*t*	p-value	equation
Agency / Corporation	.041	.231	.819	.238 + .000x
Agency / NGO	-.136	-.776	.443	.327 - .001x
Quality / Corporation	-.699	-6.033	.000***	.307 - .005x
Quality / NGO	-.438	-2.960	.005***	.299 - .004x
Popular / Corporation	.222	1.406	.168	.179 + .001x
Popular / NGO	-.205	-1.254	.218	.351 - .002x
Free / Corporation	.094	.411	.686	.256 + .002x
Free / NGO	-.157	-.781	.443	.432 - .002x

Note: Statistics of linear regression model, effect of x (time in quarters of a year) on Churnalism Index.

From Table 4 we can infer that for the quality newspapers, time is indeed a statistically significant predictor for the Churnalism Index. The negative coefficients point related to the corporate and NGO sources to a decrease of the churnalism score, implying that the similarity with the source texts are actually decreasing with -.005 and -.004 per quarter of a year. For the other relationships, no statistically significant differences are found. This implies that overall, time is largely unrelated to degrees of churnalism. In sum then, *RQ2* and *RQ4* can also be fully addressed: Both the news agency as well as the newspapers reproduce press releases to a very limited extent, the similarity is significantly lower in the case of releases from NGOs, and the ratios remain overall stable over time for all media except for the quality newspapers, whose churnalism scores are statistically significantly decreasing over time.

INTERMEDIA AGENDA SETTING RATIO

Our final research question concerned the question to what extent newspapers rely on agency copy for their content, and how this has evolved over time. To assess these questions Figure 5 depicts the intermedia agenda setting ratio over time per newspaper type.

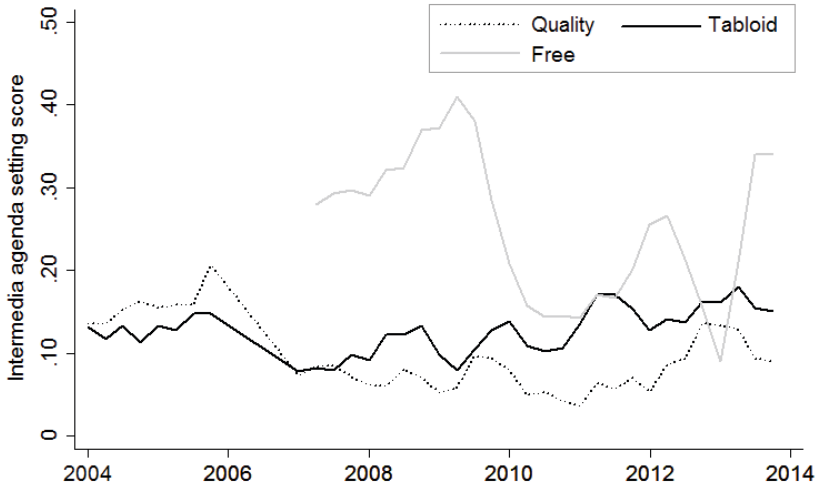


Figure 5. Intermedia agenda setting score per newspaper type, 2004-2013*.
Note. Quarterly-level aggregated data. Plotted lines represent the moving average of three time periods. Quality newspaper NRC ended the agency subscription in January 2011 and has been excluded from the analysis over the period 2011-14.

From Figure 5 we can infer that the content of the free newspapers is most strongly based on agency copy, with a mean intermedia agenda setting ratio of .26 (*SD* = .11). This score implies that 26 percent of the articles published in free newspapers, consists mainly of agency copy. For the quality and the popular newspapers, this is respectively .09 (*SD* = .05) and .13 (*SD* = .13). Simple linear regression analyses indicate no across-the-board trend of increased reliance on agency copy: A marginal statistically significant equation was found for populars ($F(1, 32) = 3.28$, $p = .08$), with an R^2 of .09. The intermedia agenda setting score for populars increases with .001 for each quarter of a year. A marginal statistically significant equation was also found for the quality newspapers, albeit in the opposite direction: ($F(1, 32) = 5.26$, $p = .03$), with an R^2 of .14. The intermedia agenda setting score decreases with .002 for each quarter of a year. No statistically significant trend has been found for the free newspapers ($F(1, 22) = 2.03$, $p = .17$).

DISCUSSION

The relation between sources and news media has been a key point of interest for journalism scholars for decennia. Recently concerns have been expressed about a "PR-isation within news media" (Jackson & Moloney, 2015, p. 2), where the PR-industry increasingly dominates journalism. Scholars arguing from a political economy perspective find an explanation in the pressure on media to produce more output at a faster pace, with fewer resources. This article has introduced two measures to empirically assess two outcomes of the alleged PR-isation of the news: stronger agenda-building capacities of sources, and increased reproduction or even verbatim use of subsidized content by the media.

The results differ from the alarming findings from studies from the UK and US. Instead, they sketch a nuanced picture for the Netherlands, with significant differences in terms of source reliance between types of newspaper. First, results of the agenda building ratio show that overall, only about one in ten news articles are initiated by a press release. For the agency, this lays around 16 percent. For neither the agency nor the newspapers there is a trend of increased agenda building capacities of sources. Second, when a media release is indeed initiated by a press release, the Churnalism index score overall indicates that the content of both the newspapers as well the agency differs substantially from this press release. This is in contrast to earlier findings on agencies' heavy reliance on and nearly verbatim use of press release material (Johnston & Forde, 2013). It is illustrative that less than one percent of the media texts is a verbatim copy of a press release. Here too, no strong positive trends over time have been found. Instead, the only statistically significant trends point to a decrease in churnalism levels. In a similar vein, the findings on the intermedia agenda setting ratio reject assumptions of an increased reliance on agency copy. While not surprisingly, free newspapers reproduce agency copy most often (about one in every four articles is de facto an agency article), there are no signs that the news agency dominates the news agenda.

Overall then, the results refute the notion that media passively process news material (Davies, 2008), as has been documented for other countries. As such, the findings are reassuring for those worrying about news being dominated by PR-influences – at least for the Dutch (print media) case. From a political economists' perspective this is an interesting finding: The economic hardship does not appear to have led to a routine of copy-pasting press releases, nor an increased reliance on agency copy. It

might be the case that while pressures have increased, journalists are adapting to this increasing speed and quantity of the news cycle. An alternative explanation for the fact that no trend has been found is that the study's time frame is too limited to capture the impact of recent mechanisms held responsible for these trends. Indeed, the influence of economic pressures on newsroom capacities – "sacrificing journalistic values to keep profit levels high" - has been signaled in the 1980's already (Curtin, 1999, p. 55). In other words, it might be that the critical point is already beyond us. Yet even when this is the case, the general results do not give an impression of a highly problematic dependency on subsidized content. The near absence of churnalism that we found is consistent with earlier research in the Dutch context (Hijmans et al., 2011) and also with findings in the Belgian context (Van Leuven et al., 2014), which in many respects has a comparable media system as the Netherlands. Arguably the most plausible explanation is that Dutch and Belgian media do not face the same pressures, or with the same intensity, as the media markets that political economists typically consider (most notably the United States and the United Kingdom). These markets are characterized by higher levels of competition and economic pressures than the Dutch market. To make more precise inferences on the relation between media system characteristics and agenda building and churnalism practices, a cross-national study that takes factors on different levels of influence into account is warranted. There is an emerging consensus that the political and economic environment are the main forces driving journalism cultures and media systems (Hallin & Mancini, 2004; Hanitzsch & Mellado, 2011). Previous studies have reported substantial cross-national variation in levels of political and economic influences experienced by journalists (Hanitzsch & Mellado, 2011), and it would be an important step forward to link journalists' perceptions of influence as well as economic indicators with the measures of agenda building and churnalism presented here.

Although there is compelling evidence that press releases are an important information subsidy, a limitation of the study is that it focuses only on this one type. Consequently, we are cautious with making generalizations of the impact of information subsidies in general on print media. Apart from other traditional types, the era of digital communication has brought along new types of subsidies and may well have led press releases to become less important. Research from the US and the UK shows that today's PR-professionals – often ex-journalists - employ increasingly sophisticated media relations practices. Content is now often tailored to different media newspapers, or even come in a whole package, including editorial suggestions, third-part commentary, and case studies. According to some observers, the days of the "monolithic media release" are largely over (Jackson & Moloney, 2015: 13). Sources

may for instance provide media with information through direct mailing, but also through blogs, Twitter, or Youtube (Kiousis et al., 2014; Parmelee, 2013). Nonetheless, various studies show that press releases and other classical news gathering channels remain central factors in news production (see Van Leuven et al., 2014). Additionally, the finding that reliance on press releases has remained stable over time does not support the thesis that the new channels have replaced the traditional press release as information source for journalists.

While the automated approach has provided valuable insights, it can be optimized. For instance, it currently does not distinguish between different parts of the media texts. Previous textual analyses have demonstrated that most of journalistic transformations in source material occur later in the text, leaving the primary message intact (Maat & de Jong, 2012). While journalists thus may alter the press release and add or contextualize information, radical transformations are rare. News reports typically are structured along an inverted pyramid, starting with the most important information. A future expansion of the tool would ideally distinguish the headline, lead and intro from the body of the text. This would refine the measure on the degree to which the meaning of a media text resembles the meaning of a source text.

One finding that is worth considering in greater depth is that overall, NGOs appear more successful in accessing the media than corporate sources. Publicity on NGOs is significantly more often initiated by subsidized content. Furthermore, their content is relatively more reflected in news coverage than corporate content. Contrasting to the traditional notion of NGOs being at a disadvantageous position compared to elite sources, the results are more in line with the findings of the Belgian news coverage study (Van Leuven & Joye, 2014) that NGOs have enhanced agenda-building capacities. An explanation for their relative success compared to corporate sources is that NGOs typically appeal to public interest (Erjavec, 2005). Consequently, it is more likely that their messages are picked up by media. The findings support more optimistic notions that the changing news ecology offers new possibilities for non-established sources to access the news arena (Castells, 2008; Greenberg, Knight & Westersund, 2011). It should be noted however that the NGOs in the current study are all relatively large and well-resourced organizations. Additional research is required to assess to what extent smaller, less resourced NGOs enjoy access to the media.

While quantitative 'input-output' analyses are all but new in the field of agenda-building, the current study has a distinct novelty: it relies on automated content analysis. The advantage of this is not to be underestimated: large-scale data analysis – for instance in a cross-country design – is now possible without a substantial budget. The statistical parameters can serve as yardsticks to assess the impact of subsidized content across countries and media markets. The approach also opens up venues for time series analyses that take into account factors on the meso level – such as media organizations' financial performance, the newsroom capacities - and the macro level - like ownership concentration and the degree of competition on the media market, advertising models, and the volume and sophistication of the PR-industry. At a time when concerns about increasingly sophisticated sources manipulating ever less equipped journalists are thriving, the proposed automated approach can offer valuable empirical insights to the discussion.

CHAPTER THREE

NUCLEAR VOICES IN THE NEWS:

A Comparison of Source, News Agency and Newspaper Content About Nuclear Energy over Time

Abstract

While news media are frequently criticized for their alleged increasing reliance on 'subsidized content' provided by sources and news agencies, this claim is seldom empirically verified. Based on insights from computer science, this study proposes an approach to quantitatively compare source, news agency and newspaper content over time. Including press releases from two corporate actors and one nongovernmental actor as well as articles of news agencies and newspapers, the approach is applied to the debate about nuclear energy in the Netherlands (2003-12). Results show no indication of an increased similarity of newspapers' content with either source content or news agency content, thus providing no justification for the concerns about an increasing dependency of newspapers on subsidized content. Contradicting literature, we found that media content is most similar to the nongovernmental organization's content, with the exception of one regional newspaper that strongly reflects the local corporation's content.

Contributing authors:
J.W. Boumans, R. Vliegenthart, & H.G. Boomgaarden

A slightly revised version of this chapter has been accepted for publication in European Journal of Communication

The current economic hardship of news media allegedly affects the quality of their performance. Decreasing newsroom capacity, faster news cycles, high levels of competition, declining and unstable readership and falling advertising revenues are just some of the challenges that newspapers face these days (Davies, 2008; Fenton, 2011; Lewis et al., 2008a). Investors and the capital market in general are placing newspapers under further pressure (McChesney, 2008). Concerns with minimizing costs and enhancing profits have led to cuts in investigative journalism and to a greater reliance on official sources, the PR-industry and news agencies in the production of news (Davies, 2008; Manning, 2013; Moloney & Jackson, 2015). Parallel with this weakening of traditional journalistic practices, critical observers have witnessed an impressive growth of the public relations industry and diverse forms of 'spin' and news management forms, who deliver this 'subsidized content' to the media (Davis, 2000; Prenger et al., 2011). These dynamics arguably have resulted in a shift of the power balance between sources and journalists in favor of the former (Davis, 2000; Franklin, 2011). This paper compares organizational press releases with news media content over a period of ten years to assess whether the growing dependency of journalists on sources has led to an increasing similarity between news and source contents.

Sources are an indispensable part of the news production process and studying these dynamics provides valuable insights into the performance of the news media. News coverage has aptly been described as "*a sampling of sources' portrayals of reality, mediated by news organizations"* (Sigal, 1986, pp.27-8). The key concern nowadays is that news coverage involves increasingly less mediation and more sampling, even to an extent that source information is literally copied (Lewis et al., 2008). This is believed to compromise media's societal duty to provide an independent flow of information and interpretation. While many journalism studies have shed light on the practices of journalists and the processes involved in the production of news, the origins of news contents - sources and their PR activities - have received less attention (Manning, 2001). Although recent research has demonstrated the pervasive role of the PR-industry in the news media (Jackson & Moloney, 2015; Lewis et al., 2008; Reich, 2010; Sissons, 2012), longitudinal empirical research into the relationship between source content and news content remains scarce. A basic assumption in this study is that if indeed sources are becoming more influential over time, news content will increasingly mirror the content of press releases.

Apart from sources, news agencies are another fundamental category in the infrastructure of news (Lewis et al., 2008). Wire content traditionally steers a large part of the daily press and research indicates that its impact is growing (Scholten and Ruigrok, 2009; TNO, 2011). At the same time, agencies are increasingly struggling to provide quality and maintain profitable at the same time (TNO, 2011; Boyer, 2011). There are a number of reasons the role of agencies deserves critical attention. First, agency material is generally taken for granted by journalists: fact checking of news agency copy is not common practice as journalists assume that the stages of fact-checking, background research and verifications have already been completed (Davies, 2011; Forde and Johnston, 2011). Yet it has been shown that agencies frequently distribute PR releases with little or no checking or verification of content involved (Forde and Johnston, 2011; Scholten and Ruigrok, 2009). The unchecked reproduction of agency content increases the chances of media disseminating false or distorted information. Second, the fact that newspapers to a large extent rely on agency copy for their coverage of an event is likely to limit the diversity of voices and perspectives in the media landscape. Although agency content in itself may very well be diverse, without doubt the overall diversity would be greater in a situation where each newspaper produces its own coverage of an event.

Despite their central role in the infrastructure of the news, news agencies have enjoyed remarkably little academic interest. Aside the cited studies above, empirical research is mostly confined to the international agencies (Boyd-Barrett and Rantanen, 2000) and longitudinal research is absent. The absence of empirical data about potential trends makes the often heard claims on increased source dependency hard to maintain. Thus, research on sourcing practices from agencies as well as journalists is in need of a longitudinal perspective.

We rely on insights from the fields of artificial intelligence and information retrieval to empirically address the concerns on media's alleged dependency on subsidized content. The central research questions are *a) to what degree does the content of the output of sources, agencies and newspapers overlap?* and *b) how does the degree of content similarity evolve over time?* The Dutch debate on nuclear energy between 2003-12 will function as a case study to address these questions. The dataset includes the contents of three domains: 1) press releases from two business sources and one nongovernmental organization, 2) articles from the two leading Dutch news agencies, and 3) articles from five national and two regional newspapers. The regional newspapers are included because concerns are expressed that their economic hardship would make

them particularly reliant on subsidized content (Franklin, 2008; O'Neill and O'Connor, 2008). Considering that regional newspapers are an important factor in the media landscape (Bakker and Scholten, 2011), a better understanding of their relation with sources is vital.

SOURCES: JOURNALISTS' GRATEFUL PARTNERS

For more than four decades now, studies of the sociology of news production have shed light on the interaction between sources, journalists and news media organizations. Based on observations of journalistic practices, Tuchman (1973) described the task of news organizations as 'routinizing the unexpected' by developing processes that ensure a steady amount of news in a timely fashion. One such routine is the systematic reliance on sources (Cottle, 2003), whose press releases are designed to support the journalist's linear, routine writing process (Van Hout & Jacobs, 2010).

With controversial societal or political issues like nuclear energy, involved parties strive for media coverage that reflects their - often conflicting - perspectives. Through coverage in line with their positions, these stakeholders aim to indirectly influence public opinion (Curtin and Rodenbaugh, 2001). Such strategic communication directed at generating media-attention is coined agenda building (Curtin, 1999) and has attracted wide scholarly attention. Previous agenda building studies have found that not all sources are equally successful in influencing the media-agenda. Typically, business and institutional sources tend to be attractive sources for journalists because they are well accessible, well resourced, and provide a regular supply of 'information subsidies' (Davis and Cottle, 2003). The perceived legitimacy of a source is also determined by the 'cultural capital' (differing competences, skills and assets) that sources possess (Anderson, 1997: 9). Being regarded newsworthy in their own right, the dominance of institutional sources has been widely demonstrated (Hall et al., 1978; Manning, 2001; Rafter, 2014; Shehata, 2010). Similarly, a study that differentiated between different source types showed that corporate PR material is nearly four times as likely to be reflected in general news stories as the press releases of NGO's or pressure groups (Lewis et al., 2008). This study concerned general domestic news however: when it comes to environmental issues, the media agenda has been found to be highly dependent on particularly government agencies and NGOs, with corporate sources being less successful in promoting their 'green agendas' (Curtin and Rodenbaugh, 2001). The specific case of nuclear energy is an environmental as much as an economic and a safety issue. While it is not the paper's core interest to assess which source type

is the primary definer, the multifaceted nature of nuclear energy makes it an interesting issue to see which type of source is most successful in promoting its agenda.

Empirical agenda-building studies usually concentrate on two major independent variables affecting media agendas: real-world conditions and events, and the activities of actors. Focusing on the degree of content overlap between actor's information provision and the media, our study falls under the latter category. The importance of these press releases for journalists has been confirmed in various news contexts, including health issues, the environment, and national and foreign policy (Denham, 2010). While the relationship between information subsidies and news media has thus generated wide scholarly attention, longitudinal analyses of this relationship are scarce, if at all existent. By comparing press releases from two industrial sources and one NGO with news output over a period of ten years, this paper sheds light on media-source dynamics over time.

When assessing newspapers' reliance on sources and news agencies, a categorization of different types of newspapers is useful. This study differentiates between national and regional newspapers, as previous investigations have shown significant differences between national and local/regional coverage (Hansen, 2011). Regional and local newspapers traditionally fulfill a unique function in democracy, namely to hold local or regional powers to account (Freer, 2007; O'Neill and O'Connor, 2008). The current economic situation could make it harder for these newspapers to perform this task. Operating on narrow markets, the steady decrease in circulation rates and advertising revenues weights heavy on regional newspapers (Kerrigan and Graham, 2010). The Dutch regional newspaper sector is in decline: The number of journalists working for the Dutch news cooperation Wegener, the largest publisher of regional newspapers, has decreased from 2,000 in 2001 to 900 in 2012 (Dohmen, 2012).This trend is visible in other countries as well: in the UK for instance, the number of regional titles has halved over the last fifty years (Franklin, 2008). Local newsrooms are characterized as 'a pressurized and demoralized working environment' where it is 'all too easy for journalists to become dependent on the pre-fabricated, pre-packaged 'news' from resource-rich public relations organisations' (O'Neill and O'Connor, 2008). We hypothesize this reliance to be strong in particular with regard to a local (in our case nuclear) organization, because previous research has demonstrated a strong reliance of regional newspapers on local sources (Prenger et al., 2011).

H1: Regional newspapers show greater similarity with the content of news agencies and regional sources than national newspapers.

THE CENTRAL ROLE OF NEWS AGENCIES

News agencies are a dominant factor in the world of the news (Boyd-Barrett and Rantanen, 2000; Forde and Johnston, 2013). Traditionally they occupy a unique position: as no other party, they monitor newsworthy events all over the world and generate news from a wide variety of sources. To illustrate, a journalist at the German department of news agency AP receives at an average day about 4,000 to 5,000 press releases that strive for agency coverage (Boyer, 2011). The perceived prominence of a source is a key filtering principle: the more prominent the politician, public figure, organization or expert is, the more likely their statements would be turned into agency releases (Boyer, 2011). This tendency to rely on pre-packaged information provided by credible sources has increased along with the numerous challenges that agencies face (Boyer, 2011; Carlson, 2009; Forde and Johnston, 2013): First, technological developments have made news omnipresent and online accessible in a split second. The agencies have thus lost their traditional monopoly on 'raw' news and are now faced with multiple competitors that have entered their market. Second, in the digital age it is increasingly difficult to secure exclusivity of the content: as soon as a medium publishes content online, other media are able to profit for free. Third, news organizations have put the screws on the agencies. This is particularly the case in the Netherlands, where the exceptionally high level of concentration of the newspaper market places the agencies in a disadvantageous position at the negotiation table (TNO, 2011: 4). The agencies seek to remain profitable by cutting substantively in personnel and the pioneering of new markets (Manning, 2008): The Dutch market leader *ANP* for instance offer their expertise to organizations in writing successful press releases (TNO, 2011).

While the number of employees has been steadily reduced over the past years, the level of output that *ANP* produces has increased with forty percent over the period 2004-2009. These developments have led the president of *ANP* to express his concern that the agency has reached a "critical borderline of providing required quality" (TNO, 2011: 25). The precarious situation of the agencies have not withheld news media to strongly rely on news agencies for their content (Hijmans et al., 2011). On the basis of a content analysis involving nearly 60,000 articles from nine national newspapers between 2006-2008, Otto Scholten and Nel Ruigrok (2009) found an across the board trend of increased reliance on agency copy for news and background information. Within three years, the percentage of newspaper articles that was (partly) based on agency content increased from 23,9 percent to 27,6 percent. The study furthermore demonstrated that in the largest national newspapers, twenty to thirty percent of the articles containing agency content did not attribute the agency as a source.

The sections above motivate our expectation that newspapers increasingly rely on news agency content and that both agencies as well as newspapers increasingly rely on source content. This translates into the following hypothesis:

H2: The degree of content similarity between the three domains has increased over time.

Furthermore, if the reliance on pre-fabricated news is indeed related to the economic decline of the newspaper sector, we may expect this trend to be strongest for regional newspapers, who are, as argued above, particularly troubled by the economic developments:

H3: The degree of content similarity with sources and agencies over time has increased more strongly for regional newspapers than for national newspapers.

THE CASE OF NUCLEAR ENERGY

The enduring issue of nuclear energy is a suitable topic to assess the overlap of content from different domains, since it has generated public debate for decennia. The extraordinary complexity of the issue, in technical as well as psychological terms (Van Dam, 2003), ensures a wide-ranging spectrum of viewpoints. A further advantage is that with both corporate as well as nongovernmental stakeholders, different source categories are involved. The environmental organization Greenpeace has been a declared opponent of the nuclear industry and actively seek media attention for their viewpoints. The corporate sources are represented by the two largest nuclear organizations that are active in the Netherlands: global nuclear service provider *Nuclear Research and consultancy Group* (NRG), whose core business is the production of nuclear isotopes for the pharmaceutical market, and *EPZ*, an energy producer that exploits a nuclear power plant. In light of the literature, the inclusion of an 'official' governmental source would have had our preference. Yet due to the government's highly fragmentized communication on the issue, no single governmental source published enough textual material to allow for a reliable comparison.

METHOD

The central aim of the paper is to compare the content of press releases from sources with the contents of news agencies and newspapers by means of an automated technique. This section describes the technique, which has shown to produce results

that are consistent with a traditional, manual content analysis (authors, 2014). For the remainder of the paper, the three categories *sources, agencies and newspapers* will be referred to as 'domains'. Within these domains, the collection of texts from each specific source, agency or newspaper is called a 'corpus'. Thus, all articles of news agency *ANP* together form one corpus, which combined with the corpus of *Novum* forms the domain of the agencies.

CONSTRUCTING MEANING THROUGH WORD CHOICE

The meaning of a text can be derived from the words that the text consists of. Our research aim is to compare similarity between collections of texts. To assess this, we rely on a 'bag of words' representation of the texts, meaning that the order of the words is not taken into consideration. Theoretically, word order can substantially change the nature of a text and the meaning of a specific word depends on the context in which it appears. In practice however, for many purposes the list of words that a text is reduced to is an adequate reflection of the general meaning of that text (Grimmer and Stewart, 2013). We argue that the higher the content similarity between an agency corpus and a newspaper corpus, the more likely it is that the general meaning of their content overlaps. Likewise, the higher the content similarity between the corpus of a source and that of an agency or newspaper, the more successful the source has been in promoting its perspective. In the next section we shall describe our proposed technique, and illustrate with an example that the similarity score introduced below is a realistic indicator of the degree to which two texts share meaning.

TF-IDF AND COSINE SIMILARITY

The most straightforward technique to assess similarities between (collections of) texts is to compare the presence and frequency of terms (i.e. words) in these texts. However, reliance on plain term frequency to uncover the meaning of a text can be problematic, as insights from the fields of information retrieval and artificial intelligence have shown (Robertson, 2004). The most pressing obstacle is that by merely counting term frequency, all terms are considered to be equally important and informative. Yet in reality, certain terms are so commonly used in language that they tell us little about the meaning of a specific text: they are unlikely to be a useful discriminator for a text's meaning. A specific example of a common term for our case – nuclear energy - would be 'nuclear'. Looking at plain frequency of occurrence, this term is likely to appear in most of the documents and would thus have a high ranking. In fact, since

all the documents are about nuclear energy, in this situation the term tells us little about the actual meaning of the texts. It is thus advisable to assign weights to words (Robertson, 2004). The Term Frequency Inverse Document Frequency (tf-idf) measure does precisely this. Tf-idf has been developed in the early 1970s (Jones, 1972) and has since then proved valuable for numerous purposes in a variety of disciplines including linguistics (Orăsan, 2009), information retrieval (Monroe, Colaresi, and Quinn, 2009; Salton, 1991), and artificial intelligence (Pazzani, 1999). The principle of tf-idf is straightforward: common terms across documents score relatively low, uncommon terms relatively high. An example of an uncommon term in our case would be 'isotope', a very specific term that is not frequently encountered, but very informative nonetheless for the identification of an implicit frame or the similarity of texts based on cosine similarity scores. While numerous extensions and variations on the measure have been proposed, tf-idf has proved 'extraordinarily robust' (Robertson, 2004, p. 1), as illustrated by the fact that it is still at the core of many ranking methods used in search engines.

Tf-idf is calculated by looking at the number of times the word appears in a document and multiplying that number by the log of the total number of documents, divided by the number of documents that the word resides in. Consequently, it increases with the frequency of the term j, but decreases as the term occurs in more documents (k) in the set of (*N*) documents (Leydesdorff and Welbers, 2011). The formula for the tf-idf score of a term w in document j reads as follows:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \qquad (1)$$

where $tf_{i,j}$ is the number of occurrences of w in j, df_i is the number of documents containing and *N* is the total number of documents.

The tf-idf procedure results in a term-document matrix where the columns consist of all the terms and the rows are formed by the documents. The cells indicate the tf-idf score of each word that may or may not appear in the document. To determine the degree to which two articles are similar (hence, share the same words in the matrix), a standardized measures is required. To this end we rely on one of the most frequently applied tools in the discipline of data mining: *cosine distance* (Tan, Steinbach and Kumar, 2006). An important explanation for the measures' popularity - its applications can be found in the areas of business, medicine, science, and engineering, among others (see Tan et al., 2006) - is the fact that the outcome is very easy to interpret. The similarity score can take any value between zero and one, depending on the degree of

similar (tf-idf weighted) terms. If the cosine similarity value is 0, two documents (or sets of documents, in our case) do not share any terms. If the cosine similarity is 1, the two (sets of) documents are identical (Tan et al., 2006).

The cosine score informs us on how similar two matrices are: in other words, to what extent the values in the matrices are the same. Regular measures such as the Pearson correlation coefficient are biased due to the high number of corresponding zero-zero pairs in the matrix. The similarity measure thus should not depend on the number of shared zero values, but on shared ***non***-zero values. The cosine measure does exactly this, and produces an indication of the degree of similarity on the basis of all the 'positive hits'. The cosine similarity is defined as follows (where x and y represent the term frequencies of the two (sets of) documents):

To get a concrete insight in how the similarity score relates to actual texts, Appendix B provides an example of three texts and their joint cosine scores.

$$1 - \cos(x, y) = \frac{x \cdot y}{||x||||y||} \quad (2)$$

A software tool has been developed to calculate the tf-idf scores and create two-dimensional 'word-document' matrices that represent the collections of documents per corpus. In a word-document matrix, each row describes a document and each column corresponds to a word. The tool then uses these matrices to calculate the cosine score for a specified time period and specified relation (e.g. Greenpeace and *ANP*). These data are read into SPSS, after which analyses of correlation and regression are performed to answer our research questions.

DESCRIPTION DATASET

One of the advantages of a computer assisted approach is that it allows analyzing the entire population of the material rather than just a sample. This is particularly valuable because the dynamics of the relationship between sources and media tend to form patterns: Sources often dominate in the news discovery phase, while in the 'news gathering phase', journalists do (Dimitrova and Stromback, 2009). Because similarity scores between the press releases and media articles over the period 2003-12[2] are aggregated to a monthly level, these short-term fluctuations in the degree of similarity are accounted for.

All articles of both national wire services, *ANP* (*n* = 753) and *Novum* (*n* = 194) are included, as well as a wide spectrum of newspapers the contents of which are digitally available through the Lexis Nexis database: national quality newspapers (*NRC* (*n* = 374), *Volkskrant (n = 377))*, a popular newspaper (*Telegraaf, n* = 170), and two regional papers: *Noordhollands Dagblad (n = 491)* and *Provinciale Zeeuwse Courant (*n = 720*)*. The selection of the regional papers is based on the region where the two nuclear reactors in the Netherlands are situated: EPZ is active in Borssele, Zeeland (*Provinciale Zeeuwse Courant*) and NRG in Petten, North-Holland (*Noordhollands Dagblad*). Relevant articles are retrieved through a search term that included 'atomic energy' OR 'nuclear!' OR 'reactor', AND 'Borssele' OR 'Borsele' OR 'Petten'.[3]

Press releases of the following sources are included: Energy producer EPZ (*n* = 110), nuclear research and service provider NRG (*n* = 100) and non-governmental environmental organization Greenpeace (*n* = 183). The press releases of the sources are retrieved from the organizations' websites. For EPZ and NRG, all press releases on the websites are included. In the case of Greenpeace, only the releases that were labeled by the organization with the tag 'nuclear energy' were selected. The data have been aggregated to a monthly level of analysis.

---

[2] *Not all media data are available for the entire period: content of Novum is available since 2004 and Noordhollands Dagblad since 2007.*

[3] *Dutch equivalents for these terms are used ('kernenergie' for atomic energy, 'nucleair!' for nuclear!). The selection of relevant articles is based on the presence of the combined search terms and is limited to domestic news. Editorials and letters are excluded from the analysis.*

RESULTS

The result section is structured on three types of analysis. First, a correlational analysis will deliver insight in the extent to which the publication of information subsidies and media attention is correlated, offering a first understanding of the media-source dynamics. Second, the cosine similarity measure is applied to gain insight in the extent to which the content of the different corpora overlaps. Third, the similarity levels between the domains will be regressed on time to assess possible trends in reliance on source content by agencies and newspapers.

PROMINENCE OF THE DEBATE

Figures 1a and 1b show the amount of published articles on nuclear energy per source and per media domain per year.

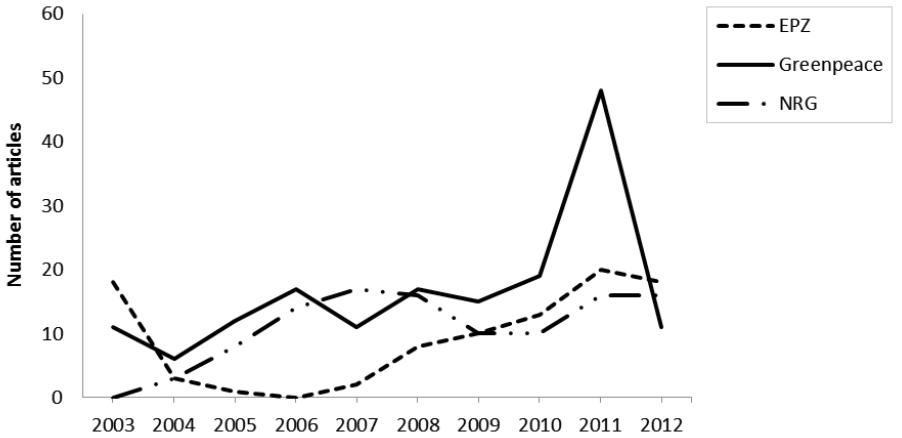


Figure 1a. Number of articles per source, 2003 - 2012

Figure 1b. Number of articles per media domain, 2003 - 2012

Figure 1a shows that overall, the three sources are publishing more stories over time. In most years, Greenpeace has the highest publication frequency, up to 48 articles in 2011: more than the combined output of the two indrustrial organizations. From Figure 1b we can infer that in this same year, media-attention was at its peak as well. Table 1 provides insight into the degree of correlation between the specific corpora (individual sources, agencies and newspapers), thus to what degree the visibility of output of these sources overlaps within months.

Table 1. Correlation of amount of articles between corpora

		Sources		Agencies		National newspapers			Regional newspapers		
		EPZ	Green-peace	*ANP*	*Novum*	*NRC*	Volks-krant	*Telegraaf*	Noord-hollands Dagblad	Provinciaalse Zeeuwse Courant	***mean***[a]
Sources	NRG	-.05	.14	.36**	.32**	.19*	.10	.22**	.30**	.38**	***.27***
	EPZ		.11	.20*	.07	.17*	.11	.08	.09	.15	***.12***
	Greenpeace			.32**	.35**	.33**	.31**	.36**	.30**	.44**	***.34***
	mean[b]			.29	.25	.23	.17	.22	.23	.32	.24
Agencies	*ANP*				.60**	.58**	.59**	.69**	.33**	.46**	
	Novum					.36**	.42**	.44**	.25**	.55**	
Newspapers	*NRC*						.78**	.55**	.14	.45**	
	Volkskrant							.63**	.13	.36**	
	Telegraaf								.35**	.41**	
	Noordhollands Dagblad									.30**	
	Df	141	141	141	141	141	141	141	141	141	

Note. [a]Mean correlation of sources with news agencies and newspapers. For EPZ and NRG, only the regional newspaper that corresponds with the location of the company is included in the calculation. [b]Mean correlation of news agencies and newspapers with sources.

* $p < .05$. ** $p < .01$.

Table 1 shows that the amounts of press releases from Greenpeace, NRG and EPZ are not correlated. In other words, the sources do not appear to communicate their message in the same time frame. When we assess the correlation of the sources with the agencies and newspapers, it can be noticed that the amount of press releases of Greenpeace correlates substantially higher with the amount of articles of the agencies and newspapers compared to the other two sources (see last column). The strongest correlations between agencies and newspapers can be found between *ANP* and the national newspapers. It becomes clear that Greenpeace, the agencies and the newspapers often share the same time frame to publish on nuclear energy. Whether there is also a shared understanding in terms of the contents of the articles is addressed in the next section.

COSINE SIMILARITY

The cosine similarity is a measure that informs about the degree of similarity between the collections of texts. Table 2 depicts the cosine values of the various relationships.

Table 2. Mean cosine similarity between sources, agencies and newspapers 2003-12

		Sources		Agencies		National newspapers			Regional newspapers		
		EPZ	Green-peace	ANP	Novum	NRC	Volks-krant	Tele-graaf	Noordhollands Dagblad	Provinciaalse Zeeuwse Courant	Mean[a]
Sources	NRG	0.19	0.22	0.33	0.42	0.3	0.39	0.29	0.53	0.21	*0.38*
	EPZ		0.22	0.28	0.47	0.38	0.36	0.33	0.12	0.45	*0.32*
	Greenpeace			0.45	0.59	0.59	0.55	0.50	0.19	0.51	*0.48*
	mean[b]			0.35	0.49	0.42	0.43	0.37	0.41	0.48	*0.39*
Agencies	ANP				0.53	0.71	0.72	0.66	0.5	0.65	
	Novum					0.64	0.64	0.59	0.36	0.61	
Newspapers	NRC						0.88	0.81	0.48	0.76	
	Volkskrant							0.83	0.58	0.76	
	Telegraaf								0.52	0.74	
	Noordhollands Dagblad									0.44	

Note. [a]Mean cosine similarity of sources in relation to news agencies and newspapers. For EPZ and NRG, only the regional newspaper is included in the calculation. [b]Mean cosine similarity of news agencies and newspapers in relation to sources. For the regional newspapers, only the similarity scores with the regional corporation and Greenpeace is included in the calculation.

We start our discussion of Table 2 by looking at the relationships between the sources. While the low similarity between Greenpeace and the industrial sources EPZ and NRG (.22 for both relations) will not come as a surprise, the even lower similarity between NRG and EPZ (.19) is interesting. The finding seems to indicate that the two organizations, apart from choosing separate publication moments, also share little commonalities in their messages concerning the issue of nuclear energy. In other words, in their external communication the two leading nuclear companies do not promote the nuclear issue from a uniform point of view, but rather from their own unique perspectives.

With respect to the content overlap between sources and agencies, results suggest that there is a moderate degree of overlap between the content of the two domains, with considerable differences between the two agencies. The content of *Novum* is overall more in line with the sources than the content of *ANP* is (.49 versus .35). Of the three sources, Greenpeace's content is generally most similar to the agencies and their content shows a relatively high similarity with the content of *Novum* (.59).

Regarding the relation between sources and newspapers, the results are consistent with the previous findings: Greenpeace's content is overall most well reflected by the newspapers. There is one notable exception however: The content of regional newspaper *Noordhollands Dagblad* scores very low on similarity with Greenpeace (.19), but fairly high with respect to NRG (.53). Examination of the terms with the highest tf-idf score confirms that *Noordhollands Dagblad* covers the issue of nuclear energy in largely the same language as NRG. Appendix C provides an overview of the most prominent terms per corpus and shows that the abbreviation 'NRG' is on top of *Noordhollands Dagblad*'s ranking. Among the most prominent terms that NRG communicates in their press releases are 'isotopes', 'medical' and 'research', terms that stress the organizations activities in the pharmaceutical world and that are also highly-ranked in the coverage of *Noordhollands Dagblad.* These terms are absent in Greenpeace's list, which consists of terms that refer to waste, danger (Chernobyl) and transport. While generally speaking the content of Greenpeace is thus most closely associated with that of the newspapers, there are noticeable differences in the source-newspaper relationships when we differentiate between type of newspaper. The quality newspapers *NRC Handelsblad* and *Volkskrant* hold the highest scores regarding content similarity with the sources (means of .42 and .43, respectively). Apparently, in their reporting on the issue of nuclear energy there are commonalities with the content of the sources. This relatively high similarity between the two papers, with .88 being the

highest similarity score in the matrix, suggests that the content of *NRC Handelsblad* and *Volkskrant* is very similar regarding nuclear energy.

As is the case with the relationship between NRG and *Noordhollands Dagblad*, so do EPZ and the regional newspaper *Provinciale Zeeuwse Courant* share a moderate degree of similarity (.45). Unlike *Noordhollands Dagblad* however, the content of *Provinciale Zeeuwse Courant* is even stronger related to the content of Greenpeace (.51). From Table 2 it has become clear that in most instances, the content of Greenpeace is considerably more reflected in particularly agency and national newspaper content than the industrial content.

Concerning the relation between agencies and newspapers, the results show that the degree of overlap between newspapers and news agencies is varied, with similarities ranging between .25 (*ANP* / *Noordhollands Dagblad*) and .63 (*Novum* / *Volkskrant*). Counter intuitively, of all newspapers the quality newspapers are the ones most similar with the news agencies: *Novum* / *NRC Handelsblad* scores .61). The regional papers have the least in common with the agencies in terms of contents, which is contradictory to our expectation that regional newspapers would be more similar to agency content than their national counterparts (*H1*).

TRENDS OVER TIME

The section above has given insight in the general strengths of the various relationships over the entire research period. As set out in the theoretical section, literature suggests that similarities between corpora may have increased over time. Regression analyses are applied with the independent variable time (***x***) explaining the degree of similarity between two domains (***y***). Data of the three domains have been aggregated to a yearly level, thus every analysis is based on ten data points. Possible trends in the degree of similarity can be established in this way. Figure 1 presents an overview of the cosine similarity score of the organizations with the agencies and newspapers.

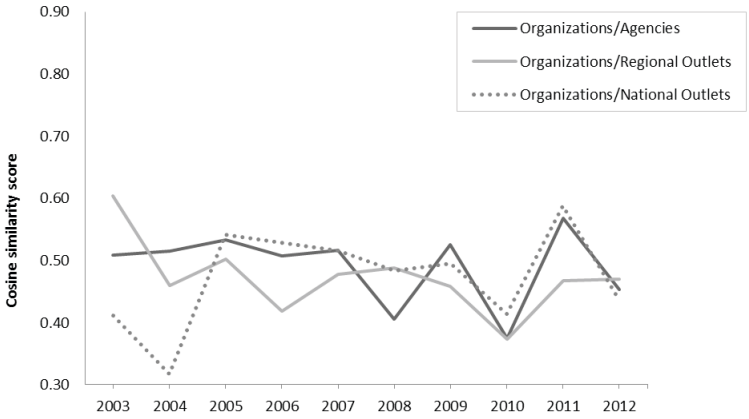


Figure 2. Cosine scores of organizations with media, 2003 - 2012

Figure 2 can partly address our expectation that the degree of content similarity between the three domains has increased over time (*H2*). As becomes apparent from the Figure, there is no clear trend for either the three relations of the organizations with the media. While the similarity score between organizations and newspapers has increased marginally between 2003 and 2012, regression analyses show that this increase is not statistically significant. Concerns about sources becoming increasingly influential are thus not confirmed by this analysis. From Figure 2 it becomes clear that the degree of similarity differs considerably between years and we will briefly zoom in on one 'dip' (2010) and one 'peak' (2011). In 2010, the media agenda was dominated by the discussion on a potential new nuclear plant, as an assessment of the media headlines reveals. The headlines of the industry's press releases indicate that only a fraction of the releases address this issue. And while judged from the NGO's headlines, Greenpeace does aim to partake in this discussion, the NGO's relatively low similarity scores with the media indicate the organization is less successful in accessing the media agenda than in most other years. The 'peak' year 2011 is marked by the nuclear incident in Fukushima, Japan, resulting in closely aligned agenda's that were largely dominated by this disaster – particularly so in the cases of the media and Greenpeace. It appears that the incident provided a 'window of opportunity' for the NGO's oppositional standpoint. To address the second part of our hypothesis, regarding our expectation that newspapers and news agencies are increasingly similar in content, we turn to Figure 3.

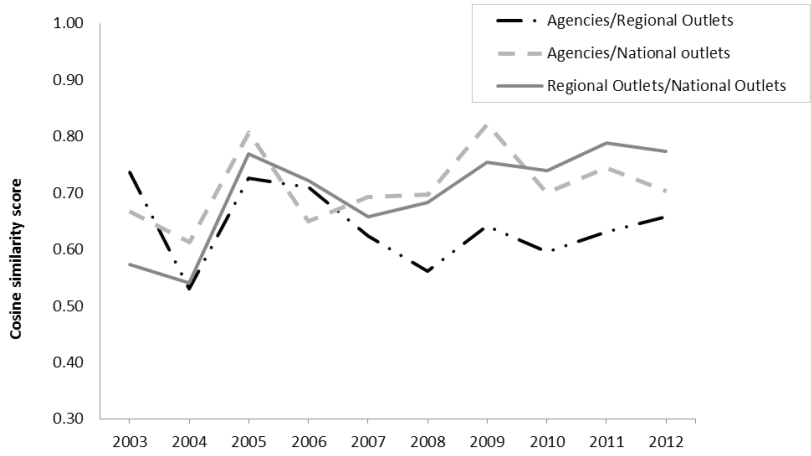


Figure 3. Inter-media cosine scores, 2003 - 2012

Figure 3 gives a visual insight into the inter-media relations. The first thing than can be noted is that the scores are structurally higher than the scores in Figure 2. Regarding the individual relations, one score is significantly increasing over time: between regional and national newspapers *(r .02, p<.05)*. There is no indication that newspapers – either regional or national - are increasingly relying on news agency content. The results of the regression analyses imply a rejection of our hypotheses on an increased content overlap between sources, agencies and newspapers (*H2* and *H3*).

DISCUSSION

Studies of journalism and news production have shown that in the relationship between sources and journalists, sources increasingly lead the dance (Carlson, 2009; Prenger et al., 2011). This study demonstrates that news agencies structurally join these ballroom sessions as well. Particularly national newspapers reflect agency content to a substantial degree. This finding is not in line with the rationale that quality newspapers' relatively large resources leads to more unique content (Scholten & Ruigrok, 2009). It may well be that the current financial and time constraints limit the newspaper's aspirations to provide unique coverage. The findings confirm the central role of news agencies in the world of the news and evoke concerns on the plurality of voices that the mainstream media offer. The ultimate consequence of these developments could be a homogenous and uncritical news landscape. The troubling economic situation of the Dutch news agency market notwithstanding, this study found no signs of national news agencies increasingly relying on the content of the selected sources. An interesting

finding is that throughout the entire period, there are considerable differences between the two rivaling agencies in terms of content similarity to the sources' press releases: the content of challenger *Novum* is significantly more similar to source content.

Apart from the role of news agencies, the impact of source content was central to the study. The content of Greenpeace was most well reflected in both the agency copy as well as the newspapers' content. This finding contradicts notions that nongovernmental organizations are generally less successful in securing media access than business sources, yet is in line with earlier findings of Greenpeace being a gatekeeper or primary definer of issues (Anderson, 1997, p. 167). As Deacon (2003) remarked, Greenpeace is increasingly "attuned to the news values of large media organizations" (p. 32). As such, it can be considered representative for the limited group of globalized, well-resourced NGOs that have tailored their strategies to attract mainstream media's interest. The vast majority of NGOs and pressure groups however does not enjoy these resources and is less successful in gaining media access (Waisbord, 2011).

It is worth restating that written press releases are only one aspect of the complex and multifaceted relationship of sources and journalists (Reich, 2010). Relationships between journalists and public relations practitioners are more complex and involve a range of written as well as verbal interactions, many of which are not easy to trace (Lewis et al., 2008). The increasingly sophisticated practice of PR-professionals to 'spin' news is a prime example of PR-influence that is difficult to examine systematically. However, as press releases still account for a substantial part of PR-activities (Forde & Johnston, 2013) they provide inevitably conservative, yet verifiable evidence. Furthermore, it is likely that the content of the press releases is in line with the messages that PR-professionals promote through other channels like face-to-face contact. Thus, while it should not be seen as an absolute measure for sources' impact on journalistic content, we expect the similarity score to provide an accurate indication for sources' success in promoting their interests.

Another point that should be taken into account is the study's time frame. Since the organizations' press releases have only been digitally archived since 2004, the current study has focused on the last decennium. Finding no significant increase of media reliance on the content of the agencies or included sources during this period, it must be noted that relevant factors like decreasing newsroom capacities and a professionalization of sources' communication have been at play for some time before the start of our research period. It is well possible that a study covering a longer period will sketch a different picture.

From a methodological viewpoint, the application of cosine similarity to compare texts from different domains has lived up to its expectations and provided useful insights. The approach reduces the complexity of the texts by evaluating and comparing the most significant terms. While this is admittedly a simplification of the texts' meaning, we posit that in general, the significant terms are indicative of the organizations' key message. When these terms are also present in media content, there is thus an overlap in agendas.

The next step would be to sophisticate the research design in such way that causality can be established. While analyses like these typically require larger datasets than the one used for the current study, time series analysis for instance could clarify the extent to which economic, technological and organizational factors are accountable for the degree in content similarity. It could also model the impact of key societal or political events on content overlap.

While in this specific case it has been shown that the news media do not increasingly rely on the press releases of the selected organizations, this does not imply that there is no growing dependency on sources in general: the number of included sources requires caution when making generalizations. Additionally, the hypothesis is tested in a very specific context – nuclear energy. Empirical research on media-source dynamics in different contexts, including a wider array of potential sources, is much needed. While nuclear energy is an important political issue as such and we have no specific reasons to expect patterns for other issues to be very different, it remains an empirical question whether we find the same similarity patterns for cases such as foreign or even celebrity news. Keeping in mind that source checking plays only a "modest role in the journalistic process" (Diekerhof & Bakker, 2012, p. 252) and that source crediting is not always as transparent as ought to be (Reich, 2010; Ruigrok & Scholten, 2009), critical research on the dynamics between sources and media remains of vital importance. The approach presented in this paper can aid in this crucial task.

Finally, concerning our findings on the regional newspapers, the relatively low similarity with the national news agenda confirms that the regional newspapers add distinctly different perspectives to the media landscape. The content is likely to be more oriented towards local and regional rather than national developments. The relatively high similarity with the corporations' content is in line with O'Neill and O'Connor's (2008) observation that regional journalism may well have become all too dependent on pre-packaged news from resource-rich public relations organizations. If anything, this finding legitimatizes further research on the relation between sources and journalists on a regional level.

CHAPTER FOUR

INTRODUCING COSINE SIMILARITY TO ASSESS FRAME OVERLAP

Abstract

Automated tools of analysis can greatly enhance research on media content. Implicit framing analysis for instance is a powerful tool to automatically extract meaning from large collections of texts. A limitation of this technique is that comparisons of content overlap between different collections of texts are typically based on a qualitative interpretation of the frames' contents. This paper complements the implicit framing technique with a quantitative measure that creates insight in the degree to which different text collections overlap in their content: cosine similarity. Analysis of the Dutch nuclear energy debate (2003-2012), including content of three organizations, two agencies and two newspapers, shows that the two approaches can be considered complementary to each other: The implicit frames provide a qualitative understanding of the content overlap while the cosine similarity scores provide quantitative insight based on a standardized measure. Wider applications of the cosine similarity measure for communication research are discussed.

This chapter has been submitted as

Boumans, J.W., Boomgaarden, H.G., & Vliegenthart, R. Application of automated content analysis and cosine similarity for framing and journalism research

Over the past decennia, mass communication research has greatly benefitted from insights and tools developed in information science. Examples can be found in the fields of political text analysis (Amazeen, 2015), social network analysis (Cho & Lee, 2010) and organizational communication (Corman, Kuhn, McPhee & Dooley, 2002; for an extensive overview see Boumans & Trilling, 2016). The advantages of computer-assisted approaches – including the capability to process large volumes of data with limited resources and increased reliability (Grimmer & Stewart, 2013) - have also drawn the attention of framing scholars (Matthes & Kohring, 2008; Van Atteveldt, Kleinnijenhuis, & Ruigrok, 2008). One particularly promising venue is to reveal the meaning of texts by looking at semantic networks, or the co-occurrence of words. Previous research has demonstrated that this technique is well capable of extracting 'implicit' frames from a variety of texts such as scientific articles (Lucio-Arias & Leydesdorff, 2007), media output (Jonkman & Verhoeven, 2013) and online public debate (Leydesdorff & Hellsten, 2005). The ability to handle such a variety of different domains of texts is a notable strength and makes the approach interesting for any scholar aiming at uncovering meaning from collections of texts. What the approach currently lacks however, are practical tools to systematically ***compare*** content similarity across different domains. Comparisons based on automated implicit framing research are currently mainly done on the basis of a qualitative interpretation of the frames' content (Van der Meer & Verhoeven, 2013). The implicit framing approach would greatly benefit from a less subjective technique to assess content overlap using a standardized measure. This article proposes such a similarity measure. The measure is based on cosine distance, which is commonly applied in information science. Since implicit frame extraction and cosine scores share the same basis – weighted word frequencies –the two approaches can be considered complementary to each other, with the implicit frames providing a qualitative account and the cosine similarity scores providing a quantitative insight in the degree of overlap between text domains. This article thus presents a methodological contribution and discusses the application and usefulness of cosine similarity in the context of a prominent field in journalism and mass communication research: framing research.

The article is structured as follows. First, the benefits of an automated text comparison tool for journalism and mass communication studies in general are discussed. Second, the implicit frame extraction technique will be contextualized within the general framing tradition, focusing on the operationalization of frames. Third, the technique and the cosine similarity measure will be discussed in detail in the method section. Fourth, there is a brief description of the specific case that serves as a substantial example

to illustrate our methodological approach: the Dutch debate on nuclear energy between 2003-2012 in press releases, news agency articles, and national and regional newspaper articles. Fifth, the results are described and compared to a manual analysis. Since the application of the cosine similarity measure is not constrained to implicit framing, the sixth section discusses wider applications of the cosine similarity index in mass communication and journalism research. Finally, the overall performance of our approach is placed into perspective and promises and pitfalls are discussed.

MASS COMMUNICATION AND THE BENEFITS OF AUTOMATED TEXT COMPARISON

The most evident benefits of automated content analysis (ACA) methods stem from their ability to extract detailed characteristics from large quantities of data, thereby contributing to both the depth and the scale of an analysis. But there are more reasons that justify the exploration of ACA. The once so surveyable mass communication landscape is now characterized by a variety of new communication channels, including blogs, news aggregators, social media, apps and newsgroups. Unlike the traditional channels, these new channels are characterized by their fluid and often interactive nature. Traditional approaches have difficulties dealing with these characteristics. Additionally, technological advances in journalism have stimulated the birth of new 'quantitative' forms of content creation, such as computer-assisted reporting, data journalism, and computational journalism (Coddington, 2014). ACA is argued to be typically better capable to identify patterns in these data than traditional content analysis (Flaounas et al. 2013).

While this is by no means a plea to abandon manual content analysis altogether, we do believe that when both the subject as well as technological opportunities changes, exploring new methodical grounds is much needed. This study presents such an exploration in the context of framing research.

POSITIONING IMPLICIT FRAMING ANALYSIS IN THE FRAMING RESEARCH TRADITION

The paper builds on the classic and widely accepted notion that framing is generally concerned with the way meaning is constructed through the presentation of an issue, by emphasizing certain aspects of an issue and ignoring others (de Vreese, 2005). Given the article's central aim – to introduce an innovative automated content analytic approach to journalism and mass communication research – it is well beyond the scope of this paper to discuss the rich and diffuse legacy of framing research here (see e.g. Matthes, 2009 for an overview). Instead, the account below focuses on the operationalization of the framing concept.

A useful distinction for the actual measurement of frames in texts is between an inductive or a deductive approach (Semetko & Valkenburg, 2000). The latter presupposes that certain frames will be explicitly present in the data and examines the occurrence of these predefined frames. Often, these frames have a generic nature, such as 'conflict' or 'human interest' framing. When previous research provides sufficient reasons to expect the presence of certain frames, the deductive approach performs well (de Vreese, 2005). A limitation of this approach is the inability to identify alternative, unanticipated frames or the emergence of new frames over time. The inductive approach, by contrast, refrains from defining frames ***a priori*** and aims to identify all the possible frames that emerge from the material during the analysis. Those frames are issue-specific by nature: they apply specifically to the particular topic under investigation. The technique proposed here contributes to the latter line of research.

Studies that rely on an inductive technique have been criticized on a number of aspects, including the labor intensity, the limitations due to small samples, and replication difficulties (Semetko & Valkenburg, 2000). A major strength of the implicit framing approach is that it can tackle these shortcomings. Since the coding is conducted automatically, both high labor intensity and small samples are no longer an issue. Furthermore, provided that the researcher is transparent in the choices made regarding the applied clustering techniques, the systematic nature of these techniques perfectly allows for replication.

As Matthes' (2009) meta-analysis of framing research showed, computer-assisted framing studies were rare: only eight out of 131 studies on framing in the period 1990-2005 were based on automated content analysis. In more recent years however, there have been a number of successful attempts to automatically extract frames (Van Atteveldt et al, 2008; Hellsten, Dawson, & Leydesdorff, 2010; Jonkman & Verhoeven,

2013; Van der Meer, Verhoeven, Beentjes, & Vliegenthart, 2014; Burscher, Odijk, Vliegenthart, de Vreese, & de Rijke, 2014). While automated content analysis can be applied on predefined frames using dictionary based approaches (e.g. Vliegenthart & Roggeband, 2007), many of the computer-assisted studies rely on an inductive approach and analyze the semantic networks of the texts. This line of research starts from the observation that words are the fundaments of language and that meaning is constructed through selection and co-occurrence of words (Hellsten et al., 2010). Word clustering techniques like factor analysis are able to reveal patterns of co-occurrence and can identify latent constructs of data (Korenius, Laurikkala & Juhola, 2007). These constructs are indicative of the meaning of the text and have been described as 'implicit frames' (Hellsten et al., 2010). This implicit framing approach has successfully been applied to various data, including the codification of scientific texts (Lucio-Arias & Leydesdorff, 2007), the discourse of public debate over time (Leydesdorff & Hellsten, 2005), and media discourse about airport risks (Jonkman & Verhoeven, 2013). Commonly, the frames arising from these analyses are visualized as semantic 'maps' where the nodes represent words and lines between the nodes represent the correlations between words. These semantic maps give a visual understanding of what the frames consist of and how they are inter-related (Leydesdorff & Hellsten, 2005).

While this visualization can be helpful, it does place limitations on the amount of data used for the determination of implicit frames: including many terms makes those semantic maps unreadable (Hellsten et al., 2010). Additionally, visualizing changes over time is a bothersome process. Since the datasets in these studies typically consist of thousands of articles that span decades, the richness of the dataset is thus often compromised by visual representation. Insightful results have also been generated however without the reliance on a visualization of the frames. For instance, the convergence of frames from different domains has successfully been traced by statistically comparing the overlap between frames in PR messages, media messages and public discourse over time (Van der Meer & Verhoeven, 2013; Van der Meer et al., 2014). Apart from the recent contributions by Van der Meer and colleagues however, implicit framing studies typically do not include a statistical comparative component. When a study includes the comparison of different domains or texts over time, it is often done subjectively on the basis of the word visualizations (Hellsten et al., 2010) or the dominant terms in the factors (Jonkman & Verhoeven, 2013). As such, this final and crucial step of contextualizing and comparing the implicit frames between different domains of text undermines the technique's key strengths: its objective character and systematic nature.

This paper advances a quantitative measure of similarity to overcome this hurdle and create statistical insight into the extent to which texts are comparable. It enables scholars to fully capitalize on the opportunities that the implicit framing approach offers, namely to analyze and compare large volumes of data. The similarity component expands the applicability of the implicit framing approach in at least two directions: over time and between domains. As such, the approach is valuable for many framing scholars, be they concerned with frame building in political documents, organizational rhetoric in press releases or interpersonal communication (e.g. comparing the discourse among different participants of social media). Applications of the cosine similarity approach in journalism and mass communication ***beyond*** framing are discussed below.

METHOD

In this section the implicit framing procedure and the cosine similarity score are explained. Both the implicit framing procedure as the cosine similarity score use the tf-idf weighted scores of the terms. This weighting procedure is described in detail in Chapter 3.

MEASURING IMPLICIT FRAMES THROUGH ROTATED FACTOR ANALYSIS

The collected texts are represented in a term-document matrix. In this matrix, the columns consist of all the terms and the rows are formed by the documents. The cells consist of the tf-idf score of each word that may or may not appear in the document. Exploratory factor analysis enables us to make sense of this term-document matrix. The total amount of terms per corpus typically well exceeds the thousands. For the sake of comprehensibility, only the 250 highest tf-idf ranked terms per corpus are included in the factor analysis. Including more terms makes the evaluation of the factors diffuse while previous research has demonstrated that their impact on a frame is very limited (Van der Meer et al., 2014). Additionally, Varimax rotation is applied to minimize the complexity of the factors (Grice, 2001). This rotation eases the interpretation of the factors because after this rotation, each original variable tends to be associated with one (or a small number) of factors, and each factor represents only a small number of variables (Abdi, 2011). The resulting factors resemble issue specific frames, which can be given a substantial qualitative interpretation on the basis of the factor loadings of the words: the higher the loading, the more explanatory the terms are for the latent underlying implicit frame.

MEASURING SIMILARITY THROUGH COSINE DISTANCE

The steps above, when carried out for each domain specific (i.e. creating a term-document matrix for each domain to be compared), create insight in the implicit frames per domain. The next goal is to obtain a statistical indication of the degree to which the domains overlap. Unlike factor analysis, which is based on a selection of the most important terms, the comparative analysis includes all terms of the domains. The proposed measure is based on a very basic assumption, namely that the degree to which two collections of texts share the same frames (indicated by key terms) is related to the degree to which all (tf-idf weighted) terms in those collections overlap. In other words: texts that have many words in common are likely to be discussing the same thing. To standardize the degree of similarity, one of the most frequently applied tools in the discipline of data mining is applied: ***cosine distance*** (Tan et al., 2006). An important explanation for the measures' popularity - its applications can be found in the areas of business, medicine, science, and engineering, among others (see Tan et al., 2006) - is the fact that the outcome is very easy to interpret. The similarity score can take any value between zero and one, depending on the degree of similar (tf-idf weighted) terms. If the cosine similarity value is 0, two documents (or sets of documents, in our case) do not share any terms. If the cosine similarity is 1, the two (sets of) documents are identical (Tan et al., 2006).

Like the factor analysis, the technique is based on term-document matrices. The cosine score informs us on how similar two matrices are: in other words, to what extent the values in the matrices are the same. Regular measures such as the Pearson correlation coefficient are biased due to the high number of corresponding zero-zero pairs in the matrix. The similarity measure thus should not depend on the number of shared zero values, but on shared ***non***-zero values. The cosine measure does exactly this, and produces an indication of the degree of similarity on the basis of all the 'positive hits'. The cosine similarity is defined as follows (where x and y represent the term frequencies of the two (sets of) documents):

$$1 - \cos(x, y) = \frac{x \cdot y}{||x||||y||} \qquad (2)$$

To illustrate this measure in practice, Appendix B provides an example of three articles and their mutual similarity scores.

DESCRIPTION DATASET: DUTCH DEBATE ON NUCLEAR ENERGY

The implicit framing approach and the test of text similarity using cosine similarity scores as described above are illustrated on the Dutch nuclear energy debate (from 2003 to 2012) with a focus on similarities of frame appearances in different domains of texts, specifically releases of organizations, news agencies and newspapers. Organizational press releases play a central role in the daily information exchange between sources and journalists. The substantial influence of press releases on news coverage has been documented repeatedly (e.g. Ervajec, 2005; Reich, 2010; Sissons, 2012). Similarly, the importance of news agencies as information brokers has been highlighted (Boyd-Barrett & Rantanen, 2000; Forde & Johnston, 2013). The enduring issue of nuclear energy with its declared opponents and proponents is an excellent topic to assess the overlap of source content and media content. The extraordinary complexity of the issue, in technical as well as psychological terms (Van Dam, 2003), leaves ample room for a wide-ranging spectrum of viewpoints and frames. The two largest Dutch nuclear organizations are selected to represent the corporate category: global nuclear service provider *Nuclear Research and consultancy Group (NRG)*, whose core activities are to supply products for the pharmaceutical and medical sector, and *EPZ*, an energy producer that exploits a nuclear power plant. The non-governmental environmental organization *Greenpeace,* a prominent opponent of the nuclear industry, completes the data from sources. All press releases for the entire period are retrieved from the organizations' websites. For EPZ (n = 110) and NRG (n = 100), all press releases published on the websites are included. For Greenpeace (n = 183), only the releases labeled with the tag 'nuclear energy' were selected. Aside press releases, the dataset contains wire service content and newspaper articles. Relevant media articles are retrieved through a search term that included 'atomic energy' OR 'nuclear!' OR 'reactor', AND 'Borssele' OR 'Borsele' OR 'Petten'.[4] All relevant articles of both national wire services, *ANP* (n = 753) and *Novum* (n = 194) are included, as well as one national and one regional Dutch newspaper, the contents of which are digitally available through the Lexis Nexis database: the national quality newspaper *Volkskrant* (n = 377) and a regional newspaper: *Noordhollands Dagblad* (n = 491*)*. The selection of this regional paper is based on the region where the largest nuclear reactor, *NRG,* is situated.

[4] *Dutch equivalents for these terms are used ('kernenergie' for atomic energy, 'nucleair!' for nuclear!).*

While an extensive comparison of our approach with other ways of (automated) content analysis is beyond the aim and scope of this paper, we do offer a concise comparison with a manual content analysis of the same dataset that focused on the presence of themes[5] (Authors, 2014). The sample for the manual coding is constructed through a stratified sampling technique, with the criterion that at least ten percent of every subpopulation must be included. The total sample consisted of 453 texts.

RESULTS

The results section is structured along two types of analysis. First, the implicit frames will be extracted from the corpora (collections of texts per source/medium). Second, the cosine similarity measure is applied to gain insight in the extent to which the content of the different corpora overlaps. Both analyses are based on the tf-idf weighted scores of the terms.

IMPLICIT FRAME EXTRACTION

As discussed, factor analysis is a suitable technique to identify associative networks, or implicit frames, which reflect the dominant meaning in texts. Table 1 depicts the results of the factor analyses. Every factor indicates a network of terms that co-occur in these texts. The factors with the highest eigenvalues explain most of the variance in the collection of terms. To keep the description succinct, only the four most prominent factors of every corpus will be discussed. The content of each frame is illustrated by a selection of key terms. These key terms are among the highest loading terms on the factor and have a factor loading of at least .50, which means they are highly indicative of the frame's substantive content.

[5] *A total of fifteen themes have been operationalized. The mean intercoder reliability score for these themes was satisfactory: Krippendorff's Alpha .83. One theme (context of national energy supply) scored lower than desirable (K' Alpha .62). Analysis indicates a very low presence of this theme in the manual dataset and the variable is discarded from comparison.*

Table 1. Overview Implicit Frames

	Organizations			News agencies		Newspapers	
	NRG	EPZ	Greenpeace	ANP	Novum	Volkskrant	NHD
factor 1	medical	nuclear transport	waste transport	production stop; isotope shortage	waste transport	production stop; isotope shortage	new nuclear plant
Key terms	therapy, diagnosis, medicine, treatment	transport, departure, track, radiation, ministry, application	contamination, waste, plutonium, recycle, France	hospital, medicine, maintenance, isotopes	transport, nuclear waste, activists, track, Greenpeace	isotope, shortage, medical, maintenance	permit application, procedure, precondition, location
Eigenvalue	16,31	17,16	10,27	11,49	8,4	10,06	16,97
Explained variance in %	8,11	9,98	4,82	4,14	3,78	3,84	6,79
factor 2	waste transport	waste storage	alternatives	production stop; isotope shortage	production stop; isotope shortage	NRG production stop	production stop; isotope shortage
Key terms	transport, radioactive, waste, stored	plutonium, permit, reprocess, storage Greenpeace	biomass, solution, wind, sustainability, climate change	NRG, reactor, leakage, shutdown, investigation,	medicine, isotope, shortage, problems, closed, production	NRG, safety, maintenance, legal proceeding, service	isotope, medical, production, hospital, treatment
Eigenvalue	10,41	10,14	8,5	9,22	8,38	8,36	7,18
Explained variance in %	5,18	5,9	3,99	3,32	3,78	3,22	2,87
factor 3	advocacy	advocacy	waste dump	waste transport	risk and safety	waste transport	regional policy
Key terms	expertise, Europe, important, sustainable	research, important, economic, safe, IAEA	uranium, Russia, recycling, transported, waste	nuclear transport, Greenpeace, waste, protesters	stress test, safety measures, Fukushima, disaster	transport, Greenpeace, waste, radio-active, nature	motion, council, preserve, employment, job
Eigenvalue	9,86	6,96	7,55	8,18	7,55	7,06	6,07
Explained variance in %	4,9	4,05	3,55	2,94	3,6	2,72	2,43
factor 4	risk reports	closure	accidents	miscellaneous	NRG production stop	new nuclear plant	NRG production stop
Key terms	risk, authority, report, measured, event	future, close, open, sustainability, energy supply	Chernobyl, disaster, accident, contaminated, Fukushima	nucleonic, uranium enrichment, technique, government	restart, cooling system, postponed, problems, NRG	Delta, develop, RWE, billion, Zeeland	repair, halted, cooling system, restart
Eigenvalue	7,93	6,44	6,72	7,71	7,1	6,07	5,92
Explained variance in %	3,94	3,74	3,15	2,77	3,38	2,34	2,37
Cumulative expl. variance	22,13	23,67	15,51	13,17	14,54	12,12	14,46

This section describes the findings as presented in Table 1 in detail.

Organizations. From interpreting the key terms in the upper left cell we can infer that *NRG* emphasizes the 'medical' aspect of the company's activities. The second factor deals with the transport and storage of nuclear waste; while the third factor can be considered a general nuclear 'advocacy' frame, stressing *NRG's* prominent role on the international market and their contribution to sustainability. Finally, the fourth factor reflects *NRG's* regular safety reports. In the content of EPZ (column 2), terms as 'transport', 'departure', 'rails' and 'container' form the most prominent factor. This factor can be typified as being concerned with 'nuclear transport'. The second factor reflects the matter of waste reprocessing and storage. Greenpeace, EPZ's most prominent opponent, is explicitly mentioned in this context. The third factor consists of terms that point at the technical and economic value of the research that is conducted ('technical', 'economic', 'research', 'world', 'important'). Finally, the fourth factor of EPZ stresses the consequences of an eventual closing down ('open', 'close') on the long term ('sustainability', 'future'). In the communication of Greenpeace (column 3), the matter of nuclear reprocessing and the waste transport is most prominent. The second factor reflects their discussion on alternatives for nuclear energy ('wind', 'biomass', 'solution'). The third factor again deals with waste, but this factor is concerned with the dumping of Dutch uranium in Russia. The NGO's fourth factor captures the nuclear accidents of the (recent) past, a topic that is absent in the industry's frames.

Agencies. The first two factors of *ANP* are concerned with the problems of NRG and consequences for the production of isotopes. The third factor focuses on the waste transport and includes references to Greenpeace's protests. *ANP*'s fourth factor is ambiguous, referring to both technical nuclear aspects as well as politics. *Novum*'s first factor resembles the 'waste transport' frame that both NRG and Greenpeace use and includes some terms that refer explicitly to Greenpeace and their opposition ('Greenpeace', 'activists', 'objection'). The second frame represents the consequences of a production stop ('closed', 'reparation', 'problems', 'production') for the medical world ('medicine', 'isotope', 'hospital', 'cancer', 'medical'). It has clear overlap with *NRG's* most prominent factor. The third factor discusses safety tests following the disaster in Japan and shows a fair degree of overlap with Greenpeace's 'disaster' frame. Finally, in the fourth factor *NRG's* production stop is discussed again.

Newspapers. As apparent from the sixth column of Table 1, *de Volkskrant* frames *NRG's* production problems in terms of medical (factor 1) and legal and safety (factor 2) consequences. The third factor covers the waste transports and has explicit references to Greenpeace and the organization's core concern: nature and environment. Finally, the fourth factor is related to the plans of the German energy provider RWE and Dutch subsidiary Essent of developing a new nuclear power plant (labeled 'Delta') in Zeeland. This subject also appears to form the most prominent factor of *Noordhollands Dagblad*. Factor 2 includes terms related to *NRG's* activities in supplying medical products ('isotopes') and thus represents the 'medical' factor that is also present in *NRG's* communication. The third factor includes references to employment and regional policies. Finally, the fourth factor discusses *NRG's* technical problems and production stop and restart. The key themes of *Greenpeace* - accidents and protest - are absent in the regional newspaper's most dominant frames.

Table 1 shows that the three organizations have one frame (waste transport) in common, although particularly Greenpeace applies a different discourse than the industry's organizations do. While the latter organizations use neutral terms to inform about the transports (EPZ even avoids the term 'waste'), Greenpeace prominently links the transport to 'contamination'. Apart from the waste frame, the organizations predominantly communicate unique frames. NRG stresses the organization's activities for the medical branch, their expertise and the sustainable character of the technology, while EPZ stresses the importance of their research and the safety of their operations. Greenpeace however focuses on accidents and alternatives to nuclear energy, two frames that are absent in the industry's communications. On the basis of this qualitative comparison, the similarity between the organizations' communication appears rather low.

Looking at the similarities with the media coverage, it can be noted that *NRG's* production stop and medical consequences are visible in both agency as well as newspaper content, whereas the frames of EPZ appear to resonate least in the media. Explicit references to Greenpeace as well as her key topics (particularly waste, and in the case of *Novum* also the disaster frame) are present among both agencies as well as the national outlet, but are absent in the regional outlet. This latter outlet appears to have more substantive content in common with the regional organization NRG, for which a higher similarity score could thus be expected.

COMPARISON WITH MANUAL CODING

The extracted implicit frames of all corpora show a strong overlap with the results of a manual content analysis study using the same dataset (Boumans & Vliegenthart, 2014), in which themes are coded. The dominant themes in Greenpeace's texts for instance are 'waste dump' (referred to in 53 % of the coded articles) and 'accidents and disasters' (present in 81 %).These themes correspond logically with three out of four of Greenpeace's implicit frames: waste transport, waste dump and accidents. References to alternative energy resources - the fourth frame - are found in 33% of the organization's articles. Equally, the manual coding showed that, for example, the content of the local newspaper *NHD* predominantly dealt with the option of a new nuclear plant (55,3%), medical implications of the production stop (52,6%) and economic of financial aspects (39,5%). These themes are again consistent with *NHD*'s implicit frames (new nuclear plant, medical implications, and regional policy). Given this strong consistency with the outcomes of the manual analysis, we are confident that the automated technique yields results that are an accurate reflection of the text's main contents.

COSINE SIMILARITY

The implicit frame extraction has provided an understanding of the different frames that are communicated by the domains. The second step is to create a quantifiable insight of their content overlap. Table 2 depicts the cosine values of the various relationships.

Table 2. Mean Cosine Similarity Scores between Organizations, Agencies and Newspapers, 2004-2013

		Organizations		News agencies		Newspapers		
		EPZ	Greenpeace	*ANP*	*Novum*	*Volkskrant*	NoordHollands Dagblad	*Mean*[a]
Organizations	NRG	0,17	0,20	0,29	0,42	0,39	0,49	0,40
	EPZ		0,22	0,23	0,34	0,33	0,11	0,25
	Greenpeace			0,36	0,51	0,50	0,19	0,39
Mean				*0,29*	*0,42*	*0,41*	*0,26*	*0,35*
Agencies	*ANP*				0,47	0,58	0,25	
	Novum					0,63	0,34	
Newspapers	*Volkskrant*						0,37	

Note. [a] Mean cosine similarity score of the organizations with the agencies and newspapers

We start examining Table 2 by looking at the relationship between the organizations. The low similarity scores between the 'anti-nuclear' NGO Greenpeace and the 'pro-nuclear' organizations NRG and EPZ (respectively .20 and .22) do not come as a surprise and have already become clear from the framing analysis. However, the even lower similarity between NRG and EPZ (.17) is interesting. This indicates that the two leading nuclear companies have little in common in their messages. Thus, while the organizations both promote the same topics to a certain extent (e.g. on waste transport and an advocacy frame, see Table 1), they each do so using a very different semantic network.

With respect to the content overlap between organizations and agencies, results suggest that there is a moderate degree of overlap, with noticeable differences between the two agencies. The content of *Novum* is overall considerably more in line with the organizations than the content of *ANP* is (mean cosine scores of respectively .42 versus .29). This corresponds with the analysis of *Novum*'s frames, which could all directly be related to the organizations' content. Of the three organizations, Greenpeace's content is generally most similar to the agencies. The NGO's content particularly shows a fairly high similarity with the content of *Novum* (.51), arguably caused by the disaster frame that both corpora share.

Regarding the relation between organizations and the national newspaper *de Volkskrant*,

again Greenpeace's content is most well reflected. In the case of the regional outlet *Noordhollands Dagblad* however, the content similarity with Greenpeace is very low (.19). These findings are consistent with the automated frame extraction, which show explicit references of *de Volkskrant* to Greenpeace and her frames, as well the absence of these frames in the regional newspaper. With respect to the regionally active organization NRG on the other hand, the similarity score with *Noordhollands Dagblad* is considerably higher (.49). This supports the indications of the framing analysis about a strong relationship between the content of the regional newspaper and that of the nuclear organization.

Concerning the relation between agencies and newspapers, the results show that the content of the national newspaper *de Volkskrant* and the press agencies are considerably higher than is the case for the regional outlet, with mean similarities being .61 for *de Volkskrant* and .30 for *Noordhollands Dagblad* (not depicted in table).

On the basis of our examination of the frames and the cosine scores, it can be confirmed that the degree of content similarity between organizations, agencies and newspapers is reflected in the dominant semantic networks. The higher the cosine similarity between two corpora, the more likely it is that there are corresponding frames. Likewise, corpora that have a low cosine similarity score have little or no similarity in frames. The regional newspaper *Noordhollands Dagblad* forms the clearest illustration of the fact that the two techniques are complimentary. The dominant frames of this newspaper all have a strong local orientation and are often explicitly directed at NRG. Consequently, the similarity score with this actor is relatively high, compared to the other scores. Equally, the fact that the newspaper's frames do not overlap with either EPZ or Greenpeace is also reflected in the relatively low cosine score between the corpora of these organizations and the newspaper.

WIDER APPLICATIONS OF COSINE SIMILARITY

The paper's focus has been on extracting frames from texts and how the cosine similarity measure is complementary in offering a standardized measure to assess similarity in framing. However, the tool's general nature and its main feature - to assess similarities in texts - make it applicable to a wider context within journalism and mass communication research. It is, for example, also well able to empirically assess gatekeeping routines. Gatekeeping - the process by which source messages are selected, transformed and transmitted as news content - has been subject of decennia of research. While the power of the traditional gatekeeper has decreased considerably in modern society (Singer, 2006), most research continues to focus on the role of the journalists and editors rather than the role of those gated, and analyze new gatekeeping phenomena with old tools (Barzilai-Nahon, 2009, pp. 32—37). By comparing the contents of multiple sources, the cosine similarity measure can facilitate research that takes into account the wide variety of sources attempting access. It can be argued that sources that overlap most in their communication with the subsequent media content are most successful in 'passing the gates'.

Research on content homogenization is another area where the measure can provide valuable insights. For instance, calls have been made to assess the effect of chain ownership models on the content of regional and national newspapers across time and between owners (Sjovaag, 2014), and this tool can facilitate such large-scale research. In a similar vein, the tool can aid research on content diversity across platforms (e.g. comparing print and online news) or on 'churnalism', the journalistic practice of reproducing subsidized content provided by PR-professionals (Davies, 2008). Scholars interested in the relation between (partisan) media and agendas of various parties, as for example reflected in parliamentary questions, could also find the measure useful. Basically, the cosine similarity measure is an interesting tool to investigate any theoretical notion for which the level of similarity across communication is relevant.

DISCUSSION

With ever increasing amounts of data to consider, research in mass communication and journalism has a lot to gain from adopting and advancing automated techniques. The implicit frame extraction approach pioneers such a technique, and has proven its value across different fields of interest. While it is useful for extracting meaning from large collections of texts, the implicit framing approach currently lacks an insightful, empirical measure to assess similarities between texts. This paper has proposed such a measure. The contribution in terms of quantifying content overlap has been demonstrated by applying the measure on a dataset including content of three organizations, two agencies and two newspapers. Its applicability goes well beyond studying frames in the media however: it can be used to study a wide range of different types of messages. The increased scope that the approach allows for promotes more comprehensive research over time, enabling for instance a media scholar to assess not only similarities of news content across newspapers, but also the consistency of a single outlet's coverage over-time.

Results demonstrate that the proposed cosine similarity scores offer a refined insight into the degree of content overlap. The results of the implicit frame analysis are broadly in line with those of a manual content analysis. This consistency has also been reported in earlier studies that applied a combination of manual and automated methods (see De Graaf & Van der Vossen, 2013, for a more elaborate discussion). In some instances, the two methods resulted in slightly different outcomes that can be traced back to the inductive versus deductive nature of the methods. In practical terms, the data preparation phase for the automated method requires a relatively large time investment. In the data gathering and analysis phase however, the method is far more efficient, allowing for the analysis of a larger amounts of material than in the manual case.

Although the approach clearly has a contribution to the automated framing field, it has its limitations as well. One limitation arguably concerns the interpretation of the cosine score: while the two extremes - unique versus identical - are unambiguous, a cosine score of for instance .34 is harder to interpret. Future studies could systematically compare the levels of overlap between texts for a range of cosine similarity scores and provide more qualitative insight in the meaning of various scores.

It has been proposed that framing research would benefit from a dynamic rather than a static perspective (de Vreese, 2012), and computer-assisted approaches are the optimal instrument to study data in more comprehensive fashion. Combining the cosine similarity approach with time series analysis for instance, provides the researcher with a powerful tool to link media theories with empirical data. Time series analysis could create insight in the influence of meso,- and macro factors on the degree of content similarity. The ability to make causal inferences about the existence and overlap of frames is a promising enrichment of framing research.

REFERENCES

Abdi, H. (2003). Factor Rotations in Factor Analyses. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds.), Encyclopedia of Social Sciences Research Methods. Thousand Oaks (CA): SAGE Publications.

Allan, S. (2004). *News culture* (2nd ed). Maidenhead: Open University Press.

Amazeen, M. A. (2015). Developing an Ad-Reporting Typology. A Network Analysis Approach to Newspaper and Fact-Checker Coverage of the 2008 Presidential Election. *Journalism & Mass Communication Quarterly, 92*(3), 617-641.

Anderson, A. (1997). *Media, culture and the environment*. London: UCL Press

Anderson, P. J., & Ward, G. (Eds.). (2007). *The future of journalism in the advanced democracies*. Ashgate Publishing, Ltd..

ANP bepaalt in grote mate de nieuwsagenda (2015, December 9). Retrieved from https://www.svdj.nl/de-stand-van-de-nieuwsmedia/ANP-bepaalt-in-grote-mate-de-nieuwsagenda/

Baker, C. E. (2009). Viewpoint diversity and media ownership. *Federal Communications Law Journal*, *61*, 651-671.

Bakker, P. (2002). Free daily newspapers; business models and strategies. *Journal of MediaManagement, 4,* 180-187.

Bakker, P. (2008). The simultaneous rise and fall of free and paid newspapers in Europe. *Journalism Practice*, *2*(3), 427-443.

Bakker, P., & Scholten, O. (2011). *Communicatiekaart van Nederland: overzicht van media en communicatie.* (8th ed). Deventer: Kluwer.

Barzilai-Nahon, K. (2009). Gatekeeping: A critical review. *Annual Review of Information Science and Technology, 43*(1), 1-79.

Bereik nu.nl. (*n*.d.). Retrieved from http://www.sanoma.nl/merken/bereik/nunl/

Berger, G. (2000). Grave new world? Democratic journalism enters the global twenty-first century. *Journalism Studies, 1*(1), 81-99.

Bergman, T. (2014). The case for a Dutch propaganda model. *International Journal of Communication, 8,* 20.

Bevan, T., Coen, E., & Coen, J. (1998). The Big Lebowski. United States of America: Polygram Filmed Entertainment.

Boczkowski, P. J., & De Santos, M. (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.

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism, 4*(1), 8-23.

Boumans, J. W., & Vliegenthart, R. (2014). 'Safety first' versus 'op de barricaden'. Een inhoudsanalyse van het nucleaire debat in Nederland. [Safety first versus on the barriers. A content analysis of the nuclear debate in the Netherlands] *Tijdschrift voor Communicatiewetenschap, 42*(4), 358-377.

Boyd-Barrett, O. (1980). *The international news agencies (Vol. 13).* London: SAGE Publications.

Boyd-Barrett, O., & Rantanen, T. (2000). European national news agencies: The end of an era or a new beginning? *Journalism, 1*(1), 86-105.

Boyer, D. (2011). News agency and news mediation in the digital era. *Social Anthropology, 19*(1), 6-22.

Brandenburg - van de Ven, T. (2015, May 18). Verlies voor ANP. Retrieved from https://www.villamedia.nl/artikel/verlies-voor-ANP

Brants, K., & Van Praag, P. (2006). Signs of media logic half a century of political communication in the Netherlands. *Javnost-the public, 13*(1), 25-40.

Broersma, M. (2009). Waarheid in tijden van crisis. In B. Ummelen (Ed), *De journalistiek in diskrediet* (pp. 23-40).Nijmegen: KIM.

Broersma, M., den Herder, B., & Schohaus, B. (2013). A question of power. The changing dynamics between journalists and sources. *Journalism Practice, 7*(4), 388-395.

Brüggemann, M., Engesser, S., Büchel, *F.*, Humprecht, E., & Castro, L. (2014). Hallin and Mancini revisited: Four empirical types of western media systems. *Journal of Communication, 64*(6), 1037-1065.

Burscher, B., Odijk, D., Vliegenthart, R., de Rijke, *M.*, & de Vreese, C. H. (2014). Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis. *Communication Methods and Measures, 8*(3), 190-206.

Carlson, C. (2009). Dueling, Dancing or dominating? Journalists and their sources. *Sociology Compass, 3(*4), 526–542.

Carlson, M. (2011). Whither anonymity? Journalism and unnamed sources in a changing media environment. *In B. Franklin &* M. *Carlson (Eds.), Journalists, Sources and Credibility: new perspectives* (pp. 37-48). New York: Routledge.

Castells, M. (2008). The new public sphere: Global civil society, communication networks, and global governance. *The Annals of the American academy of Political and Social Science, 616*(1), 78-93.

Cho, H., & Lee, J. (2010). Collaborative information seeking in intercultural computer-mediated communication groups. Testing the influence of social context using social network analysis. *Communication Research, 35,* 548-573.

Coddington, M. (2015). Clarifying journalism's quantitative turn: a typology for evaluating data journalism, computational journalism, and computer-assisted reporting. *Digital Journalism, 3*(3), 331-348.

Corman, S. R., Kuhn, T., McPhee, R. D., & Dooley, K. J. (2002). Studying Complex Discursive Systems. *Human communication research, 28*(2), 157-206.

Cottle, S. (2000). Rethinking news access. *Journalism Studies, 1,* 427–48.

Cottle, S. (2003). *News, public relations and power.* London: SAGE Publications.

Curran, J. (1996). Mass media and democracy revisited. In: J. Curran and M. Gurevitch (Eds), *Mass media and society* (2nd ed.) pp. 81-119. London: Edward Arnold.

Curran, J., Iyengar, S., Brink Lund, A., & Salovaara-Moring, I. (2009). Media system, public knowledge and democracy: A comparative study. *European Journal of Communication, 24*(1), 5–26. doi:10.1177/0267323108098943.

Curtin, P. A., & Rhodenbaugh, E. (2001). Building the news media agenda on the environment: A comparison of public relations and journalistic sources. *Public Relations Review, 27*(2), 179-195.

Curtin, P.A. (1999). Reevaluating Public Relations Information Subsidies. *Journal of Public Relations Research 11*(1), 53–90.

Czarniawska, B. (2011). *Cyberfactories.How news agencies produce the news.* Cheltenham: Edward Elgar.

Davies, *N.* (2008). *Flat Earth News.* London: Chatto & Windus.

Davis, A. (2000). Public relations, news production and changing patterns of source access in the British national media. *Media, Culture & Society 22*(1): 39–59.

Davis, A. (2002). *Public Relations Democracy: Public Relations, Politics and the Mass Media in Britain.* Manchester: Manchester University Press.

Davis, Aeron. 2008. “Public Relations in the News.” In Pulling Newspapers Apart, edited by Bob Franklin, 256–264. Abingdon: Routledge.

De Graaf, R., & Van der Vossen, R. (2013). Bits versus brains in content analysis: comparing the advantages and disadvantages of manual and automated methods for content analysis. Communications: The European Journal of Communication Research, 38(4), 433-443.

de Vreese, C. (2005). News framing: Theory and typology. *Information Design Journal + Document Design, 13(1), 51-62.*

de Vreese, C.H. (2012). New avenues for framing research. *American Behavioral Scientist, 56**(3),*** 365-375. doi: 10.1177/0002764211426331

Deacon, D. (2003). Non-governmental organizations and the media. In S. Cottle (Ed.), News, public relations and power (pp. 99–115). London: SAGE Publications.

Denham, B. E. (2010). Toward conceptual consistency in studies of agenda-building processes: A scholarly review. *The Review of Communication, 10*(4), 306-323.

Diekerhof, E., & Bakker, P. (2012). To check or not to check: An exploratory study on source checking by Dutch journalists. *Journal of Applied Journalism & Media Studies, 1*(2), 241-253.

Dimitrova, D. V., & Strömbäck, J. (2009). Look who's talking: Use of sources in newspaper coverage in Sweden and the United States.*Journalism Practice, 3*(1), 75-91.

Dohmen, A. (2012, February 29). [news article *NRC* Next]. Nog meer banen schrappen? Onmogelijk'. ['Cutting even more jobs? Impossible'].

Doudaki, V., & Spyridou, L. P. (2013). Print and online news: Remediation practices in content and form. *Journalism Studies, 14*(6), 907-925.

Doyle, G. (2015). Multi-platform media and the miracle of the loaves and fishes. *Journal of Media Business Studies, 12*(1), 49-65.

Erjavec, K. (2005). Hybrid public relations news discourse. *European Journal of Communication, 20*(2), 155-179.

Fenton, *N.* (2010). *New media, old news: Journalism & democracy in the digital age.* London: SAGE Publications.

Fenton, *N.* (2011). Deregulation or democracy? New media, news, neoliberalism and the public interest. *Continuum, 25*(1), 63-72.

Flaounas, I., Ali, O., Lansdall-Welfare, T., De Bie, T., Mosdell, *N.*, Lewis, J., & Cristianini, *N.* (2013). Research methods in the age of digital journalism: Massive-scale automated analysis of news-content. *Digital Journalism, 1*(1), 102-116.

Forde, S. & Johnston, J. (2013). The news triumvirate. *Journalism Studies, 14*(1), 113-129.

Franklin, B. (2011). Sources, credibility and the continuing crisis of UK journalism. *In B. Franklin &* M*. Carlson (Eds.), Journalists, Sources and Credibility: new perspectives* (pp. 90-107). New York: Routledge.

Franklin, B. (2012). The future of journalism: Developments and debates.*Journalism studies, 13*(5-6), 663-681.

Franklin, B. (Ed.). (2008). *Pulling Newspapers Apart: analysing print journalism.* New York: Routledge.

Freedman, D. (2010). The political economy of the 'new' news environment. In: A. Fenton (Ed.), *New media, old news* (pp. 35-50). London: SAGE Publications.

Freer, J. (2007). UK regional and local newspapers. In P. Anderson & G. Ward (Eds.), *The Future of Journalism in the Advanced Democracies* (pp. 89-103). London: Ashgate.

Gans, H. (1979). *Deciding What's News.* New York: Pantheon.

Golan, G. (2006). Inter-media agenda setting and global news coverage: Assessing the influence of the New York Times on three network television evening news programs. *Journalism studies, 7*(2), 323-333.

Gold, D., & Simmons, J. L. (1965). News selection patterns among Iowa dailies. *Public Opinion Quarterly,* 425-430.

Golding, P. & Elliott, P. (1979). *Making the News.* New York: Longman.

Greenberg, J., Knight, G., & Westersund, E. (2011). Spinning climate change: Corporate and NGO public relations strategies in Canada and the United States. International Communication Gazette, 73(1-2), 65-82.

Grice, J. W. (2001). A comparison of factor scores under conditions of factor obliquity. *Psychological Methods, 6,* 67-83.

Grimmer, J. & Stewart, B. (2013). Text as data. The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis,* 1-31.

Haakman, D. (2008, January 21). ANP is vaak sneller, maar *Novum* is goedkoper. [News article]. Retrieved from http://vorige.*NRC*.nl/article1879852.ece/ANP_is_vaak_sneller,_maar_*Novum*_is_goedkoper

Hafez, K. (2007). *The myth of media globalization.* Cambridge: Polity Press.

Hall, S., Critcher, C., Jefferson, T., Clarke, J., & Roberts, B. (1978). *Policing the Crisis: Mugging, the State, and Law and Order.* London: McMillan.

Hallin, D. C., & Mancini, P. (2004). *Comparing media systems: Three models of media and politics.* Cambridge University Press.

Hanitzsch, T., & Mellado, C. (2011). What shapes the news around the world? How journalists in eighteen countries perceive influences on their work. *The International Journal of Press/Politics, 16(*3), 404-426.

Hanitzsch, T., Anikina, M., Berganza, R., Cangoz, I., Coman, M., Hamada, B., ... & Mwesige, P. G. (2010). Modeling perceived influences on journalism: Evidence from a cross-national survey of journalists. *Journalism & Mass Communication Quarterly, 87*(1), 5-22.

Hanitzsch, T., Hanusch, *F.*, & Lauerer, C. (2014). Setting the Agenda, Influencing Public Opinion, and Advocating for Social Change: *Det*erminants of journalistic interventionism in 21 countries. *Journalism Studies*, 1-20.

Hansen, A. (2011). Communication, media and environment: Towards reconnecting research on the production, content and social implications of environmental communication. *International Communication Gazette, 73*(1-2), 7-25.

Harcup, T. (2015). *Journalism: Principles and Practice* (3rd ed.). London: SAGE Publications.

Hellsten, I., Dawson, J., & Leydesdorff, L. (2010). Implicit media frames: Automated analysis of public debate on artificial sweeteners. *Public Understanding of Science, 19(*5), 590-608.

Hijmans, E., Schafraad, P., Buijs, K., & d'Haenens, L. (2011). Wie schrijft ons nieuws? *Tijdschrift voor Communicatiewetenschap, 37*(2), 77-91.

Hopmann, D. *N.*, Elmelund-Præstekær, C., Albæk, E., Vliegenthart, R., & de Vreese, C. H. (2012). Party media agenda-setting How parties influence election news coverage. *Party Politics,18(*2), 173-191.

Humprecht, E., & Büchel, *F.* (2013). More of the same or marketplace of opinions? A cross-national comparison of diversity in online news reporting.*The International Journal of Press/Politics, 18(*4), 436–461.

Jackson, D., & Moloney, K. (2015). Inside Churnalism: PR, journalism and power relationships in flux. *Journalism Studies*, (ahead-of-print), 1-18.

Jacobs, G. (1999). Self-reference in press releases. *Journal of Pragmatics, 31,* 219–242.

Johnston, J. (2009). 'Not wrong for long': The role and penetration of news wire agencies in the 24/7 landscape. *Global Media Journal-Australian Edition, 3*(1), 1-16.

Johnston, J., & Forde, S. (2011). The silent partner: news agencies and 21st century news. *International Journal of Communication, 5,* 20.

Jones, J., & Salter, L. (2012). *Digital journalism.* London: SAGE Publications.

Jones, K.S. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation, 28(*1), 11-21.

Jonkman, J. & Verhoeven, P. (2013). From risk to safety. Implicit frames of third-party airport risk in Dutch quality newspapers between 1992 and 2009. *Safety Science, 58,* 1-10.

Joye, S. (2010). Reflections on Inter Press Service Evaluating the importance of an alternative news voice. *Global Media and Communication, 6*(1), 121-125.

Jukes, S. (2013). A perfect storm. In: K. Fowler-Watt & S. Allan (Eds). *Journalism: New Challenges* (pp. 222-241). Retrieved from https://microsites.bournemouth.ac.uk/cjcr/publications/journalism-new-challenges/

Karlsson, M. (2010). Rituals of transparency: Evaluating online news outlets' uses of transparency rituals in the United States, United Kingdom and Sweden. *Journalism studies, 11*(4), 535-545.

Karlsson, M. (2011). The immediacy of online news, the visibility of journalistic processes and a restructuring of journalistic authority. *Journalism,12(*3), 279-295.

Karlsson, M., & Sjøvaag, H. (2016). Content Analysis and Online News: Epistemologies of analysing the ephemeral Web. *Digital Journalism, 4*(1), 177-192.

Kerrigan, *F.*, & Graham, G. (2010). Interaction of regional news-media production and consumption through the social space. *Journal of marketing management, 26*(3-4), 302-320.

Kim, J. Y., Kiousis, S., & Xiang, Z. (2015). Agenda building and agenda setting in business: Corporate reputation attributes. *Corporate Reputation Review, 18*(1), 25-36.

Kiousis, S., Kim, J. Y., Ragas, M., Wheat, G., Kochhar, S., Svensson, E., & Miles, M. (2014). Exploring new frontiers of agenda building during the 2012 US Presidential election pre-convention period: Examining linkages across three levels. *Journalism Studies,* 1-20.

Kiousis, S., Popescu, C. and Mitrook, M. (2007). Understanding influence on corporate reputation: An examination of public relations efforts, media coverage, public opinion, and financial performance from agenda-building and agenda-setting perspective. Journal of Public Relations Research,19(2),147–165.

Klinenberg, E. (2005). Convergence: News production in a digital age. *The Annals of the American Academy of Political and Social Science, 597*(1), 48-64.

Korenius, T., Laurikkala, J., & Juhola, M. (2007). On principal component analysis, cosine and Euclidean measures in information retrieval. *Information Sciences, 177,* 4893-4905.

Kroon, A., & Schafraad, P. (2013). Copy-paste of journalistieke verdieping?. *Tijdschrift voor Communicatiewetenschap, 41*(3), 283.

Lee, B., Lancendorfer, K. M., & Lee, K. J. (2005). Agenda-setting and the Internet: The intermedia influence of Internet bulletin boards on newspaper coverage of the 2000 general election in South Korea. *Asian Journal of Communication, 15(*1), 57-71.

Lewis, J. M. W., Williams, A., Franklin, R. A., Thomas, J., & Mosdell, *N.* A. (2008). *The quality and independence of British journalism.* Cardiff: Cardiff University.

Lewis, J., Williams, A., & Franklin, B. (2008a). Four rumours and an explanation: A political economic account of journalists' changing newsgathering and reporting practices. *Journalism Practice, 2*(1), 27-45.

Lewis, J., Williams, A., & Franklin, B. (2008b). A compromised fourth estate? UK news journalism, public relations and news sources. *Journalism Studies,9*(1), 1-20.

Leydesdorff, L., & Hellsten, I. (2005). Metaphors and diaphors in science communication. Mapping the case of stem cell research. *Science Communication, 27*(1), 64-99.

Leydesdorff, L., & Welbers, K. (2011). The semantic mapping of words and co-words in contexts. *Journal of Infometrics, 5,* 469-475. doi:10.1016/j.joi.2011.01.008

Livingston, S., & Bennett, W.L. (2003). Gatekeeping, indexing, and live-event news: Is technology altering the construction of news? Political Communication, 20(4), 363-380.

Lopez-Escobar, E., Llamas, J. P., McCombs, M., & Lennon, *F.* R. (1998). Two levels of agenda setting among advertising and news in the 1995 Spanish elections. *Political Communication, 15*(2), 225-238.

Lucio-Arias, D., & Leydesdorff, L. (2007). Knowledge emergence in scientific communication: From "fullerenes" to "nanotubes". *Scientometrics, 70*(3), 603-632.

Maat, H. P. (2007). How promotional language in press releases is dealt with by journalists: Genre mixing or genre conflict?. *Journal of Business Communication, 44*(1), 59-95.

Maat, P.H. (2008). Editing and genre conflict: How newspaper journalists clarify and neutralize press release copy. *Pragmatics 18*(1), 85–111.

Maat, H. P., & de Jong, C. (2012). *How newspaper journalists reframe product press release information. Journalism, 14*(3), 348-371.

MacGregor, P. (2013). International news agencies: global eyes that never blink. *Journalism*, 35-63.

Macnamara, J. (2014). Journalism–PR relations revisited: The good news, the bad news, and insights into tomorrow's news. *Public Relations Review, 40*(5), 739-750.

Manning, P. (2001). *News and news sources. A critical introduction.* London: SAGE Publications.

Manning, P. (2008). The Press Association and News Agency sources. In: Franklin B (Ed.) *Pulling Newspapers Apart.* London: Routledge.

Manning, P. (2013). Financial journalism, news sources and the banking crisis. *Journalism, 14*(2), 173-189.

Marsh, K. (2013). Investigative Journalism. In: K. Fowler-Watt & S. Allan (Eds), *Journalism: New Challenges* (pp. 222-241). Retrieved from https://microsites.bournemouth.ac.uk/cjcr/publications/journalism-new-challenges/

Matthes, J. (2009). What's in a frame? A content analysis of media-framing studies in the world's leading communication journals, 1990–2005. Journalism and Mass Communication Quarterly, 86, 349-367.

Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. Journal of Communication, 58, 258-279.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective (Vol. 1). Psychology Press.*

Mazzoleni, G., & Schulz, W. (1999). " Mediatization" of politics: A challenge for democracy? *Political communication, 16*(3), 247-261.

McChesney, R. (2003). The problem of journalism: a political economic contribution to an explanation of the crisis in contemporary US journalism. *Journalism Studies, 4(*3), 307-309.

McChesney, R. (2008). *The political economy of media. Enduring issues, emerging dilemmas.* New York: Monthly Review Press.

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly, 36*(2), 176-187.

McCombs, M., & Funk, M. (2011). Shaping the agenda of local daily newspapers: A methodology merging the agenda setting and community structure perspectives. *Mass*

Communication and Society, 14(6), 905-919.

McManus J.H. (1994). Market-driven journalism: Let the citizen beware? In H. Tumber (Ed.), *News: A Reader* (pp. 180-190)*.* Oxford: Oxford University Press.

McQuail, D. (2000). Mass communication theory (4th ed.). London: SAGE Publications.

Minority employment in daily newspapers (2015). Retrieved from http://asne.org/content.asp?pl=140&sl=129&contentid=129

Mitchelstein, E., & Boczkowski, P. J. (2010). Online news consumption research: An assessment of past work and an agenda for the future. *New Media & Society, 12(7),* 1085-1102.

Moloney, K. (2006). *Rethinking public relations: PR propaganda and democracy.* New York: Routledge.

Moloney, K., Jackson, D., & McQueen, D. (2013). News journalism and public relations: a dangerous relationship. In: Fowler-Watt, K., & Allan, S. (Eds), *Journalism: New Challenges*. CJCR: Centre for Journalism & Communication Research, Bournemouth University.

Monroe, B.L., Colaresi, M.P., & Quinn, K.M. (2009). Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis, 16,* 372-403. doi:10.1093/pan/mpn018

Mosco, V. (1996). *The political economy of communication*. London: SAGE Publications.

Mujica, C., & Bachmann, I. (2015). Beyond the public/commercial broadcaster dichotomy: Homogenization and melodramatization of news coverage in Chile. *International Journal of Communication, 9,* 210-230.

Nel, *F.* (2010). *Laid off. What do UK journalists do next?* [Research report]. Retrieved from https://www.journalism.co.uk/uploads/laidoffreport.pdf

Netto mediabestedingen in 2013 opnieuw gedaald (2014). Retrieved from http://www.nielsen.com/nl/nl/insights/news/2014/netto- mediabestedingen-in-2013-opnieuw-gedaald-3-5.html [Media expenditures have decreased again]

Niewold, B., van der Werf, B., Punt, B., Yperlaan, O., Wynia, J., den Ouden, M., . . . Bronner, *F.* (2010). *Onderzoek naar mediabestedingen in Nederland.* Retrieved from http://dare.uva.nl/document/2/94939 [Study of the media expenditures in the Netherlands]

O'neill, D., & O'Connor, C. (2008). The passive journalist: How sources dominate local news. *Journalism Practice, 2*(3), 487-500.

Orăsan, C. (2009). Comparative evaluation of term-weighting methods for automatic summarization. *Journal of Quantitative Linguistics, 16*(1), 67-95.

Papieren oplage daalt verder (2015, May 12). [News article]. Retrieved from http://nos.nl/artikel/2035268-papieren-oplage-kranten-daalt-verder.html [Circulation rates decrease further]

Parmelee, J. H. (2013). The agenda-building function of political tweets. *New Media & Society.* doi: 10.1177/1461444813487955

Paterson, C. (2001). Media imperialism revisited: The global public sphere and the news agency agenda. *News in a Globalized Society, Nordicom, Gotëborg,* 77-92.

Paterson, C. (2006). News agency dominance in international news on the internet. Retrieved from http://icswww.leeds.ac.uk/papers/cicr/exhibits/42/cicrpaterson.pdf

Pedregosa, *F.*, Varoquaux, G., Gramfort, A., Michel, V.,&Thirion, B. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830

Pew Research Center (2015). Retrieved from http://www.journalism.org/2015/04/29/newspapers-fact-sheet/

Phillips Davison, W. (1974). News media and international negotiation, *Public Opinion Quarterly,* Summer.

Phillips, A. (2010a). Old sources: new bottles. In *N.* Fenton (Ed.), *New media, old news. Journalism & democracy in the digital age* (pp. 87-102). London: SAGE Publications.

Phillips, A. (2010b). Transparency and the new ethics of journalism. *Journalism Practice,* 4(3), 373-382.

Plaisance, P. L. (2007). Transparency: An assessment of the Kantian roots of a key element in media ethics practice. *Journal of Mass Media Ethics, 22*(2-3), 187-207.

Powers, M., & Benson, R. (2014). Is the Internet Homogenizing or Diversifying the News? External Pluralism in the US, Danish, and French Press.*The International Journal of Press/ Politics, 19*(2), 246-265.

Prenger, M., van der Valk, L., van Vree, *F.*, van der Wal, L. (2011). *Gevaarlijk spel. De verhouding tussen pr & voorlichting en journalistiek. Diemen: AMB.*

Preston, Paschal. 2009. *Making the News: Journalism and News Cultures in Europe.* London: Routledge.

Protess, D., & McCombs, M. E. (Eds.). (1991).*Agenda setting: Readings on media, public opinion, and policymaking.* New York: Routledge.

Quandt, T. (2008). (No) News on the World Wide Web? A comparative content analysis news in Europe and the United States. ***Journalism Studies, 9***(5), 717-738.

Rafter, K. (2014). Voices in the crisis: The role of media elites in interpreting Ireland's banking collapse. *European Journal of Communication,* 1-10.

Ragas, M. W., & Kiousis, S. (2010). Intermedia agenda-setting and political activism: MoveOn. org and the 2008 presidential election. *Mass Communication and Society, 13*(5), 560-583.

Reich, Z. (2006). The process model of news initiative: Sources lead first, reporters thereafter. *Journalism Studies, 7*(4), 497-514.

Reich, Z. (2010). Measuring the impact of PR on published news in increasingly fragmented news environments: A multifaceted approach. *Journalism Studies, 11*(6), 799-816.

Reich, Z. (2011). Source credibility and journalism. *Journalism Practice, 5*(1), 51-67.

Reich, Z. (2015). Comparing news reporting across print, radio, television and online: Still distinct manufacturing houses. *Journalism Studies,* 1-21.

Reich, Z., & Godler, Y. (2014). A Time of Uncertainty: The effects of reporters' time schedule on their work. *Journalism Studies, 15*(5), 607-618.

Rennen, A. A. M. (2000). *Journalistiek als kwestie van bronnen: Ontwikkeling en toepassing van een bron-georiënteerde benadering van journalistiek*. Delft: Eburon.

Roberts, M., & McCombs, M. (1994). Agenda setting and political advertising: Origins of the news agenda. *Political Communication, 11*(3), 249-262.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation, 60*(5), 503-520.

Salton, G. (1991). Developments in automatic text retrieval. *Science, 30*(253), 974-980.

Scholten, O. & Ruigrok, *N.* (2009). Bronnen in het nieuws. Een onderzoek naar ANP-berichten in nieuws en achtergrondinformatie in Nederlandse Dagbladen 2006-2008. In W. Jongbloed, E. Lauf, & R. Negenborn (Eds.), *Mediamonitor. Analyse en verdieping #1* (pp. 37-62). Hilversum: Commissariaat voor de Media.

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication, 50,* 93-109.

Sigal, L. V. (1986). Who? Sources make the news. In R. K. Manoff & M. Schudson (Eds.), *Reading the News* (pp. 7-33). New York: Pantheon.

Singer, J. B. (2006). Stepping back from the gate: Online newspaper editors and the co-production of content in campaign 2004. Journalism & Mass Communication Quarterly, 83(2), 265-280.

Sissons, H. (2012). Journalism and public relations: A tale of two discourses. *Discourse & Communication, 6(3), 273-294.*

Sjøvaag, H. (2014). Homogenisation or Differentiation? The effects of consolidation in the regional newspaper market. *Journalism Studies, 15*(5), 511-521.

Skovsgaard, M. (2014). A tabloid mind? Professional values and organizational pressures as explanations of tabloid journalism. *Media, Culture & Society,36*(2), 200-218.

Soroka, S. (2000). Schindler's List's intermedia influence: exploring the role of" entertainment" in media agenda-setting. *Canadian Journal of Communication, 25*(2), 211.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining.* Boston: Pearson Education.

Tiffen, R., Jones, P. K., Rowe, D., Aalberg, T., Coen, S., Curran, J., ... & Rojas, H. (2014). Sources in the news: A comparative study. *Journalism Studies, 15*(4), 374-391.

Tijdelijke Commissie Innovatie en Toekomst Pers [TCITP]. (2009). *Adviesrapport De volgende editie. Den Haag: OCW. [Advisory report 'The next edition'].*

TNO. (2011). Zijn de persbureaus te verslaan? De positie van Nederlandse persbureaus in de nieuwsketen. Retrieved from http://www.tno.nl/content.cfm?context=thema&content=inno_publicatie&laag1=897&laag2=919&item_id=860

Trilling, D., & Schoenbach, K. (2015). Investigating people's news diets: How online news users use offline news. *Communications: The European Journal of Communication Research, 40*(1), 67–91. doi:10.1515/commun-2014-0028

Tuchman, G. (1973). Making news by doing work: Routinizing the unexpected. *American Journal of Sociology,* 110-131.

Van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, *N.* (2008). Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles. *Political Analysis, 16*(4), 428-446.

Van Dam, H. (2003). *Kernthema's* (valedictory speech). [Core Themes]. TU Delft.

Van der Meer, T. G., & Verhoeven, P. (2013). Public framing organizational crisis situations: Social media versus news media. *Public Relations Review, 39*(3), 229-231.

Van der Meer, T. G., Verhoeven, P., Beentjes, H., & Vliegenthart, R. (2014). When frames align: The interplay between PR, news media, and the public in times of crisis. *Public Relations Review, 40*(5), 751-761.

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering, 13*(2), 22-30.

Van der Wurff, R. (2012). The Economics of Online Journalism. In E Siapera & A. Veglis (Eds.), *The Handbook of Global Online Journalism* (pp. 231-250). Wiley-Blackwell, Oxford, UK. doi: 10.1002/9781118313978.c*H1*3

Van Hout, T., & Jacobs, G. (2008). News production theory and practice: Fieldwork notes on power, interaction and agency. *Pragmatics, 18*(1), 59.

Van Leuven, S., & Joye, S. (2014). Civil society organizations at the gates? A gatekeeping study of news making efforts by NGOs and government institutions. *The International Journal of Press/Politics, 19*(2), 160-180.

Van Leuven, S., Deprez, A., & Raeymaeckers, K. (2013). Increased news access for NGOs? How Médecins Sans Frontières' press releases built the agenda of Flemish newspapers (1995–2010). *Journalism Practice, 7*(4), 430-445.

Van Leuven, S., Deprez, A., & Raeymaeckers, K. (2014). Towards more balanced news access? A study on the impact of cost-cutting and Web 2.0 on the mediated public sphere. *Journalism, 15*(7), 850-867.

Van Leuven, S., Deprez, A., & Raeymaeckers, K. (2015). Het journalistieke brongebruik in tijden van churnalism en Twitter. *Tijdschrift voor Communicatiewetenschap, 43*(1), 64-83.

Vermaas, K. & Jansen, *F.* (2009). *Het persbureau in perspectief: Rol, functies en kernwaarden van Nederlandse persbureaus.* Diemen: AMB Press.

Vliegenthart, R., & Roggeband, C. (2007). Framing immigration and integration: Relationships between press and parliament in the Netherlands. *International Communication Gazette, 69*(3), 295-319.

Vliegenthart, R., & Walgrave, S. (2008). The contingency of intermedia agenda setting: A longitudinal study in Belgium. *Journalism & Mass Communication Quarterly, 85*(4), 860-877.

Vogt, *N.* & Mitchell, A. (2016). Crowdfunded journalism: a small but growing addition to publicly driven journalism. Retrieved from http://www.journalism.org/2016/01/20/crowdfunded-journalism/

Waisbord, S. (2011). Can NGOs change the news?. *International Journal of Communication, 5,* 142-164.

Weaver, D., & Bimber, B. (2008). Finding news stories: a comparison of searches using LexisNexis and Google News. *Journalism & Mass Communication Quarterly, 85*(3), 515-530.

Weaver, D., McCombs, M., & Shaw, D.L. (2004). Agenda-Setting Research: Issues, Attributes, and Influences. In L. Kaid (Ed.) *Handbook of Political Communication Research* (pp. 257–81). London: Lawrence Erlbaum Associates.

White, D. M. (1950). The gatekeeper: A case study in the selection of news. *Journalism Quarterly, 27*(4), 383-390.

Whitney, D. C., & Becker, L. B. (1982). "Keeping the gates" for gatekeepers: The effects of wire news. *Journalism Quarterly, 59*(1), 60-65.

Witschge, T., Fenton, *N.*, Freedman, D. (2010). *Protecting the news. Civil society and the media.* Retrieved from http://www.carnegieuktrust.org.uk/getattachment/1598111d-7cbc-471e-98b4-dc4225f38e99/Protecting-the-News--Civil-Society-and-the-Media.aspx

World Association of Newspapers and News Publishers [WAN-IFRA] (2015). Retrieved from http://www.slideshare.net/WAN-IFRA/250515-wpt-2015-final

Number of press releases and media articles in dataset

NGOs		Corporations		Agency		Quality Newspapers		Popular Newspapers		Free Newspapers	
Artsen zonder Grenzen	254	Aegon	597	*ANP*	6142	*NRC*	3984	*Telegraaf*	3944	Spits*	775
Greenpeace	556	Ahold	223			*Volkskrant*	3944	Algemeen Dagblad	2552	*Metro**	941
KWF	183	Apple	483			Financieel Dagblad	6788				
Unicef	245	Nuon	634								
Vluchtelingenwerk	213										
WWF	1067										
Total	2518		1937		6142		14716		6496		1716

Note. *Data collected from 2004 - 1013, except where indicated *: data from 2008-2013.*

APPENDIX B – CHAPTER THREE
Cosine example

Press release Greenpeace, July 2008

Yet another French nuclear plant leak
On July 7, 30,000 liters of radio-active water leaked from another nuclear plant.

Again, a leak has been discovered at a nuclear plant in France on Friday July 18. A crack in a pipeline of the plant in Romans-sur-Isere, south of Lyon, caused the leakage of uranium. It is the second leak within two weeks at a nuclear plant. The new leak once again demonstrates that nuclear energy remains dangerous and can lead to severe contamination.

The French nuclear watchdog ASN decided on the 11th of July that the Tricastin-plant temporarily needs to be shut down. Because of a leak, 30,000 liters radio-active water spilled into the soil and the rivers Gaffiere and Lauzon. All activities in the plant have been cancelled and new safety measures are introduced to avoid future incidents.

As a consequence of the accident, all recreational areas are closed, the agriculture can no longer use water from the rivers and fishing is forbidden. The two leaks are yet another example of the risks that come along with nuclear energy. While the nuclear industry attempts to convince everyone that nuclear energy is safe, accidents like these continue to prove them wrong. Rather than wasting time on nuclear energy, we ought to invest in a real energy revolution. Greenpeace has been campaigning internationally against nuclear plants for years. Earlier this year, Greenpeace-activists protested against the realization of two nuclear plants in Finland and France.

Press release NRG, July 2008

NRG prepared for development new nuclear plants

Nuclear expertise center NRG is ready to actively contribute to expanding the share of nuclear energy in the Dutch energy mix. The recently published annual report makes clear that NRG is planning to build a new nuclear plant in Finland, a fusion reactor in France and prolonging the lifespan of nuclear plants in England. This is the concrete result of the new strategy to shift *NRG's* activities from maintaining expertise to the development of new nuclear plants. The company's new position neatly fits the government's standpoint on the role of nuclear energy in the Netherlands, as expressed in the Energy Report of June 2008. The government works out different scenarios for nuclear energy in order for the next cabinet to make a decision on the development of one or more new nuclear plants in our country. NRG views this as a promising step towards the sustainability of our national energy supply and to keep up with the latest developments around nuclear energy in Europe. The current share of nuclear energy in Europe amounts up to 30 percent and will need to be maintained to secure a safe, reliable and affordable future energy supply. This implies that in the next twenty years, 60 to 70 third generation nuclear plants will need to be build. This generation produces a smaller volume and shorter active waste and uses around 20 percent less uranium.

NRG focuses her expertise on supporting the government and the international market. The realization of new plants in the coming years involves an investment of tens of billions of euros. The time has come for the Dutch business world to position itself into a share in this.

NoordHollands Dagblad article, July 2008

NRG shifts strategy towards new nuclear plants Petten, July 15 2008. The Netherlands will build new nuclear plants in the near future. Nuclear Research & consultancy Group (NRG, Petten) is convinced of this. To avoid missing the boat, the organization has drastically changed course. In the past years, without prospects of more nuclear energy, NRG among others maintained nuclear expertise in the Netherlands. Currently, NRG focuses more prominently on developing new plants. The daughter of Energy Research Centre Netherlands (ECN) has plans for a new to build her own new plant well. NRG, owner of the high flux reactor (hfr) in the dunes, wants to build a new reactor in Petten after 2015, the so-called Pallas project.

According to NRG, it provides nuclear expertise for a large new plant in Finland, and it is involved in prolonging the lifespan of existing nuclear plants in England. In the shape of a material study, NRG also contributes to an experimental fusion reactor in Cadarache, in the South of France.

The optimism about the future is great in Petten. The atom scientists have calculated that in the period til 2030, around 70 new nuclear plants will be built in Europe. NRG is growing, although it is hard to find enough qualified personnel. President R. Stol has faith however. At the University of Delft for instance, there is an increase of students applying for nuclear studies.

Note: Mutual terms are underlined.

Cosine scores: Greenpeace – NRG .260; Greenpeace – NHD .122; NRG – NHD .394

Appendix B presents three articles: one from Greenpeace, one from NRG and one from *PZC*. Upon reading the texts, we notice that the content of NRG and *PZC* appears similar; they both discuss the organization's future strategy of building new plants. The contents of the two texts are far from identical, but the texts do show a fair degree of overlap both in terms of vocabulary (see mutual terms, underlined) as well as with regard to their meaning. This is reflected in the cosine similarity score of .394 for the two texts. The Greenpeace text, on the other hand, has less in common with the other two texts. This becomes clear from the relatively low number of mutual terms and is reflected in low similarity scores of .122 (Greenpeace-*NHD*) and .260 (Greenpeace-NRG). The fact that this latter score is higher than the previous is explained by the higher amount of terms that Greenpeace and NRG share: both mention 'nuclear energy' multiple times and refer to 'uranium', the 'future' as well as 'safety'. The news article does not include these latter terms and refers less often explicitly to energy. Because the terms 'nuclear', 'energy' and 'plant' occur frequently across all three texts, these terms have a relatively lower individual tf-idf score than for terms that occur less often.

Most prominent tf-idf terms per corpus

		Sources		Agencies		National newspapers			Regional newspapers	
	EPZ	NRG	Greenpeace	ANP	Novum	de Volkskrant	NRC	De Telegraaf	Noordhollands Dagblad	Provinciaalse Zeeuwse Courant
1	nuclear fuel	flux	nuclear waste	Greenpeace	Delta	ANP	wasWte	Delta	NRG	Zeeland
2	transport	authorities	waste	isotopes	isotopes	Europe	Europe	Essent	Pallas	Delta
3	used	isotopes	radioactive	Delta	Greenpeace	minister	cabinet	RWE	isotopes	EPZ
4	container	hoge	clean	NRG	NRG	waste	minister	cabinet	zijpe	Zeeuws
5	France	medical	activists	minister	ministry	cabinet	ANP	minister	million	Vlaanderen
6	company	research	Russia	waste	research	energy supply	environment	million	bibu	nuclear
7	COVRA	HFR	Delta	transport	construction	environment	Delta	construction	medical	CDA
8	recycling	waste	Verhagen	cabinet	EPZ	state	Greenpeace	secretary of state	dunes	Bevelanden
9	minister	production	uranium	Haag	minister	economy	state	NRG	research reactor	waste
10	measurement	events	contamination	Verhagen	cda	NRG	energy supply	environment	flux	construction
11	unit	Pallas	France	EPZ	spokesman	Greenpeace	economy	waste	ECN	Walcheren
12	waste	company	sustainable	nuclear waste	RWE	closure	construction	CDA	high	Covra
13	hague	radioactive	Chernobyl	spokesman	safety	isotopes	research	European	construction	Duiveland
14	recycling factory	European	construction	CDA	cabinet	ministry	EPZ	energy company	research	Schouwen
15	inspection	performed	Teuling	medical	Zeeuwse	European	CDA	EPZ	production	Tholen

AUTHOR CONTRIBUTIONS

AUTHOR'S INITIALS

Jelle Boumans – JB
Rens Vliegenthart – RV
Hajo Boomgaarden – HB
Damian Trilling – DT

Chapter One: The Agency Makes the (Online) News World Go Round. The Impact of News Agency Content on Print and Online News

Study design: JB. Acquisition of data: JB. Analyses and interpretation of data: JB and DT. Method development: JB and DT. Writing manuscript: JB. Critical revision of the manuscript: DT, RV and HB. Supervision of entire process: RV and HB.

Chapter Two: Subsidizing the News? Organizational Press Releases' Influence on the Agenda and Content of News Media

Study design: JB. Acquisition of data: JB. Analyses and interpretation of data: JB. Writing manuscript: JB. Critical revision of the manuscript and supervision of the process: RV and HB.

Chapter Three: Nuclear Voices in the News. A Comparison of Source, News Agency and Newspaper Content About Nuclear Energy over Time

Study design: JB, RV and HB. Acquisition of data: JB. Analyses and interpretation of data: JB and RV. Writing manuscript: JB. Critical revision of the manuscript and supervision of the process: RV and HB.

Chapter Four: Introducing cosine similarity to assess frame overlap

Study design: JB, RV and HB. Acquisition of data: JB. Analyses and interpretation of data: JB. Writing manuscript: JB, RV and HB. Critical revision of the manuscript and supervision of the process: RV and HB.

SUMMARY

The challenges and economic pressures that newspapers worldwide faced, and are still facing, have been widely documented for years. While circulation rates and advertising revenues are steadily declining, competition from other media is rampant, and a profitable online news model has not yet been found. News organizations respond to these setbacks by increasingly rationalizing the news production process: expensive activities like foreign news reporting and investigative journalism are being cut and fewer journalists are forced to produce more output, at a faster pace. To be able to cope with these demands, it has been argued that journalists become less proactive in searching for news and instead rely on 'subsidized' content provided by news agencies and by sources that aim for coverage that advances their interests. This form of journalism has been called churnalism and is believed to lead to less accurate and trustworthy news. Yet a review of the literature reveals that churnalism has been subject to a plethora of normative considerations, but received considerably less empirical attention. An additional striking observation is that the largest provider of input, the news agency, is structurally overlooked in research on news production processes. Given the centrality of the agency in the news landscape, the general scarcity of detailed academic analyses of the performance of news agencies and their exact impact on the news agenda and content is remarkable.

Relying on sophisticated automated content analyses, this dissertation explores the relationship between organizational press releases, agency content and content from newspapers and online news media in the Dutch context. The dissertation is organized as follows. Chapter One empirically investigates to what extent the news agency steers national print and online news, looking at differences both within and between these two categories. The chapter introduces two measures: The intermedia agenda-building ratio indicates what percentage of news articles is initiated by the news agency, while the churnalism index indicates to what extent a news article consists of replicated agency material. Chapter Two expands the scope of the analysis by considering the impact of subsidized content in the shape of organizational press releases on both news agency and print content. Tracing this impact for ten organizations over a period of ten years, the study investigates

whether indeed media have become more reliant on subsidized content over the recent past. A related concern expressed in the literature is that some sources are structurally more successful in accessing the news agenda than others. By comparing the influence of press releases from NGOs and corporations, this point is addressed in both Chapter Two and Chapter Three. The focus of both studies differs however: while the second chapter – like the first chapter – investigates to what extent news media reproduce literal content, the third chapter assesses similarity in themes between organizations and news media. To this end it analyzes a highly contested issue: nuclear energy. Chapter Three furthermore takes the role of the regional newspapers as well as the second largest news agency into account. Finally, Chapter Four expands the applicability of the proposed approach of automated text comparison to the field of framing research.

This empirically grounded dissertation has generated a number of important insights into the role of press releases and news agencies in the news production process. First, contrasting widespread concerns it has been shown that Dutch (print) newspapers are no copy factories: about one in ten news articles that feature an organization is based on a press release by that organization. Furthermore, when a press release is used for a news article, the news content generally contains other information as well. It is illustrative that no more than three of the 4,455 press releases included in the second study have been literally replicated by the newspapers. A second important insight is related to the role of the news agency and again contrasts a common assumption in the literature: the longitudinal datasets showed no increased presence of agency copy in the print news. The recent economic challenges thus do not appear to have led to a stronger reliance on the cost-efficient services of the agency. For online news, the picture looks radically different: here, the agency is to a great extent steering both the news agenda as well as the literal content of the online news that is not behind a paywall. This dominant role of a single news provider fuels concerns on the level of homogeneity and possible lack of diversity of viewpoints in online news. It is in this light reassuring that the second study rejects the notion of an unhealthy dependency of the news agency on PR-material as reported in some other countries. The results show that the national agency does redistribute press releases relatively more than newspapers, yet rarely verbatim. A final result worth highlighting is that the systematic comparison between NGOs and corporations indicates that NGOs are more successful than corporations in their efforts to create media coverage through distributing press releases. This is line with the view that NGOs – particularly the large international organizations - are increasingly able to cater to the media's needs for news content, for instance by presenting themselves as credible experts on societal issues.

In conclusion, the overall picture of Dutch journalism's reliance on sources that arises from this dissertation is not as dark as one would expect on basis of the literature. There is no such thing as a strong journalistic routine of copying and pasting subsidized content– neither for news agencies nor for newspapers. What has been demonstrated is the central role of the news agency in the news production process. In the case of online news, the dominance of agency content may have potential negative side-effects. First, in a society where news is increasingly consumed online, the agency has a large influence on the boundaries of public debate. If an event does not pass the agency's filter, it is less likely to be part of the online news agenda. Furthermore, the way an issue or controversy is presented online will strongly be funneled by the way the agency has chosen to present it. Although our analysis has shown that the news agency by no means can be viewed as a mouthpiece for organizations, the agency's precarious financial situation justifies a continuous close monitoring of the agency's performance. Already in 2011, ANP's president expressed concerns of the possible negative consequences the economic hardship has. The most recent financial results and rounds of layoffs indicate that the agency is still struggling to cope with the pressures – most notably caused by declining incomes from newspapers and broadcasters and the illegal spread of agency content on the internet. This dissertation has laid bare what an impact the collapse of the agency would have on, in particular, the online news landscape.

SAMENVATTING

Dat dagbladen niet de meest florissante tijden doormaken is een understatement. Wereldwijd dalen de lezersaantallen en oplagen al tijden en de primaire bron van inkomsten voor dagbladen, advertentie-inkomsten, stroomt inmiddels naar andere kanalen. Waar Nederlandse dagbladen in 1990 nog meer dan de helft van alle mediabestedingen ontvingen is dat aandeel nog geen twintig jaar later gemarginaliseerd tot 5,2 %. Behalve de structurele uitdagingen die de immense concurrentie van andere media met zich meebrengt hebben dagbladen ook conjuncturele problemen. In weinig sectoren zijn de klappen van de recente economische crisis harder gevoeld dan in de printsector: de media bestedingen van adverteerders zijn ruim gehalveerd tussen 2008 en 2013.

Critici vrezen dat de benarde positie van dagbladen negatieve consequenties heeft voor de kwaliteit van de nieuwsvoorziening. Een specifiek punt van zorg betreft de mogelijk toegenomen afhankelijkheid van informatie die wordt aangeleverd door persbureaus en PR-professionals in dienst van organisaties. Deze toegenomen afhankelijkheid zou een logisch gevolg zijn van krimpende redacties en een stijgende werkdruk. De substantiële groei en professionalisering van het 'PR-apparaat' dat de media van informatie voorziet zou verder bijdragen aan een ongelijk speelveld waarin de aanbieders van informatie domineren. De journalistieke routine van het vrijwel ongewijzigd en ongecontroleerd doorsturen van aangeleverde kopij staat wel bekend als *churnalism*. Het doel van dit proefschrift is empirisch vast te stellen in welke mate deze praktijk zichtbaar is in het Nederlandse nieuwslandschap.

Het onderzoek beschreven in dit proefschrift levert vier concrete bijdragen aan het debat over de rol van de PR-industrie en het persbureau in het journalistieke proces. Hoewel er regelmatig met zorg gesproken wordt over een ongezond grote rol en toenemende van het persbureau in de productie van het nieuws, is er verbazingwekkend weinig wetenschappelijk onderzoek verricht om deze invloed daadwerkelijk vast te stellen. De *eerste* bijdrage wordt dan ook gevormd door een gedetailleerde empirische analyse van de mate waarin nieuwsmedia vertrouwen op de berichten van het persbureau.

Om meer inzicht te genereren in de journalistieke routine van het persbureau wordt ook onderzocht in welke mate PR-berichten worden gebruikt. Om verschillende redenen wordt aangenomen dat persbureaus, meer nog dan journalisten, nadrukkelijk steunen op 'gesubsidieerde berichten' voor hun content. Zo is de tijdsdruk – een voorname verklaring voor het gebruik van gesubsidieerde berichten – een stuk hoger voor journalisten van het persbureau dan voor dagbladjournalisten. Met behulp van een grootschalige inhoudelijke vergelijking wordt vastgesteld of persbureaus inderdaad vaker en in grotere mate PR-berichten overnemen dan dagbladjournalisten. De *tweede* bijdrage komt voort uit de nationale context van de studie: waar het overgrote deel van de alarmerende studies afkomstig is uit de Verenigde Staten en Groot-Brittannië, richt dit onderzoek zich specifiek op de Nederlandse situatie. Landenvergelijkend onderzoek toont dat de staat van de journalistieke sector aanzienlijk verschilt per nationale context. Er zijn aanwijzingen dat de afhankelijkheid van informatie afkomstig van de PR-industrie en het persbureau in Nederland weliswaar toeneemt en aanleiding geeft tot lichte zorg, maar niet zo problematisch is als in eerdergenoemde landen. De *derde* bijdrage komt voort uit de unieke datasets die zijn samengesteld om de relaties te onderzoeken. Aangezien de schaarse empirische aanwijzingen voor churnalism voortkomen uit kwalitatieve of kleinschalige kwantitatieve analyses, is grootschalig onderzoek hard nodig. Een uniek onderdeel van het proefschrift in dit verband is het longitudinale karakter van twee van de drie datasets. Dit maakt het mogelijk om vast te stellen of er inderdaad sprake is van een toenemend gebruik van persberichten en kopij van het persbureau. Het ontwikkelde onderzoeksinstrument dat de verwerking en analyse van dergelijke grote datasets mogelijk maakt vormt de *vierde* bijdrage van het project. Automatische inhoudsanalyse-technieken maken het mogelijk om met beperkte financiële middelen grootschalige analyses uit te voeren. Een ander voordeel is dat dergelijke technieken in staat zijn om patronen in data te identificeren die traditionele analysemethoden niet kunnen identificeren, of alleen met grote moeite. Dit proefschrift illustreert dit laatste punt door de mate van tekstuele overlap inzichtelijk te maken aan de hand van gestandiseerde maten.

OVERZICHT VAN DE STUDIES EN BEVINDINGEN

Hoofdstuk 1 richt zich op de mate waarin nationale kranten en nieuwssites gebruik maken van kopij van het Algemeen Nederlands Persbureau (ANP). De dataset omvat alle in 2014 verschenen print- en online nieuwsberichten van de *Volkskrant*, *De Telegraaf* en *Metro*, alsmede de nieuwsberichten van de meest bezochte nieuwssite, nu.nl. Daarnaast zijn alle in 2014 verschenen berichten van het ANP

geselecteerd. Gezien de interesse in actualiteiten zijn verschillende typen berichten en katernen uitgesloten, waaronder columns, redactioneel commentaar, ingezonden brieven en katernen als reizen en lifestyle. Hierna bleven 119,452 ANP-berichten en 247,161 nieuwsberichten over die met elkaar zijn vergeleken door middel van een geautomatiseerde inhoudsanalyse. Het analyse-instrument maakt gebruik van de cosinus gelijkheid, een vergelijkingsmaat die in de informatica (*information retrieval*) veel wordt toegepast. De cosinus maat geeft op een schaal van 0 tot 1 aan in welke mate twee documenten overeenkomen: hoe meer woorden van twee documenten overeenkomen, hoe hoger de cosinus score. Op basis van deze maat kan worden vastgesteld *of* er sprake is van overlap tussen twee teksten, en *hoe sterk* deze overlap is. Resultaten wijzen uit dat online nieuwssites in sterke mate gebruik maken van ANP-berichten: gemiddeld 66 procent van de online nieuwsberichten komt voort uit een persbureaubericht. De churnalism-score voor online nieuws duidt erop dat het gemiddelde online artikel voor het overgrote deel uit letterlijk overgenomen ANP-inhoud bestaat en weinig toegevoegde informatie bevat. Voor print nieuws ligt deze score beduidend lager, waarbij opgemerkt moet worden dat de gratis krant *Metro* significant meer content letterlijk overneemt dan de betaalde kranten.

Het tweede hoofdstuk richt zich op een stap eerder in het nieuwsproductieproces door ook de invloed van PR-berichten in de analyses te betrekken. Het is daarmee één van de weinige studies die inzicht verschaft in de triangulaire relatie tussen bronnen, persbureaus en kranten. Het onderzoek maakt gebruik van dezelfde methode als hoofdstuk 1 en vergelijkt 4,455 persberichten van tien organisaties met 6,142 ANP-berichten en 22,928 krantenartikelen over deze organisaties over een periode van tien jaar (2004-2013). Het maakt daarnaast onderscheid tussen de afzender van het persbericht (nongouvernmentele organisaties (NGO's) versus bedrijven) en type krant (kwaliteit, populair en gratis). De resultaten wijzen uit dat gemiddeld één op de tien geanalyseerde krantenberichten gebaseerd is op een persbericht. Zoals verwacht ligt de mate van overname in het geval van het persbureau iets hoger (16 procent). Dit impliceert dat het persbureau, wanneer het over een organisatie bericht, dit relatief vaker doet op basis van een persbericht van deze organisatie. Nieuwsberichten over NGO's zijn significant vaker geïnitieerd door de organisatie zelf dan nieuwsberichten over bedrijven. Ervan uitgaande dat organisaties over het algemeen persberichten uitbrengen met het streven publiciteit te genereren, kan gesteld worden dat NGO's beter slagen in dit streven dan corporaties. Daarnaast vertonen mediaberichten die gebaseerd zijn op persberichten van NGO's meer overlap met het oorspronkelijke persbericht dan mediaberichten die gebaseerd zijn op persberichten van bedrijven.

Hiermee is echter niet gezegd dat media de aangeleverde informatie integraal overnemen: de gemiddelde churnalism-score tussen persberichten en mediaberichten duidt erop dat zowel het persbureau als de dagbladjournalist informatie toevoegt. De dagbladjournalist doet dit in grotere mate dan het persbureau. Het letterlijk overnemen van persberichten is zowel in het geval van kranten als het persbureau vrijwel niet aangetroffen.

In hoofdstuk drie dient het debat rond kernenergie in de periode 2003-2012 ter illustratie van de wijze waarop PR-berichten hun weerslag vinden in de nieuwsmedia. Daarbij wordt zowel de inhoud van de persberichten van een prominentie tegenstander (Greenpeace) als voorstanders van nucleaire energie (energieproducent EPZ en nucleair dienstverlener NRG) in beschouwing genomen. Bijzondere aandacht gaat daarnaast uit naar de mate waarin de twee nationale persbureaus, ANP en Novum (per 2015 ingelijfd door het ANP), alsook regionale kranten informatie overnemen van deze persberichten. Op basis van de cosinus gelijkheidsmaat kan worden vastgesteld dat mediaberichtgeving omtrent kernenergie sterker overeenkomt met de inhoud van persberichten van Greenpeace dan van de bedrijven. Anders gezegd: de thema's die Greenpeace benadrukt (nucleaire risico's, rampen, afvalproblematiek) zijn meer zichtbaar in de media dan de thema's van het bedrijfsleven (veiligheid, voordelen, vooruitgang). Een uitzondering wordt gevormd door het regionale *Noordhollands Dagblad*, wiens berichtgeving nauwelijks overeenkomt met Greenpeace, maar wel een relatief sterke overlap vertoont met de persberichten van het lokaal actieve bedrijf NRG. Een vergelijking tussen de twee persbureaus leert dat de berichten van 'nieuwkomer' Novum aanmerkelijk meer overeenkomsten vertonen met de persberichten van alle drie de organisaties dan het geval is voor ANP.

Hoofdstuk vier maakt gebruik van dezelfde dataset als hoofdstuk drie en heeft een methodologische focus. Gedemonstreerd wordt dat de cosinus gelijkheidsmaat een waardevolle aanvulling vormt op de 'impliciete frame' benadering, een geautomatiseerde framingsanalysetechniek. Het idee achter deze techniek is dat de betekenis van een tekst tot stand komt door combinaties van woorden die voorkomen in een tekst. Zo staan de woorden 'wind', 'biomassa', en 'oplossing' symbool voor het 'alternatieve energie' frame dat Greenpeace hanteert in een deel van haar persberichten. Door voor elke collectie van teksten de meest voorkomende frames vast te stellen kan een inzicht worden verkregen in de wijze waarop verschillende organisaties en media over een issue praten. De cosinus maat maakt het vervolgens mogelijk de mate van overeenkomst tussen de verschillende tekstcollecties ook

in cijfers uit te drukken. De resultaten tonen aan dat de twee benaderingen elkaar inderdaad aanvullen: hoe hoger de cosinus score, hoe groter de kans dat twee tekstcollecties gelijksoortige frames bevatten.

CONCLUSIES

Uit deze dissertatie kan een aantal belangrijke conclusies worden getrokken over de relatie tussen persberichten, persbureaus en nieuwsmedia. Ten eerste is – in tegenstelling tot wat verwacht zou worden op basis van de literatuur – aangetoond dat Nederlandse kranten verre van kopieerfabrieken zijn. Eén op de tien krantenartikelen waarin een organisatie centraal staat, is gebaseerd op een persbericht van deze organisatie. Daarnaast is vastgesteld dat er praktisch geen sprake is van het grotendeels of geheel ongewijzigd publiceren van PR-berichten. Ook is in de periode 2004-2013 geen toename in het gebruik van ANP-kopij of persberichten geconstateerd, wat erop duidt dat de slechtere financiële omstandigheden vooralsnog niet hebben geleid tot de opkomst van een 'knip- en plakcultuur' in de Nederlandse printsector. In het online nieuwslandschap daarentegen lijkt een dergelijke cultuur wel te bestaan. Niet alleen is vastgesteld dat driekwart van de online nieuwsberichten gebaseerd is op ANP-berichten, ook duidt de mate van overlap erop dat de ANP-berichten over het algemeen ongewijzigd of ingekort als nieuwsberichten worden gepresenteerd. Het feit dat driekwart van het online nieuws afkomstig is van één bron kan om meerdere redenen zorgelijk genoemd worden. Ten eerste impliceert het dat de verscheidenheid aan onderwerpen en invalshoeken die aan bod komen, grotendeels beperkt blijft tot de keuzes die het persbureau maakt. Het roept vragen op over de mate van pluriformiteit van het online nieuws. In dit licht is het een geruststellende constatering dat het persbureau niet voldoet aan het beeld van 'PR-doorgeefluik' zoals in de literatuur wordt geschetst. De analyses in deze dissertatie wijzen uit dat het persbureau weliswaar relatief vaker gebruik maakt van PR-berichten voor hun artikelen, maar dat het ANP-bericht slechts zeer zelden uitsluitend of overwegend bestaat uit de inhoud van het PR-bericht. Het is echter allerminst een vanzelfsprekendheid dat het persbureau in staat blijft om de kwaliteit van haar berichtgeving te waarborgen. Verschillende factoren, waaronder dalende inkomsten en de illegale verspreiding van content online, hebben de continuïteit van het persbureau onder druk gezet. Dit proefschrift heeft de aanzienlijke invloed van het persbureau op met name het online nieuwslandschap aangetoond. Geconcludeerd kan worden dat het bestaan van een goed functionerend persbureau van groot belang is voor de nieuwsvoorziening in Nederland.

ACKNOWLEDGEMENTS

First off, I'd like to emphasize I'm very grateful that I've been given the chance to conduct my personal research project the past years. It's a privilege that not many are given, and if this thing called life wouldn't have gotten between me and the project so often, I'm pretty sure this would've have been a terrific piece of work. Alas, things worked out differently than anticipated most of the time. Conducting a PhD project involves quite a bit highs and lows. Luckily there have been plenty more highs than lows. I will get back to the highs in a minute; I'll start with the lows here. You can feel pretty isolated when you're stuck in the twenty-seventh version of that *damn* dataset and you don't have a clue in what previous version exactly things went so terribly wrong. You realize the project has taken a hold on you the moment you're getting lost in different aggregation levels in your dreams ('Damn, how do I get back to y,q from here?'). And after yet another critical rejection of my first year paper in which I proposed that brilliant, novel research technique, I felt a bit like Dirkjan promoting his latest gadget (Figure 1).

During those occasional frustrating periods there were many people who cheered me up – foremost my family, promotors and closest friends at ASCoR. But rather than mentioning every individual person that has been important to me these years, I'd like you to do the math and find out how much you contributed to this dissertation by using the Grand Gratitude Grading System.

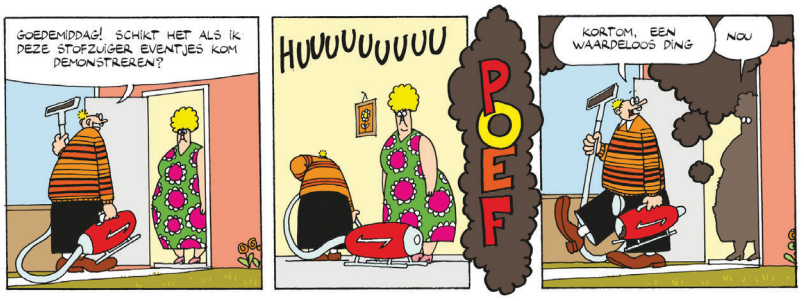


Figure 1. Dirkjan promoting his vacuum cleaner.

JELLE'S GRAND GRATITUDE GRADING SYSTEM

1. First things first: If you've been a member of the Doctorate Commission, please give yourself 50 points. Thank you for contributing to this special day.

2. If you've been my coach (some would say 'supervisor', but you've meant much more to me than just that) give yourself 100 points. Add 2 for every Skype-call, 3 for every face-to-face meeting, and .5 per email (I've counted them for you, at least 388). I'm very grateful for your patience and faith throughout all those years. Rens, 100 bonus points for your unlimited support during the final two months.

3. If you're my paranymph, you get 30 points for the job. Add 100 for being such close companions the past years.

4. If you're a colleague at ASCoR, give yourself 20 points. It is you who make ASCoR that inspiring organization that I love working for so much.

5. If you're a (junior) lecturer or in any other way involved in teaching or the administration at the Department, it goes without saying that you earn at least 20 points as well. Add 5 if you're a study advisor/drinking companion of the recent past.

6. If you've been sharing the PhD-ride, give yourself 25 points. If we've been roomies at the Bushuis, add 20 points. Corporate PhD'ers of the first hour are also eligible to a 20 point bonus.

7. Add .5 points for every time we played table football together and another .3 for every time you let me and Toni win (yes, that's an awful lot of points coming your way, Jasper and Guus. Congrats!)

8. If we've shared good times together at one of the Etmaal or ICA conferences, give yourself 25 pts.

9. Add 10 when you've been in that preconference in Seattle. I'm pretty sure even the great Dalai Lama would've felt the same way we did that day: without a clue.

10. Add 10 when you were there that night in Seattle when my wallet got nicked. Was a good band and a great night! You can give me back my wallet now.

11. Add 10 when you've been sharing the good times in Puerto Rico.

12. Add 30 when you've been sharing the bad times in Puerto Rico. Add 50 when you came out stronger.

13. Add .5 for every toast we made, whether in Leuven, Nijmegen, Puerto Rico, Seattle or Antwerp. Oh and of course Lisbon: Add .5 for every mojito or glass of port (stop counting at 10).

14. Give yourself 5 points if you've been attending one of my many conference sessions. (Yes, that's five a person for all twelve of you!)

15. Give yourself 10 points if you've enlightened my stay in picturesque Ljubljana during the summer method course.

16. Give yourself 7 points if you've edited or reviewed my articles. Add 3 for every useful suggestion to improve my paper, 3 if you gave it an R&R and another 3 for accepting it.

17. Give yourself 20 points if we've been in a PhD-club together.

    a. Add 20 if you've been chairing those sessions.

    b. Add 2 for every critical remark that improved my work.

    c. Add 1 for every joke on statistics. I know they're not easy to come up with.

18. You earn 5 points if you consider yourself an Engeltje. Add .5 for every Friday you've been at the Engelbewaarder. Alright, add 2 for CREA and 3 for Koosje. If you've been aboard of one of the infamous Engeltjes Research Missions, add 10 points.

19. If you've the one responsible for the software tool that kept me off the streets for three years straight, give yourself 100 points (and write a python script that transfers 3 points in your direction every time I use the software).

20. If you've spent time coding or collecting data for me, give yourself .5 points for every hour.

21. If you've taken care of all the administrative tasks that come along with the PhD-ship, give yourself 20 points and add five for every solved last minute request.

22. 149 points to the man who helped me create this book. Turned out quite nice, didn't it?

23. Give yourself 10 points if we've been in an NESCOR Dissertation Award Jury together. Add 5 if we've handed out the award together (or planned to and things worked out differently).

24. Give yourself (x(-b±√(b^2-4ac))/2a) *.031 points if you're that math teacher in high school that assured me I'd never pass the exams. If you've been my partner in crime since high school or even elementary school, give yourself 50 points. Who could've thought back then I'd end up as dr.!

25. Give yourself 50 points if you're the spiritual father of Dirkjan, a daily source of inspiration. Thank you for allowing Dirkjan to make his first academic appearance ever in my dissertation.

26. 20 points for every Nautiliaan. Add 80 if you've been at my side during my years as chairman. I still haven't given up faith on being on the pitch again with you guys one fine day.

27. If you're family, give yourself 10,000 pts. Your support and love means the world to me. I feel blessed knowing I can always rely on you.

28. Add 20 points for each time you took care of the children when I was stressing about a yet another deadline. Most of the time I missed them anyway but it would've been much worse if it wasn't for you. Thank you.

29. If you're the love of my life and mother of my children, give yourself 10,000 points and a tray of Haagen Dasz. Add 20 points for every day you had to do triple shifts with the pukkies because there was yet another super important deadline or conference. We've been through the best times and the worst times together and are still standing strong. I unconditional love you and can't wait to see what adventures lie ahead for us.

30. If you're my child, have mommy give you 10,000 points. Thank you for making every day so much more beautiful. If you're shining down on us from above, never mind those points. You'll forever be surrounded with a million stars. We are still with you.

WHAT'S YOUR SCORE?

25-49 points:

See, I knew I've contributed in some way to this dissertation. It's got my imprint, in a very subtle way. Glad he has noticed in all the chaos.

50-149 points:

Wow, Jelle really holds me in high regard. I never realized I meant that much to him. Maybe he should've said it a bit more often. Well, at least I've got it on paper now.

150 points and higher:

It's pretty clear that guy would've never gotten that title without me. He might as well have ended up working at the pet store selling catnip if it wasn't for the spiritual guidance, love, and incredible patience I displayed. Now that I think of it, shouldn't my name be somewhere on that fancy piece of paper? Surely there's a spot where I can sign my name, I know he wouldn't mind...

ABOUT THE AUTHOR

Jelle Boumans (October 28, 1982) started his academic career as a bachelor student of Communication Science at the University of Amsterdam (2003—2007). From January—May 2006 he studied at the Simon Fraser University in Vancouver, Canada. After graduating, Jelle continued his studies in the Research Master program in Communication Science at the same university. During this period he specialized in political communication and wrote a master's thesis about media personalization of politicians under the supervision of Hajo Boomgaarden and Rens Vliegenthart. The thesis was published in *Political Studies* in 2013. Jelle also co-authored a book chapter about the same subject with his supervisors, which was published in *Political Communication in Postmodern Democracy* (2011). After obtaining his degree in 2010, Jelle worked as a lecturer at the Communication Science Department for two years. In the second year he wrote a proposal for a PhD project on the relationship between sources, news agencies and news media that was selected for funding as part of the NWO Graduate Program. The research was conducted within the Corporate Communication Group of the Amsterdam School of Communication Research under the supervision of Rens Vliegenthart and Hajo Boomgaarden and has been accepted for publication in *European Journal of Communication* and *Tijdschrift voor Communicatiewetenschap*. The latter publication was awarded with the Young Scholar Award for Best Article 2014. Jelle has presented his research at several international conferences, including the annual conference of the International Communication Conference (ICA) and the biannual conference of the European Communication Research and Education Association's (ECREA). In his final year as PhD candidate he also wrote a methodological paper on automated content analysis techniques for communication and journalism scholars together with Damian Trilling, which has been published in *Digital Journalism*. Jelle is currently working as Assistant Professor of Corporate Communication at the University of Amsterdam.

PUBLICATIONS

Boumans, J.W., Vliegenthart, R., & Boomgaarden, H.G. (forthcoming). Nuclear voices in the news: A comparison of source, news agency and newspaper content about nuclear energy over time. *European Journal of Communication*.

Boumans, J.W., & Trilling, D.C. (2016). Time to take stock of the toolkit. An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism, 4*(1), 8-23.

Boumans, J. W., & Vliegenthart, R. (2014). 'Safety first' versus 'op de barricaden'. *Tijdschrift voor Communicatiewetenschap, 42*(4).

Boumans, J. W., Boomgaarden, H. G., & Vliegenthart, R. (2013). Media Personalisation in Context: A Cross-National Comparison between the UK and the Netherlands, 1992–2007. *Political Studies, 61*(S1), 198-216.

Vliegenthart, R., Boomgaarden, H. G., & Boumans, J. W. (2011). Changes in political news coverage: Personalization, conflict and negativity in British and Dutch newspapers. In *Political communication in postmodern democracy* (pp. 92-110). Palgrave Macmillan UK.

Journalists are increasingly blamed for the uncritical recycling of subsidized material in the form of press releases and news agency copy. This practice is called churnalism and is believed to compromise journalism's autonomy and threaten the quality of the news. While the context – rampant competition, decreasing newspaper circulation rates and advertising revenues, shrinking newsrooms, failing online business models – is well documented, it remains an empirical question what the consequences of these developments are for journalists' reliance on 'subsidized content'. A striking observation is also that the largest provider of input, the news agency, is structurally overlooked in research on news content. Filling these voids, this dissertation employs automated content analyses to assess the relationship between press releases, news agency copy and print and online news in the Dutch context.