# ABSTRACT

Title of Document: COMPUTATIONAL METHODS APPLIED TO MASS COMMUNICATION RESEARCH: THE CASE OF PRESS RELEASE CONTENT IN NEWS MEDIA

Sergey Golitsynskiy, Ph.D., 2013

Directed By: Christopher Hanson, Associate Professor, Philip Merrill College of Journalism

In this dissertation, I apply a variety of computational methods to explore new approaches to investigate the problem of news media's use of press release content.

Being used by the public relations industry in an effort to influence the media agenda, press releases often promote the organization's viewpoint on issues. Journalism scholars have expressed numerous concerns over news media using such content as a source, often without attribution.

A review of previous research has revealed a number of shortcomings, with the main problem being the lack of a reliable methodology to establish a connection between a press release and an article, which is essential for such research. This deficiency is explained by the need for in-depth textual analysis on the one hand, and the requirement

of large representative samples on the other – which is near impossible to achieve using traditional methodological approaches used in journalism research.

I propose using computational methods to address this problem. I use computation to extract large amounts of text from web sites, transform loosely structured text into well-formatted data, and reduce a data set consisting of 6,171 press releases and 48,664 related news articles to a sample of 1,643 press release/news article pairs, showing reasonable evidence that each of the press releases has been used as a source by a corresponding news article. Such evidence is established through verbatim text matches of sufficient length.

I use the constructed data sample to investigate the extent to which press release content is used by news media verbatim, how such content is used and whether proper attribution is made identifying the true source of the news. Although my findings suggest that the problem of press release content might be not as severe as presented in previous research, due to the limitations of verbatim text matching, it might be also possible that such practice remains undetected, with all content borrowed from press releases appearing in news media in paraphrased form.

Finally, my investigation leads to a discovery of a "smoking gun" – a striking example of PR influence in the form of a corporation "manufacturing" statements, getting elected officials to repeat them, and the media reporting them as a regular news story.

# COMPUTATIONAL METHODS APPLIED TO MASS COMMUNICATION RESEARCH: THE CASE OF PRESS RELEASE CONTENT IN NEWS MEDIA

By

Sergey Golitsynskiy

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Associate Professor Christopher Hanson, Chair
Professor Maurine Beasley
Assistant Professor Kalyani Chadha
Associate Professor Ira Chinoy
Professor Susan Moeller

Professor Philip Resnik, Dean's Representative

ACKNOWLEDGEMENTS

I thank my advisor, Dr. Christopher Hanson, for never doubting me – regardless of how daunting the research goals of this dissertation may have appeared. Dr. Hanson's unconditional support, his insightful comments, and the numerous conversations we have had on connecting two very different fields – mass communication and computer science – have been key to my success.

I thank Dr. Maurine Beasley for guiding me through the history of journalism and communication studies, for helping me find the right context for my methodological ideas – which eventually turned into a research proposal, and then a dissertation – thus paving the way for my academic career. The encouragement and advice Dr. Beasley has offered me over the years – from the very first time we met in class during my first semester, to my comps, proposal, and defense – is priceless.

I thank Dr. Philip Resnik for introducing me to computational linguistics, for helping me through my first steps in computational text analysis, and for providing me with access to all the resources of the Computational Linguistics and Information Processing Lab. The help Dr. Resnik offered me over the years, as well as his in-depth reviews of my research proposal and dissertation manuscript were most helpful in many ways, but especially in establishing the validity of the research design I used in my dissertation.

I thank Dr. Susan Moeller for being a mentor, colleague, and friend. Dr. Moeller's trust in my abilities, her continuing support and encouragement are hard to quantify. I am forever grateful for the experience I gained from serving as a TA for

the Media Literacy course, co-teaching the Research Methods class, and working on the numerous fun research projects we have had over the years.

I thank Dr. Ira Chinoy for his thoughtful comments on my dissertation, as well as my overall research interests. The frequent conversations we have had over the past years on computational journalism and how technology can be applied to the study of media have been both useful and inspiring; they definitely helped me find my research niche.

I thank Dr. Kalyani Chadha for introducing me to research methods and helping me understand how different research methodologies may complement each other. Most importantly, I thank Dr. Chadha for helping me find the right balance between disciplines in my interdisciplinary dissertation project, and for encouraging me to focus on methodology, as it has been my primary interest.

There are many people besides my committee whom I would like to thank for helping me in my research, which has led to this dissertation, and, ultimately, to my degree. The following list is, most likely, incomplete – and I apologize to those whom I have not mentioned.

I thank Dr. Douglas Oard for trusting my programming and data analytical skills and hiring me at UMD's E-Discovery Lab at the College of Information Studies; thus giving me an opportunity to learn to work with very large data sets, while being part of a team engaged in cutting-edge research in e-discovery.

I thank Dr. William Dardick and Dr. Hong Jiao – who helped me gain a profound understanding of the fundamental concepts in statistics – which has changed

the way I look at data analysis and has put any advanced statistical methods I choose to explore within my grasp.

I thank Leslie Walker – for her continuous trust in my programming abilities. The software development projects Leslie has offered have provided me with exciting problems to work on, as well as generous financial support – which has been a tremendous help as I worked on my dissertation research.

I thank my fellow Ph.D. students and lab mates for their help, their friendship and support. I especially thank my friend and fellow Ph.D. student Jim Baxter and wish him all the best in his dissertation research and his future career.

I thank Dr. Sarah Oates – for her kind words of support, which reinforced my confidence in my research findings and conclusions.

I thank Clint Bucco for his friendship, and his continuous support with all things IT. Thanks to Clint I had access to all the computational resources our College could provide.

I also thank my colleagues and professors at the University of Northern Iowa, who helped lay the foundations of this dissertation. I thank Dr. Dean Kruckeberg for helping me understand the complex nature of public relations and guiding me through a critical study of the field. I thank Dr. Eugene Wallingford for helping me discover the magic of computer science – which opened more possibilities to me than I could've imagined, computational text analysis in media research being one of them.

I thank Dr. John Somervill – for being my friend and mentor, and for helping me make sense of my research, scientific method, as well as life, the universe and everything.

This dissertation would have not been possible without my wife Alina, who is a more accomplished academic than I will ever be.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# INTRODUCTION

In this dissertation, I apply a variety of computational methods to explore new approaches to investigate the problem of news media's use of press release content. This critically important issue has been prominent in mass communication scholarship for at least a century, yet it has not been investigated in sufficient detail due to a variety of methodological issues. In this study, I show how using computation to collect, process, and analyze very large amounts of textual data can be utilized to address some of these issues. Specifically, I use computational methods to extract large amounts of text from web sites, transform loosely structured text into well-formatted data, and reduce a very large data set to a sample of most relevant items suitable for both automated and manual textual analysis. Through conducting such analysis, I investigate the extent to which press release content is used by news media verbatim, how such content is used and whether its positive tone affects the overall tone of the news article, and whether proper attribution is made identifying the true source of the news. Finally, my investigation leads to a discovery of a "smoking gun" – a striking example of PR influence in the form of a corporation "manufacturing" statements, getting elected officials to repeat them, and the media reporting them as a regular news story.

This is a preliminary study, and its goal is to explore the possibilities offered by applying computational methods to mass communication research; and, through that, to contribute to a broader understanding of the problem of press release content in news media.

## Problem Overview

Press releases, conceptualized in mass communication scholarship as information subsidies, have been shown to play a critical role in the media agenda-building process. Being used by the public relations industry in an effort to influence the media agenda, press releases are often described as "prepackaged" information, carefully constructed to promote the organization's viewpoint on issues, with the ultimate goal of causing news coverage to reflect a similar viewpoint.

Journalists, press critics and journalism scholars have observed that press releases are typically worded in a way that strongly favors the organization that provides them. Thus, by using them as a news source – especially when copying their text verbatim – news media incorporate such wording, carefully constructed by PR, into a news account, which is expected to be impartial. This, not surprisingly, results in mostly positive news accounts, which serve the interests of the organizations which supply that content.

Furthermore, scholars and press critics have pointed out that such news articles often provide no attribution to the press release as their source and, therefore, appear to be unbiased, thus misleading the audience by hiding the true agenda behind the story. As a result of such practice, news media unwittingly serve the needs of the public relations industry and its clients instead of serving the public.

## Methodological Issues of Previous Research

A thorough investigation of the extent of such practice – i.e., of the extent of press release use by news media, especially with little or no edits – is, therefore, critically important for mass communication scholarship, as well as for preserving the media's

ability to produce independent news accounts, serving as society's fourth estate. However, a review of previous research on press releases has revealed a number of troubling shortcomings. Most importantly, the majority of such research originates in the field of public relations and usually serves its practical needs. Thus, numerous studies have been conducted on how to make a press release more efficient as a public relations tool, including identification of criteria by which press releases are evaluated by journalists.

Other studies have examined the relationship between press releases and the news coverage they are assumed to have caused by comparing them in terms of quantity and content, using both quantitative and qualitative methods. However, such studies rarely provide any reliable evidence establishing the relationship between a press release and a news article. The few studies which provide such evidence by conducting in-depth textual analysis, often comparing individual words and phrases from the two texts side-by-side, use very small samples which are not representative of the overall population and, therefore, cannot be easily generalized. Such studies provide a detailed analysis of how press release content is used and transformed by journalists in the case of a given press release or company, or a type of corporate announcement.

Studies which use larger samples typically rely on a set of keywords used to retrieve news articles from LexisNexis or a similar source, with such keywords being the only connection between a given press release and a news article – which, however, does not indicate that a press release has been used as the source. Keyword search locates texts with relevancy defined in terms of content similarity, which is established based on the provided keywords. For example, if one searches a news database for the name of a

recently launched popular consumer product, chances are that, besides abundant news coverage, the results will include a product launch press release. Clearly, considering the breadth and richness of today's media environment, as well as the scope of tools available to and used by corporate communications, one simply cannot claim that one press release caused all the news coverage of such an event.

Content similarity does not imply causality, or content dependency – i.e., similarity alone does not indicate that the text of a given press release has been actually used as a source for the news article. A newsworthy event or fact can be announced through means other than a press release – such as an interview, a press conference, or a website – not to mention the numerous communication tools provided by social media and, undoubtedly, used by organizations to communicate their messages to the public. In other words, there are numerous ways in which a journalist can obtain information on a topic that also happens to be announced through a press release. Comparing topics is simply not enough; a better, more reliable method of establishing a relationship between a press release and a news article is needed.

## Pros and Cons of Using Verbatim Text Matches to Detect Relationships

An obvious way to establish that an article is, indeed, based on the text of a press release – i.e., that the journalist actually used the press release in some way while writing the article – is to locate text which occurs in both the article and the press release. This can be handled by relying on verbatim matches, or by using a more flexible approach and matching paraphrased text.

A verbatim text match between a press release and an article, sufficiently meaningful in terms of length and content, enables a researcher to tie a specific press release to a specific article beyond reasonable doubt. However, this approach has certain limitations. There are various ways of identifying a verbatim match, yet none of them offer a perfect balance between a definitive clue pointing to a connection between two texts, and the flexibility of a partial match suggesting the possibility of such a connection. A sequence of words, such as a sentence, may point to a connection between a press release and a news article in one case, but will fail to uncover the same connection in another case, where the sentence is broken up by the insertion of one single word. It is reasonable to expect that someone copying parts of a press release into a news article they are writing, would attempt to make the similarity of the texts less obvious; breaking up a sentence is the simplest way to do that; paraphrasing parts of the copied text is the next logical step. Thus, by using verbatim matches, a researcher is bound to miss a connection between a press release and a news article if the text borrowed from the press release is completely paraphrased.

One solution to this limitation is to take into account paraphrased text. However, without a verbatim match, it may be harder to argue that the text of the news article is, in fact, a paraphrase of the press release and not some other source. There are numerous ways in which content created by a corporation may reach a journalist, and a press release is but one possibility. Paraphrased text suggests a similar topic – a topic that might have originated from a press release, an interview, a press conference, or social media – the scope of possible sources is simply too broad to tie an article to a specific press release. Paraphrased text can be used to answer broader questions about public relations content

in general. However, using it to establish a definitive connection between a specific press release and a news article for the purpose of examining the use of press release content – not PR content in general – appears to be less reliable than using verbatim matching. Therefore, in this dissertation, I rely primarily on verbatim text matches as evidence of a direct relationship between a press release and a news article.

## Computational Methods and the Goals of This Study

Discovering explicit connections between press releases and news articles, or any texts for that matter, will be shown to be a very difficult task, which cannot be accomplished manually – regardless of whether the matches are exact (i.e., verbatim), or partial (i.e., paraphrased). This constitutes a methodological obstacle which explains the reason why such studies have not been conducted. In this dissertation, I argue that such obstacles can be overcome with the application of computational methods.

I will describe how computation can be used to process a very large data set and detect relationships between press releases and news articles through locating verbatim text matches of sufficient length. Thus, the significance of using computation in this study lies primarily in the reduction of a very large data set consisting of press releases and news articles exhibiting a similarity in topics – which is not enough to draw any conclusions about a causal relationship between two texts – to a much smaller sample, consisting of press releases and news articles, explicitly tied to each other by sequences of matching text – which makes this sample much more reliable in comparison with the initial data set. This sample, referred to as a relevance sample, contains the data most relevant to the specific research questions raised in this study. To the best of my

knowledge, this sampling approach has not been used within the field of journalism and mass communication research.

This dissertation has two distinct goals. One is to address some of the concerns expressed by journalists, press critics and journalism scholars, as well as the shortcomings identified in previous research, thus, contributing to a broader understanding of the nature and scope of news media's use of the press release. The other goal is to explore the possibilities offered by applying computational methods to the problem of press release content in news media, and to demonstrate the utility of this approach for journalism and mass communication research in general.

I address these goals by using an adequate sample, representative of US news media and public relations sources alike; computing verbatim text matches, thus reducing the initial data set to a relevance sample containing the most critical data, consisting of pairs of press releases and news articles which exhibit a direct relationship with each other; and investigating this data with regard to language, content, and proper attribution of news to public relations sources.

The primary limitation of my approach is reliance on verbatim text matches, which would have prevented me from measuring the degree of press release influence on news content. However, that is an acceptable limitation, for the goal of this study is not to measure public relations influence, but to demonstrate how computational methods can be used to collect, and then reduce a very large data set to a relevance sample – thus, putting such data within the grasp of a qualitative mass communications researcher. Through this, I explore the possibilities offered by applying computation to mass

communication and journalism research, while contributing to a broader understanding of the problem of press release content in news media.

## Research Questions

The research questions posed in this dissertation address the extent to which press release content is used by news media verbatim, how such content is used and whether its positive tone affects the overall tone of the news article, and whether proper attribution is made identifying the true source of the news. All research questions are based on the constructed relevance sample consisting of pairs of press releases and news articles, which will be identified in the data set construction phase of the study.

**RQ1.** Given a press release, which is used as a source for a news article, what is the proportion of the press release text used without any change? In other words, how much of the press release text is used verbatim?

**RQ2.** Given a news article, which uses a press release as a source, what is the proportion of the article's text not copied without any change from the press release? In other words, how much of the article's text is not copied verbatim from the press release?

**RQ3.** How does press release content compare to the content of a news article, which uses that press release as a source, in regards to the evaluative, or subjective, language they use? I hypothesize that a news article will use less evaluative, or subjective, language compared to the press release it uses as a source.

**RQ4.** How does press release content compare to the content of a news article, which uses that press release as a source, in regards to the polarity (positive versus negative) of the evaluative, or subjective, language they use? I hypothesize that the

language of the news article will be less positive compared to the language of the press release it uses as a source.

**RQ5.** Do news articles, which use press releases as a source, provide attribution to the press release? In other words, do such articles mention the press release as the source of their content?

## Structure of the Dissertation

This dissertation consists of seven chapters. Chapter One introduces the central problem addressed in this dissertation. I review scholarship representing two opposite perspectives on the issue of news media reliance on the press release. Journalists, journalism scholars and press critics voice numerous concerns over the practice of news media using press release content as a source. Public relations scholars share a different perspective, according to which both public relations and news media play an important part in the process of information dissemination. However, a review of articles offering the perspective of the public relations industry suggests that the field's tactics have not evolved qualitatively over the past hundred years.

Chapter Two serves as a theoretical foundation for the issues raised by journalism scholars and discussed in Chapter One. I show how a press release may be conceptualized as an essential component of the information exchange between news media and public relations sources, how it is used as a tool for media agenda building, and how – based on agenda-setting theory and the concept of framing – a press release may be responsible for affecting the opinion the media's audience holds of the organizations providing the press release.

Chapter Three examines research which has been conducted on press releases and their relationship with news media. This review demonstrates that although such research is conducted in multiple fields, it has considerable shortcomings preventing it from adequately addressing the central problem discussed in this dissertation.

Chapter Four analyzes the specific shortcomings of previous research on press releases, focusing on the lack of an explicitly defined connection between press release and news article. It makes an argument that an adequately sized sample of press releases and news articles having such a connection can be collected only with the help of a computational approach. The chapter proceeds to discuss the use of computation in mass communication research and proposes the utility of computational methods for this study in particular. The overall goals of this study and specific research questions are stated at the end of this chapter.

Chapter Five provides a detailed description of the research design, explaining how basic methods from computer science and mass communication research can be combined to tackle a specific problem in journalism studies. The chapter describes the sampling strategy used in this study; the steps taken to construct a data set – i.e., the extraction of text from web sites, transformation of loosely structured text into well-formatted data, and reduction of a very large data set to a sample of most relevant items suitable for both automated and manual textual analysis; and the methods applied to address the specific research questions outlined in Chapter Four.

Chapter Six consists of two different parts. The first part describes the results of designing a method for discovering relationships between press releases and news

articles, and constructing the final data set. The second part describes the results of addressing specific research questions.

Chapter Seven discusses the results of the data analysis, as well as the overall outcomes of this exploratory study. In this chapter, I examine the role of data in describing the use of press release content by news media; the limits of what quantitative method can provide, as well as the limitations of this particular study; and the overall possibilities offered by applying computational methods to investigating problems in journalism and mass communication.

# CHAPTER 1.
# PROBLEM DESCRIPTION

The focus of this dissertation is news media's use of press release content – a problem which has been prominent for at least a century. The problem can be viewed from at least two very different perspectives, which I refer to as the journalism perspective and the public relations perspective. One argues that by using press releases, the news media publish "prepackaged" information, carefully constructed to promote the sponsoring organization's viewpoint on issues (Zoch & Molleda, 2009), serving the needs of the public relations industry and its clients, instead of serving the public. However, according to the other perspective, both news media and public relations play integral roles in the mass communication process: "PR is an important part of the elaborate process of delivering accurate, unbiased and timely information to the general public ... The media's job is to transform these materials, gathered together with dozens of their own ideas and sources" (Brooks, 1999, pp. 26-27). Both of these perspectives are supported by various scholarship, a selection of which I examine in this chapter.

## Journalism Scholarship: Views on the Issue

### Selecting the Literature

There has been much research on the relationship and cross-perceptions of journalists and public relations practitioners. Numerous studies have shown that journalists have been accusing public relations practitioners for decades of being "unethical, manipulative, one-sided, and deceptive," and serving "special interests rather than the public" (DeLorme & Fedler, 2003, p. 99). However, the vast majority of such research examines the broader picture, with press releases being but one minor part

among the numerous aspects which constitute the complex relationship between the two fields. (For an overview of the relationship and cross-perceptions of journalists and public relations practitioners, see DeLorme & Fedler, 2003; Kopenhaver, 1985; Len-Ríos et al., 2009; Paluszek, 2002; Ryan & Martinson, 1988; Sallot & Johnson, 2006; Supa & Zoch, 2009).

To examine the journalism perspective on press releases, I turned to two publications which contain articles by journalists, press critics and journalism scholars. One is the *American Journalism Review*, a national magazine, which is "written and edited by respected journalists" and "covers all aspects of print, television, radio and online media, [including] ethical dilemmas in the field" ("American Journalism Review," 2012). The other is the *Columbia Journalism Review*, a similar publication which "[encourages] excellence in journalism in the service of a free society, ... monitors and supports the press, ... [offers] a mix of reporting, analysis, and commentary ... hosting a conversation that is open to all who share a commitment to high journalistic standards in the US and around the world" ("Columbia Journalism Review," 2012).

I used the Communication and Mass Media Complete online database to search for articles containing the phrases "press release" or "news release" (the two terms are interchangeable). A total of 16 relevant articles published during the period from 1974 to 2011 were identified. These articles do not describe the results of scientific studies; they mostly feature opinions, supported by anecdotal evidence and, occasionally, empirical data limited to a single case study or a single issue of a given newspaper. These are well-written fact-based stories, which lack the rigor of formal research, yet provide a compelling snapshot – from a journalist's perspective – into what is problematic about the

news media's practice of accepting press releases from the public relations industry. The concerns which are raised in these articles are plentiful, and range from individual problems of, arguably, minor significance (e.g., "A press release and a news story," 1975) to troublesome issues, suggesting major implications for a democratic society (e.g., Bagdikian, 1974).

"We know the practice of using press releases is widespread, but we're still disturbed by it, not least because the public expects that a newspaper story is an impartial account" ("A press release and a news story," 1975, p. 5) – this sentiment is central to these articles. More specifically, the expressed concerns can be summarized as the combination of the various ways of selecting and emphasizing favorable content in press releases, the amount of such "prepackaged" content supplied by the public relations industry, the little (if any) contextual information added by the journalist, the ease with which this content becomes printed news without any significant changes, and the lack of attribution to the press release as the source of the news.

**Positive Content, Favorable Narrative, and No Attribution to the Source**

Biased news reporting is one of the most commonly expressed concerns in the reviewed articles. One of the manifestations of such bias is the selection of information to report or even the arrangement of content to make it look favorable towards the sponsoring organization – i.e., the source of the news. While some authors agree that using press releases leads to the prevalence of positive content over negative content (Bagdikian, 1974), a few point out that even a balanced press release, used verbatim or without major edits, may lead to a biased news account. A short article from 1975

analyzes a press release from The Boeing Company and the subsequent news article. The author compares the first paragraph of the press release to the first paragraph of the corresponding article and finds them identical – which, the author explains, is no surprise, since "Seattle is a company town, and ... Boeing is the company" – which makes news of the company's financial performance important to readers of the local paper. However, the author argues that the positive news (the company's improved financial performance) is presented ahead of the negative news (employment reduction in the Seattle area), which makes this – according to the author – "an especially clear example of what is wrong with a practice that some editors defend" ("A press release and a news story," 1975, p. 5) – i.e., running a press release as a news story with only minor changes, resulting in the presentation of positive news ahead of negative news.

The order in which the parts of a news story are presented – i.e., positive information preceding negative information – can be viewed in a broader context. Sullivan views the problem of "original reporting often [containing] the fingerprints of government and private public relations" as a problem of public relations supplying the narrative of the news report (Sullivan, 2011, p. 37). Sullivan cites a Pew Center study which analyzed the Baltimore news market and found that "63 percent of [news about a given set of subjects] was generated by the government, 23 percent came from interest groups, or public relations, and 14 percent started with reporters" (p. 37). Upon closer examination of nineteen articles based on a press release from the University of Maryland dealing with the development of a vaccine, it was revealed that "three contained significant new information, another three had new details, and the rest either repeated the same basic facts as the press release or were identical stories appearing on a different

platform" (p. 37). The problem in this, according to the Pew Center report is that such practice "does hand a lot of control over the narrative to the institution that is peddling the story." (as cited in Sullivan, 2011, p. 37)

"Handing over the narrative" to public relations may indirectly lead to shallow news accounts (Sibbison, 1988; Smith, 1977): after all, once the narrative is constructed, it is but one step away from publication. Sibbison (1988) described the publication of press releases issued by the Environmental Protection Agency, which lead to "relaying to readers self-serving statements by EPA officials as the truth" (p. 26). Smith (1977) discussed the publication of FDA press releases which lead to news containing statements shaping general perceptions without providing any detailed context. Smith argued that the media failed to explain the scientific evidence to its audience: "...The real story was not the [subject of the press release] itself, but the climate of opinion it created... This was the story that the media and the public missed" (p. 29).

Jones (1975), speaking from the public relations side, yet from the same perspective as the journalism scholars which have been cited in this chapter, offered a curious opinion of how easy it was to get press releases to be published almost verbatim: "a surprising number of the smaller daily and weekly newspapers... are quite willing to dispense with editors and reporters whenever they can, and turn instead to innocuous feature stories and 'canned copy' ... in order to fill up the empty spaces between the advertisements" (p. 10). Describing his own personal experience, Jones recalled that "many of [their] news releases ran in newspapers all over the country, very often word-for-word, with nothing added and nothing cut, ...[that] much of the American press obligingly and uncritically told their readers exactly what [the organization behind the

press release] wanted them to know" (p. 10). Furthermore, Jones points out that the press release was by no means impartial: "It was not a very subdued release... The one-thousand-word release, exactly as [the company] wrote it, was deemed usable by editors in more than fifty American newspapers" (p. 10).

Similar issues have been observed in regards to the publication of news from Congress. Bagdikian (1974) observed that "most of the media are willing conduits for the highly selective information the member of Congress decides to feed the electorate" (p. 4). Bagdikian emphasizes the issue of attribution, or lack thereof: "This propaganda is sent to newspapers and broadcasting stations, and the vast majority of them pass it off to the voters as professionally collected, written and edited 'news'" (p. 4). The author made his claim after examining a collection of local newspaper clippings. Several of the articles appeared to have shared identical paragraphs with corresponding press releases requested by the author from a congressman's office: "the papers had run the release word for word as journalistic news" (p. 5). Bagdikian argued that such practice was not a rare phenomenon: "hundreds of press releases, paid for by the taxpayers, are sent to the media by members of Congress, and hundreds are run verbatim or with insignificant changes, most often in medium sized and small papers" (p. 5). Furthermore, Bagdikian points out that the published press releases result in a strong bias towards positive news: there is no press release when something goes wrong – therefore, only good news is announced and published.

Yet another study directly focuses on the issue of attribution to public relations as the real source of news content by the media. Ambrosio (1980) conducted an analysis of a single issue of the Wall Street Journal after the paper announced a 10% increase in

space for news content. The author requested press releases from all companies mentioned in the short articles which lacked in-depth analysis typical of the Wall Street Journal. After analyzing 70 press releases (out of 111 requested), the author found that 53 news stories "were solely based on press releases"; 32 out of them "were reprinted almost verbatim or in paraphrase," and in 21 "only the most perfunctory additional reporting had been done" (p. 36). A total of 84 stories were based on press releases, which accounted for 45% of the day's 188 news items and 27% of the paper's non-tabular "news hole." The author's main argument was about the amount of press release content presented as news collected by the newspaper's reporters and the apparent lack of attribution to the real source. The author explains that "distinctions clear to the editors and staff of the nation's largest daily may be far less clear to its readers" (p. 35). Likewise, Potter (2004) noted that "the audience deserves to know if what they are reading/seeing is a handout from a commercial or a government source" (p. 68).

**"The Grain of Salt"**

Overall, concerns over the lack of attribution, as well as those dealing with biased news in general, seem to go hand in hand with discussions of the overwhelming proportion of public relations content in news media, with the study by Ambrosio (1980) being but one example. Furthermore, such issues are sometimes generalized to arguments about the shifting balance between journalism and public relations towards public relations (Starr, as cited in Sullivan, 2011). Sullivan refers to McChesney and Nichols who "tracked the number of people working in journalism since 1980 and compared it to the numbers for public relations... [and] found that the number of journalists has fallen

drastically while public relations people have multiplied" (McChesney & Nichols, as cited in Sullivan, 2011, p. 34). Thus, the concern over press release content is often generalized into a concern over the whole practice of public relations:

> The problem is that there is a large gray zone between the truth and a lie... Skillful PR people can exploit this zone to great effect ...They are able to provide data that for journalistic purposes is entirely credible... The information is true enough. It is slanted. It is propagandistic. But it is not false (p. 36).

Sullivan concludes that "without the filter provided by journalists, it is hard to divide facts from slant" (p. 38). However, this begs the question: *is the filter provided by journalists objective?*

Public relations scholars and practitioners alike have repeatedly criticized journalists for their self-perception as the only provider of objective and balanced news, sometimes blaming journalism education for "the wariness journalists feel toward public relations practitioners, and the consequent defensiveness practitioners feel about their communication with the media" (Kopenhaver, 1985, p. 1). And indeed, books like "Public relations and the subversion of democracy" (Miller & Dinan, 2007) or the infamous "Toxic sludge is good for you: Lies, damn lies and the public relations industry" (Stauber & Rampton, 1995) serve as a reminder that sometimes journalists may demonstrate anything but balance and impartiality in expressing their views on the field of public relations. Kopenhaver (1985) describes this as "misunderstandings about the contributions both groups make to information dissemination" (p. 1). Could these misunderstandings be a result of the historical antagonism between the two professions? Should we take the grave concerns expressed by journalism scholars with a grain of salt?

## Journalism and Public Relations: the Troubled Relationship

### The Roots of the Antagonism

To better understand the nature of the problem of pre-packaged content accepted by news media, it may be helpful to briefly examine the roots of the relationship between the two professions.

Public relations historians often trace the roots of the profession to the practice of press agentry (e.g., Newsom, Turk, & Dean, 2004; Zoch & Molleda, 2009). Press agents typically would solicit media attention for their clients by any means possible, with P.T. Barnum being among the most famous, notorious for his stunts and hoaxes (Newsom et al., 2004). It has been shown (e.g., Cutlip, 1994; Lamme & Russell, 2010) that it is the practice of press-agentry that is largely responsible for the reputation the field of public relations has had with the media throughout the years.

However, although the origins of public relations are considered to be in press agentry, most public relations scholars argue that modern public relations appeared as a profession at the beginning of the 20th century and should be associated with the concept of corporate publicity (e.g., Kruckeberg & Starck, 1988; Ledingham & Bruning, 2000; Zoch & Molleda, 2009), which is distinct from the practice of press agentry. Russell and Bishop (2009) found through an analysis of the press between 1865 and 1904, that "while press agentry was connected to the circus and theater, 'corporate publicity' was linked to Theodore Roosevelt's call for the release of financial information in the public interest" (p. 91). Kruckeberg and Starck (1988) explain that the development of corporate publicity may have been prompted by the reaction of industry against the muckrakers: “Business people began asking themselves whether traditional policies of secrecy were

really the wisest course. If publicity was being used so effectively to attack business, why could it not be used equally well to explain and defend it?" (p. 6).

In 1906, Ivy Lee, a former journalist hired by John Rockefeller Jr. to help his business communicate its message to the public, declared his principles of a publicist (for a detailed account, see Hiebert, 1966). In his declaration, Lee stated:

> In brief, our plan is, frankly and openly, on behalf of business concerns and public institutions, to supply the press and public of the United States prompt and accurate information concerning subjects which it is of value and interest to the public to know about (as cited in Supa & Zoch, 2009, p. 1).

Lee's declaration of principles signified the transition of corporate attitudes from "public be damned" (Newsom et al., 2004, p. 31) or "the public be fooled (Goldman, as cited in Grunig & Grunig, 1992, p. 286) – to "the public be informed" (Hiebert, 1966, p. 48). Russell and Bishop (2009) note that Lee's statement "established publicity as an open and honest function of corporations" (p. 92). Essentially, Lee declared that his goal was to *provide truthful information* about his clients to the press – a statement which goes in sharp contrast not only with the practice of press agentry, but with all the concerns expressed by journalism scholars and discussed at the beginning of this chapter.

Nevertheless, research shows that the antagonism demonstrated by journalists towards the field of public relations goes back at least a century. DeLorme and Fedler (2003) observe that "prior academic work has found that the hostility between journalists and PR practitioners began at the end of World War I, when the newspaper industry started a campaign against 'spacegrabbers'" (p. 100), referring to the attempts of "press agents" to get free space in the newspaper – i.e., we see that journalists often did not see much of a difference between a press agent and a publicist. Rodgers (2010) refers to

Frank Cobb, the editor of the New York World, who "charged that one of the reasons for the confused state of public opinion was the press-agent system" (p. 52).

Roscoe C.E. Brown (1921) noted that "'pre-digested news' was affecting the process of news gathering as reporters became 'a race of mere retailers of ready-made intelligence'" (as cited in Rodgers, 2010, p. 52) – a statement which explicitly demonstrates that concerns about news media's reliance on the press release are easily a hundred years old. Indeed, scholarship from different years, including studies conducted in the 1980s, 1990s and 2000s (e.g., DeLorme & Fedler, 2003; Spicer, 1993), as well as going back to as far as 1906 (Rodgers, 2010, p. 52), has demonstrated the same sentiment among journalists and journalism scholars alike; and that sentiment is consistent with all the concerns expressed by journalism scholars, which were discussed at the beginning of this chapter and are the premise of this study.

**The Evolution of Public Relations**

Public relations scholars, however, have a different view of their field and its goals. The dominant paradigm is centered around an evolutionary concept of the field, described by Lamme and Russell (2010) as "a progression from 'worst' to 'best' practices in public relations" (p. 286).

Grunig and Hunt (1984) developed a theory of four models of public relations, according to which public relations evolves from a one-way asymmetrical model – i.e., press-agentry/publicity, which implies unbalanced, one-way communication between the organization and its audience. Grunig and Grunig (1992) considered the practice of hiring a "journalist in residence" to be the next stage in the development of public relations,

which is the public information model. They noted that although these journalists, hired as public relations counsel, included only favorable information in their handouts, the information was generally truthful (p. 288).

Beginning with the Creel Committee during World War I public relations practitioners began to incorporate into their work behavioral and social sciences. The approach was based on gathering information about the organization's target audience and applying it to achieve the organization's communication goals. Finally, the two-way symmetrical model, proposed by Grunig and Grunig (1992) implies the use of research to gather information about the organization's publics to facilitate understanding and communication rather than to identify messages most likely to persuade or motivate publics. In this model, understanding, rather than persuasion, is the principal objective of public relations (p. 289). Thus, according to this theory, it would seem that the scholarship discussed previously in this chapter refers to a stage in the evolution of public relations which is long gone.

The theory, however, is questionable on many levels. It theory has been criticized for being ahistorical – i.e., constructed on individual facts only loosely based on historic reality. Lamme and Russell (2010) noted that its first phase (or first model) "accounted for all the time before 1900 yet largely focused on the exploits of P.T. Barnum in the mid-19$^{th}$ century … which evolved into what became 100 years later the ethical, professional, and two-way 'public relations' model firmly rooted in postwar American business" (p. 286). Brown, most appropriately, referred to this as the 'Big Bang Barnum concept'" (as cited in Lamme & Russell, 2010, p. 286).

Scholars have also criticized individual stages in this "evolutionary process." With regard to the public information stage, Ledingham and Bruning (2000) pointed out that "the dominance of the field … by former journalists reinforced the notion of manipulation of the mass media and generating favorable publicity as the central focus of public relations practice" (p. xii). Cutlip (1994) quotes an early practitioner: "I was in the publicity business. I was a press agent. Very simply, my job was to get the client's name in the paper" (as cited in Ledingham & Bruning, 2000, p. xii). Thus, the goal of early public relations was influencing public opinion through the use of mass media (Golitsinski, 2007).

The next stage – which saw the incorporation of social scientific research into the practice of public relations – is associated with Edward Bernays (Grunig & Grunig, 1992), whose definition of public relations stated that "public relations is an attempt, by information, persuasion, and adjustment, to engineer public support for an activity, case, movement or institution" (Bernays, 1955, pp. 3-4). Thus, the theories of this approach introduced by Bernays were based on propaganda, persuasion and "engineering of consent" – which, again, can be described as manipulation of public opinion through the use of mass media, and which, again, is consistent with the concerns discussed at the beginning of this chapter.

Furthermore, historians have noted that the actions and views of P. T. Barnum, Ivy Lee, and Edward Bernays would have been considered unethical by today's standards, but those people were products of their time, and what they did was considered acceptable (Vos, 2011). Thus, there may be little historical evidence of a distinct evolution of the profession. In fact, there is evidence against it: The Creel Commission

(Pinkleton, 1994) which handled propaganda during the First World War came after Ivy Lee declared the principles of openness and publicity: "if PR did, indeed, arrive in 1905, explanations that point to the Creel Commission's work a decade later are problematic at best" (Vos, 2011, p. 121).

In fact, many public relations scholars, including Grunig and Grunig (1992), agree that today's public relation is still focused primarily on media relations and publicity. Kruckeberg and Starck (1988) argue that public relations was most commonly practiced today as persuasive communication to obtain a vested goal on behalf of a client. Ledingham and Bruning (2000) observe that, although some scholars argue that the role of "journalist in residence" has been replaced by that of the "expert prescriber" – a public relations counselor who advises the client on matters of public policy (Broom and Dozier, 1986, as cited in Ledingham & Bruning, 2000, p. xii), in reality, organizations "still view public relations primarily as a means of generating favorable publicity. Their rationale for public relations is found not in the management of reciprocal relationships between an organization and its publics, but rather in 'the credibility attached to information that has been examined by reporters through third party endorsement by the media'" (p. xii). Finally, Grunig and Grunig (1992) discovered that, contrary to their expectations, press-agentry – the first model of public relations – was still the most common form of public relations in practice (p. 305).

## Public Relations Industry: Views on the Issue

There have been numerous studies examining the relationship between public relations and news media. However, it is commonly known that there exists a

considerable disconnect between academia and practice in the field of public relations (Heath, 2010). Therefore, rather than examining more research conducted by scholars, I turn to those who practice public relations, who, among other things, write press releases for a living, and offer advice on such writing to their colleagues.

I use *Public Relations Quarterly* as my source, which is a publication whose audience is mostly public relations practitioners. The nine relevant articles I was able to locate span the period from 1976 to 2005, which, conveniently, is quite similar to the period spanned by the articles representing the opposite perspective and written by journalists, journalism scholars and press critics, which were discussed at the beginning of this chapter (1974 - 2011).

It is no surprise that the issue of press release content used by news media is viewed quite differently in this literature. From their perspective, the one and only problem appears to be in getting the news media to use the information supplied through the press release; any implications of such practice and the consequences for journalistic objectivity of the resulting news report are nonexistent in this literature.

I have identified three common themes in these articles. One theme is focused on proposing various rules dealing with form and structure which, according to the authors, increase the chances of a press release to be accepted for publication. The other theme suggests that public relations people become better salesmen and practice their sales skills and creativity at selling the press release to the editor. The last theme, rare, but distinct, deals with disguising the true intent of a press release. For lack of better terms, I refer to these themes as *the pointless, the annoying* and *the troubling*.

**The Pointless**

The negative attitude of news media towards press releases is well known in the field of public relations. Some authors even suggest doing away with press releases; Marken (1994) observed that the "one universal statement [heard] is that editors seldom look through the growing pile of news releases they receive. When they do, they find the majority a waste of time, money and efforts" (p. 46). Based on the literature under review, such negative attitudes towards press releases among journalists were common place at least as far back as 1976. Pedersen (1976) noted that editors hated press releases, that press releases were badly written, and only a few resulted in news coverage. The main complaint, according to Pedersen, was about bad writing overall.

To counter such negative attitudes, judging by the literature, the field seems to be looking for formulaic approaches to improve the form and structure of the press release, instead of worrying about its content. What's remarkable is that recommendations made by Pedersen in 1976 closely resemble those made by Applegate in 2005 – i.e., three decades seemed to have changed nothing. Pedersen (1976) recommended changing the title "news release" to "news from..." – to avoid using the term *press release* due to its poor reputation. Thirty years later, Applegate (2005) suggests placing the title "PRESS RELEASE" (in caps) at the top of each press release. In both cases the authors are convinced that these "cosmetic" measures will improve the chances of a press release being published.

Applegate (2005) also suggests writing the release as a news story, answering the who/what/when questions, using the inverted pyramid structure, as well as following a long list of format-related guidelines such as using the company letterhead and serif

typefaces – the list is quite long. Many other authors suggest similar approaches to the problem. Some authors describe detailed stylistic improvements: Ryan (1995) advises that press releases should consist of "four paragraphs, or parts, each of which is composed of one compound sentence containing two phrases" (p. 26).

To be fair, some authors do mention content as one of the criteria. Obston (2004) suggests excluding specific words from press releases (mostly adjectives like *unique*, *revolutionary*, *cutting edge*), whereas Applegate (2005) and Brooks (1999) mention specific newsworthiness criteria as one of the main criteria determining the acceptance of the press release.

**The Annoying**

Not all suggestions on how to get a press release accepted deal with form and structure: the other most common theme in this literature is taking a strategically different approach and getting better at "selling" the press release to the media – an approach where the public relations practitioner is the salesman and the editor or reporter are the customer. Williams (1994) observed that "getting a story is basically the result of negotiation, and any decent negotiator will tell you to always look at a problem from your adversary's point of view before crafting your strategy" (p. 6). In other words, designing the "right" strategy will win the negotiation with news media and will get the story published.

Another take on strategy is presented by Brooks (1999) who proposes a supply-chain management analogy for media relations that, according to the author, will make news media more effective. The author quotes a study which surveyed 2,500 business

reporters and editors and found that 65% said PR people were the least likely to be considered a useful source, 60% said PR never or rarely gave useful comments for the story (p. 26). The author's conclusion? Public relations has an image problem caused by silly promotions, empty press releases, etc. – and the solution is a better business relationship between public relations and the media: "PR is an important part of the elaborate process of delivering accurate, unbiased and timely information to the general public ... The media's job is to transform these materials, gathered together with dozens of their own ideas and sources" (pp. 26-27). Brooks believed that "reporters and publicists have one powerful, shared goal: [they] both want to [fill] the front page with colorful, interesting stories that are accurate, creative and effective" and saw the primary function of public relations as "to give journalists terrific story ideas their publication or program can use to attract and retain an audience" (p. 27). These are presented as story ideas that will pass the scrutiny of "tough editors" who "before reading a single word of copy ... will likely ask: 'Is it news? Is it timely? Is it relevant? ... And most important of all, will it amuse me?'" (p. 28) – i.e., amusement is presented as the primary criteria of newsworthiness. The author's ultimate conclusion on improving this business relationship was to move away from selling stories and, instead, to offer journalists story ideas "that sell themselves" (p. 28). How does one write press releases (or comes up with stories) "that sell themselves"? One example can be borrowed from a research paper by Levin (2002): one of the measured variables determining the success of a press release is described as "personalizing the press release with a handwritten note to a journalist" (p. 85). To be sure, there are many other sales techniques – and Brooks (1999) recommends

that public relations practitioners talk to salespeople regularly to become better at using such techniques.

Better selling techniques include more than successful negotiations. Pelham (2000) suggests using creative approaches to get a press release noticed: "Considering the plethora of news releases with which editors are bombarded, you may want to consider some methods for distinguishing yours from the rest of the crowd... If you can be a bit different (albeit in a relevant way), you will get noticed" (p. 40). The author agrees that the press release "should be kept brief and purely factual," but, at the same time, he argues that using "a non-traditional topic or headline, when appropriate, can serve as a powerful asset" (p. 40). The author even provides an example of creating such a "powerful asset": "Which story would interest you more: 'Macintosh Apple Opens Regional New York Headquarters' or 'Macintosh Sponsors Apple Plant in the Big Apple'?" (p. 40).

Such creativity aimed at getting content into the news can be extended from word choice to choice of medium. Marken (1994) suggests using other ways to get a message to the media: "If your organization is going to be involved in an event or do something that's clearly newsworthy, don't bury the facts in a tired press release. Use a media alert, a short punchy announcement of who, what, when, where and why to catch the attention of the press" (p. 46). In other words, the author proposes to send the same content disguised as something more newsworthy than a press release. Echoing ideas put forward by Pedersen (1976), the author also suggests using pitch letters, feature stories and other types of content to attract the editor's attention by "plant[ing] the germ of an idea in his or her mind" (p. 47).

## The Troubling

But why is there such a strong emphasis on using creative approaches to "push" the press release? After all, according to public relations scholars and practitioners, the press release provides journalists with "information that they need to do their jobs" (Supa & Zoch, 2009, p. 3). One could speculate that the need for better salesmanship can be explained very easily: maybe a press release is not always information the journalist needs, maybe it does not always contain the "news that's fit to print"?

An article by Williams (1994) may justify this assumption. Williams suggests that in writing a press release one must disguise the corporate strategic message as news: "The true test of a release writer – indeed, of everyone in this business – is how well he or she disguises corporate intent as news" (p. 5). The author argues that "the role that releases can and should play is establishing an organization's identity and affecting public and media attitudes toward it;" he explains that a press release should include both news and a corporate strategic message:

> [A press release] enables you to tell a story from your perspective, in a larger context that in all likelihood is alien to even your beat reporters. Of course, you have to disguise it as news, which refocuses on the issue of real news writing versus shameless self-promotion (p. 7).

One is left to wonder: if disguising a corporate message as news is not shameless self-promotion – then what is?

## Summary

In this chapter, I have examined two opposite perspectives on the issue of press release content used as a source by news media. Journalists, journalism scholars and press critics, who represent the journalism perspective, have voiced numerous concerns

over such practice. The overall concern is that the use of press release content with little or no change and no contextual information added, results in mostly positive, self-serving statements published on behalf of the organization, yet presented as an impartial news account, often with no attribution to the press release as the true source of the story. Thus, by adopting such practice, news media may be serving the needs of the public relations industry and their clients instead of serving the public.

Public relations scholars share a different perspective. It is widely believed among the public relations academy that the field has went through an evolutionary process, moving from its unethical roots in press agentry to the ethical model of today's public relations, with its primary objective being not persuasion, but understanding. According to this perspective, both public relations and news media play an important part in the process of information dissemination.

However, a review of articles offering the perspective of the public relations industry on press releases has demonstrated support for the concerns voiced by the journalism community. Public relations practitioners appear to be concerned about one thing only: getting the press release published; any implications press release content may have for journalistic objectivity in the resulting news report seem to be nonexistent. Furthermore, the nature of the advice public relations practitioners offer may suggest that the field's tactics have not evolved qualitatively since the era of press agentry.

In the next Chapter I discuss the issue of press release content appearing in news media from a theoretical standpoint. I will then examine research which has been conducted on press releases and their relationship with news media to determine whether this issue has been researched and explained in sufficient depth.

# CHAPTER 2.
# THEORETICAL FOUNDATION

In this chapter, I will examine the problem of press releases used by the media as a news source from a theoretical standpoint, employing contemporary mass communication scholarship. I will show how a press release may be conceptualized as an essential component of the information exchange between news media and public relations sources, how it is used as a tool for media agenda building, and how – based on agenda-setting theory and the concept of framing – a press release may be, in fact, responsible for affecting, at least on part, the opinion the media's audience holds of the organizations supplying the press releases.

## Information Subsidies, Agenda-Building and the Corporate Message

According to almost any textbook on public relations, press releases are only one component of media relations; whereas media relations are but one part of the much broader field of public relations (Zoch & Molleda, 2009). Nevertheless, the press release is often seen as one of the key components in the "troubled relationship" (Gower, 2007) between journalism and public relations.

### Press Releases as Information Subsidies

It is well known that the press release has existed for many decades for a simple reason: the media needed information for creating its news content, and public relations practitioners needed to communicate information about their client or employer to the public through a credible source – which is the news media (Berkowitz & Lee, 2004).

However, the first theoretical conceptualization of this information exchange was offered by Oscar Gandy (1982) in his book "Beyond Agenda Setting." Gandy used political economy as the conceptual foundation for his argument. According to Gandy, information in a capitalist world is a sought-after commodity which has a price. Journalists, and news media in general, are consumers of information in the sense that they are required to produce information artifacts to fill the newspaper. Government, corporations, or other organizations are interested in reaching the public through the media for the sake of advancing their interests (in fact, in 1981 mass media was *the only* way to reach the general public with a message). To do that, these organizations offer pre-packaged information to news media which is intended to offset the media's cost of producing their own information – which is what Gandy refers to as an information subsidy (Gandy, 1982).

The reason why information subsidies are a fact of life is two-fold: both the news media and the public relations industry have a strong need for it. Why do journalists use such information provided to them by their sources – after all, wouldn't that be inconsistent with the mission of journalism in a democratic society which is to watch over the powers that be and provide objective and impartial accounts to their audience? The "watchdog" function of news media implies a commitment to investigative journalism, which, Gandy explains, is very time-consuming and expensive: "Where the average journalist may generate one 'think piece' a week, the use of bureaucratic sources facilitates the [production of two or more routine stories each day" (Gandy, 1982, p. 12). In fact, Gandy's argument is strongly supported by research going back over a hundred years: journalists being "overworked and underpaid" has been a recurring theme and is by

no means new in today's media landscape (DeLorme & Fedler, 2003). This explains why news media have relied on information subsidies for decades (Hong 2008, Morton and Warren 1992, Turk 1985, and Walters and Walters 1992, as cited in Maat & de Jong, 2012).

The practice is, naturally, fraught with controversy; Gandy (1982) warned that such practice leads to corruption of the news gathering process: "Organizations with resources have more opportunities to offer this subsidy to the media than groups or organizations without resources … Thus, well-established groups such as businesses, government agencies, and paid experts will more often provide background for news stories and their positions on issues will appear more often in the news coverage" (as cited in Taylor, 2009, p. 24).

However, where Gandy (1982) sees a threat to the news gathering process, public relations scholars see opportunity to influence the agenda building process (e.g., Kopenhaver, 1985; Ohl, Pincus, Rimmer, & Harrison, 1995; Zoch & Molleda, 2009) – a term used in public relations scholarship to describe the process of building the media agenda. The need of the public relations industry to build the media agenda is based on its need  not only to reach the public, but to reach it through an impartial channel, which makes their information or message appear to be more credible than if it were offered directly by the source (Berkowitz & Lee, 2004).

**The Problem of the Message**

However, it is not the fact that the corporate message gets delivered to the public that is troubling: after all, it could be considered yet another source in the news gathering

process. It is *the nature of the message* which is being relayed by news media to the public that makes this an issue.

Chapter One of this dissertation showed that both journalism and public relations authors alike consider the press release to be anything but impartial. But this is not at all new: almost a 100 years ago Walter Lippmann (1922), in his analysis of the concept of public opinion, pointed out that leaders in business and politics were "compelled often to choose even at the best between the equally cogent though conflicting ideals of safety for the institution and candor to [their] public," (p. 158) and had to decide what facts and in what setting would be made available to the public. Lippmann suggested that the underlying reason for the existence of the press agent, or public relations, was the knowledge of how to "manufacture consent":

> The enormous discretion as to what facts and what impressions shall be reported is steadily convincing every organized group of people that whether it wishes to secure publicity or avoid it, the exercise of discretion cannot be left to the reporter. It is safer to hire a press agent who stands between the group and the newspapers … Many of the direct channels to news have been closed and the information for the public is first filtered thru publicity agents. The great corporations have them, the banks have them, the railroads have them, all the organizations of business and of social and political activity have them, and they are the media through which news comes (pp. 217-218).

Lippmann (1922) explained that "the publicity man" made his own choice of facts for the newspapers to print, thus saving the reporter much trouble by presenting him a clear picture of a situation. Yet, that picture was "the one he [wished] the public to see. He [was] a censor and propagandist, responsible only to his employers, and to the whole truth responsible only as it accords with the employer's conception of his own interests" (p. 218).

And indeed, public relations scholars not only see the practitioner as a "'pre-reporter' for the journalist, providing them with information that they need to do their jobs" (Supa & Zoch, 2009, p. 3), but also make no secret of their goals and describe the concept of information subsidy itself as the generation by practitioners of "prepackaged information to promote their organizations' viewpoints on issues" (Zoch & Molleda, 2009, p. 284).

Thus, I conclude that by accepting a press release as a news source, news media is accepting "prepackaged" content prepared by an interested party. Usage of this content by news media may not only lead to corporations and other organizations which have the funds having more news coverage than those who don't; more importantly, it will complicate the news media's task to provide a balanced and impartial account of reality – for the content prepared by an interested party is never impartial: it's biased by definition.

The next section shows that through building the media agenda, the corporation may be achieving more than just sending a corporate message or affecting news coverage in general: by affecting the media agenda, a corporation may be also contributing to telling us "what to think about" – a phenomenon explained by the theory of agenda-setting.

## Agenda-Setting: Telling the Audience What to Think About

### The Broader Picture: Agenda-Setting and Agenda Building

The concept of agenda building is closely related to agenda-setting theory – a theory which helps explain an organization's overall purpose and motivation for building the media's agenda.

According to McQuail (2005), agenda-setting is "a process of media influence (intended or unintended) by which the relative importance of news events, issues or personages in the public mind are affected by the order of presentation (or relative salience) in news reports" (p. 548). That means that whatever news content – i.e., events, issues, personages , etc. – the media decide to emphasize, through giving that content more space and time (or more prominent space and time), will become a more important issue on the public agenda.

**Origins of agenda-setting.** The origins of agenda setting theory can be traced to Walter Lippmann's *Public Opinion*; in fact, M. McCombs and Estrada (1997) referred to agenda-setting as Lippmann's "intellectual offspring" (p. 237). In his book, Lippmann "scrutinized the centerpiece of democratic theory: the 'omnicompetent citizen'" (Steel, 1980). According to classic democratic theory, the average citizen understood important public issues and was capable of making rational judgments about them if presented with the facts. The facts were supposed to be presented by the press (i.e., the mass media). Examining how public opinion was formed, Lippmann argued that the media was not capable of providing an accurate picture of the world; and even if it could, "the average man" had neither the time nor the ability to deal with the complexity of such information; for, in a mass society he had a direct experience and understanding only of a tiny part of the world around him. What "the average man" saw, was an image of the world, "reflected through the prism of his emotions, habits and prejudices" – a reality that fits his experience, defined through stereotypes, or "pictures in our heads," which "provide security in a confusing world." (as cited in Steel, 1980). Lippmann noted that these stereotypes "determined not only how we see but what we see," for, in creating these

images for "the average man," "every newspaper … [was] the result of a whole series of selections as to what items shall be printed, in what position, how much space, etc… There [were] no objective standards [there]." Hence, through making these choices, the press became responsible for those "pictures in our heads." (Lippmann, 1922).

Later, in 1963, political scientist Bernard B. C. Cohen (1963) coined the famous phrase that the media "may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to think about" (p. 13). The idea was reiterated by Lang and Lang (1952) in their famous analysis of how television covered MacArthur Day in Chicago: "The mass media force attention to certain issues. … They are constantly presenting objects, suggesting what individuals in the mass should think about, know about, have feelings about" (as cited in Lowery & De Fleur, 1983, p. 380).

**Establishment of agenda-setting as a mass communication theory.** The term *agenda-setting*, as well as the theory itself, were introduced by M. E. McCombs and Shaw (1972) in a paper, which reported the results of their study of the 1968 presidential election. The authors made an argument that the mass media played an important part in shaping political reality through making choices in regards to what news to publish or broadcast:

> Readers learn not only about a given issue, but also how much importance to attach to that issue from the amount of information in a news story and its position. In reflecting what candidates are saying during a campaign, the mass media may well determine the important issues—that is, the media may set the "agenda" of the campaign (p. 176).

The paper's hypothesis stated that although the mass media may have little influence on the audience's attitudes, they do "set the agenda for each political campaign,

influencing the salience of attitudes toward the political issues" (M. E. McCombs & Shaw, 1972, p. 176). The authors found a high correlation between the rank order in salience of the issues reported in the mass media and the importance rank assigned to these issues by survey responders.

Since the Chappell Hill study, agenda-setting became one of the most widely tested theories in mass communication research. According to Rogers, Hart, and Dearing (1997), approximately 360 research papers on agenda-setting have been published between 1972 and 1995. Furthermore, even though agenda-setting research started with studies focused on political communication and election campaigns in particular (Wimmer & Dominick, 1997), M. McCombs and Reynolds (2002) noted that the evidence that supports the theory is not limited to elections: "the accumulated evidence about the agenda-setting influence of the news media on the general public comes from many different geographic and historical settings worldwide and covers numerous types of news media and a wide variety of public issues" (p. 3).

**Connecting Information Subsidies with Agenda-Setting**

Gandy suggested a connection between information subsidies and agenda-setting by noting that information subsidies "increas[ed] the salience of objects in news media coverage and in turn affect[ed] public perceptions" (as cited in Ragas, 2012, p. 91). In other words, by contributing to the process of building the media's agenda, one gets to affect the public agenda. M. McCombs (1983) called Gandy's book "an important benchmark in the transition from agenda-setting at the media-audience interface to a broader analysis of the agenda-building process" (p. 377). Later, M. McCombs and

Reynolds (2009) explained that while "the core theoretical proposition underpinning agenda-setting theory is the transfer of object salience from the media to the public, ... agenda building moves a step earlier in this process, focusing on the transfer of object salience between sources and the media" (M. McCombs & Reynolds, 2009).

Carroll and McCombs later noted that "accumulated research has provided solid evidence to date of corporate public relations efforts engendering agenda-building and agenda-setting effects ... through public relations strategies to influence the media with a variety of information subsidies" (as cited in Ragas, 2012, p. 93). Thus, it is no surprise that public relations scholars see the significance of information subsidies to the public relations industry in their contribution to setting the media agenda. (Turk & Franklin, 1987).

Thus, I conclude that by getting press released published in the news – i.e., by engaging in the process of media building – the public relations industry contributes to setting the public agenda. And by way of affecting the public agenda, in the words of Bernard Cohen, a corporation may be telling the public what to think about.

## Framing: Telling the Audience *How* to Think

The consequence of media usage of information subsidies might be not limited to corporations contributing to setting the public agenda. An examination of the concept of framing as a "second dimension" (M. McCombs & Reynolds, 2002) of agenda-setting suggests that not only does the media "tell the audience what to think about" – it might, after all, also tell the audience what to think, or at least *how* to think about the issues on its agenda.

## Defining Framing

Framing is a multidisciplinary concept with its origins spanning the fields of psychology, sociology, economics, communication science, and political communication – to name a few. Goffman (1974) introduced framing as a "schemata of interpretation, … which allows its user to locate, perceive, identify, and label a seemingly infinite number of concrete occurrences defined in its terms" (p. 21). Tversky and Kahneman demonstrated that the mere selection of a different wording resulted in a different opinion of the proposed question (1979, 1984, as cited in Scheufele & Tewksbury, 2007, p. 11). The concept of framing in mass communication is based on the assumption that how an issue is characterized by mass media may have an influence on how it is understood by audiences.

In 1993, Robert Robert M.  Entman (1993) observed that, at that time, there was no "general statement of framing theory that [showed] exactly how frames become embedded within and make themselves manifest in a text, or how framing influences thinking" (p. 51). To provide "a more precise and universal understanding" of the various uses of the term, Entman developed a theory of framing, according to which frames define problems, diagnose causes, suggest moral judgments, and recommend treatment, and occurred at least four components of the communication process: the communicator, the text, the receiver, and the culture.

## Framing Through Salience and Selection

One of the central themes in Entman's (1993) conceptualization of framing was its utility "to describe the power of a communicating text" (p. 51). The text of the message

contains textual frames, which, according to Entman, are manifested by the usage (or absence) of keywords, phrases, stereotypes "... that provide thematically reinforcing clusters of facts or judgments" (p. 52). Entman explained that framing occurs through selecting, highlighting, and using the highlighted elements "to construct an argument about problems and their causation, evaluation, and/or solution" (p. 53). The concept of selection and highlighting, or selection and salience – i.e., "making a piece of information more noticeable, meaningful, or memorable to audiences" (p. 53) – was the focus of Entman's definition of the process of framing:

> To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described (p. 52).

Making some information more salient, according to Robert M. Entman (1993), can be achieved through "placement or repetition, or by associating [it] with culturally familiar symbols" (p. 53). In fact, in an earlier study, Robert M. Entman (1991) offered a detailed description of this process as constructing news frames using keywords, metaphors, concepts, symbols and images:

> Since the [news] narrative finally consists of nothing more than words and pictures, frames can be detected by probing for particular words and visual images that consistently appear in a narrative and convey thematically consonant meanings across media and time. By providing, repeating, and thereby reinforcing words … that reference some ideas but not others, frames work to make some ideas more salient in the text, others less so – and others entirely invisible. … Through repetition, placement, and reinforcing associations with each other, the words and images that comprise the frame render one basic interpretation more readily discernible, comprehensible, and memorable than others (p. 7).

Robert M. Entman (1993) also noted that frames are defined not only through assigning more salience to selected information, but also through the exclusion of other information: "frames are defined by what they omit as well as include, and the omissions

of problem definitions, explanations, evaluations, and recommendations may be as critical as the inclusions in guiding the audience" (p. 54).

However, it is also necessary to point out that selecting and highlighting is not the only, or the necessary condition of framing effects: "even a single … appearance of a notion in an obscure part of the text [can become] highly salient, if it comports with the existing schemata in a receiver's belief system" (Robert M. Entman, 1993, p. 53). At the same time, highlighting an element which does not correspond to the receiver's schemata may not help the receiver to "notice, interpret and remember" the frame. In general, Entman observed that "the presence of frames in text does not guarantee their influence in audience thinking," for "salience is a product of the interaction of texts and receivers" (Entman, 1989; Graber, 1988, as cited in Robert M. Entman, 1993, p. 53).

**Manifest content versus multiple meanings.** Reese (2007) saw Entman's approach to framing rather limiting in the sense that it was attempting to define it too strictly as manifest content. Reese argued that Entman's definition "begs the question of how [news stories] are organized 'in such a way as to promote' their effects" (p. 152). Reese's believed that "it is precisely the way that certain attributes come to be associated with particular issues that should concern framing analysis" (p. 152). Reese noted that it had been widely accepted that there were features in a news text that, "when taken together, tell a larger tale than the manifest story" (p. 152). Therefore, Reese argued for a more interpretive approach in framing analysis, which takes into account the multiple meanings which may be derived from a text by different members of the audience.

However, Robert M. Entman (1993) argued that to identify a meaning as dominant or preferred is to suggest a particular framing of the situation that is most

heavily supported by the text and corresponds to the most likely audience framing schemata. He cautioned against using non-dominant parts of the message and examining how they might be interpreted in ways that oppose the dominant meaning: "If the text frame emphasizes in a variety of mutually reinforcing ways that the glass is half full, the evidence of social science suggests that relatively few in the audience will conclude it is half empty" (p. 56).

**Framing as Second Dimension of Agenda-Setting**

The connection between framing and agenda setting has been discussed by Scheufele (1999), who conceptualized the connection as the processes of frame building and frame setting, and by M. McCombs and Estrada (1997), who observed that a new direction in agenda-setting research was addressing the issue of how news frames affect public opinion. McCombs and Estrada described framing as the second dimension of agenda-setting, dealing with salience of object attributes. The authors explain that each object, or issue has attributes – i.e., characteristics that describe it. Objects vary in salience, and so do attributes. The agenda-setting roles are, therefore, the selection of objects and selection of attributes for thinking about these objects. Perspectives and frames that the audience and journalists alike may use to think about issues are constructed by semantic devices "which draw attention to certain attributes and away from others." Furthermore, M. McCombs and Shaw (1993) argue that framing being a second dimension of agenda-setting may indicate that contrary to Cohen's (1963) famous observation, "media may not only tell us what to think about, but also how to think about it, and consequently, what to think" (p. 65).

**Framing in public relations.** Public relation scholars view framing as a way to "establish common frames of reference as a prerequisite for building mutually beneficial relationships with the public." However, they also explain that "framing and information subsidies are just tools ... to participate in the building process of the media agenda" (Zoch & Molleda, 2009, p. 290). Reiterating Entman's (1993) conceptualization, Zoch and Molleda explain that public relations practitioners contribute to the framing of a story by "highlighting or withholding specific information about a subject or issue from those covering the story" (p. 283). The scholars acknowledge that "the ideal outcome of information subsidies' efforts will be that the coverage reflects a similar viewpoint to the one presented in the subsidies" (p. 290). Thus, "although it appears that the sources provide subsidies to the media in order to gain time or space (Berkowitz, 1990, as cited in Zhang, 2004, p. 6), their ultimate goal of agenda building is not just to influence the mass media, but to create a favorable public opinion and public agenda" (Turk, 1985, as cited in Zhang, 2004, p. 6).

Thus, it becomes apparent that by publishing press releases with little or no change, the news media relay to their audience not only the objects, or issues, included in a press release, but also the objects' attributes, which constitute the frames set by the organizations supplying the press release. Based on the theories discussed in this chapter, these objects, together with their attributes – i.e., the subject of the press release, as well as the frames – embody not only "what the media tells us to think about," but also "how to think about it."

## Summary

The scholarship reviewed in this chapter provides a theoretical framework which helps explain the role of press releases in the information exchange between the public relations industry and the news media. Press releases, described as information subsidies, are supplied by the public relations industry to news media in an effort to build, or contribute to building the media agenda. A review of agenda-setting theory has shown that when news media uses press releases as a news source, the organizations supplying the press releases may be affecting the public agenda. Furthermore, examining the concept of framing, conceptualized as a second dimension of agenda-setting, has shown how the usage of press releases with little or no edits, thus, preserving the viewpoints expressed in the press release, may be affecting the opinion the media's audience holds of the organizations supplying the press releases.

This chapter serves as a theoretical foundation for the issues raised by journalism scholars and discussed in Chapter One. However, before setting specific goals for this study, I must first determine whether or not these issues have been addressed in previous research. Therefore, the next chapter will review research on press releases which has been conducted in journalism and mass communication, as well as several related fields. Based on this review, the subsequent chapter will formulate the goals of this study and specific research questions.

# CHAPTER 3.
# REVIEW OF RESEARCH

## Selecting the Literature

The goal of this review is to find out what kinds of studies have been done on the topic of press releases. Due to the exploratory nature of this review and its broad multidisciplinary scope, I am providing a description of the steps I took in searching for relevant literature, as well as a basic analysis of the results of my search. My goal was to ensure that I use a representative sample covering research on press releases conducted in various fields and from different perspectives.

My initial plan for locating relevant literature was straightforward: I would select research papers dealing with the interaction of journalism and public relations which, specifically, involved press releases and their use by the media. I hoped to select papers which dealt with this subject from different perspectives and had different research agendas. However, a preliminary review of available literature showed that such research would be hard to find: it appears that the majority of papers dealing with the subject of press releases come from the field of public relations – a field mostly focused on applied research, primarily focusing on issues serving the practical needs of the public relations profession. To counter this limitation, which would have led to a one sided review of the subject, I broadened my initial criteria for relevancy to include research on press releases without a direct reference to their effect on or relation to news media content.

The adopted broader scope immediately posed a challenge in the form of defining the boundaries of relevancy. While most of the papers examining the relationship between press releases and news coverage came from public relations or journalism studies, papers examining press releases without a direct reference to corresponding

media coverage originated from a variety of fields, including, but not limited to linguistics, rhetoric, political communication and business. Each of these fields brought forward its own set of theoretical concepts, spanning areas from agenda setting and framing to issue management and stakeholder theory; a set of distinct terminology; and an interpretation of the overall conceptual boundaries of the field of public relations, sometimes stretching these boundaries to include anything from industrial espionage (Stauber & Rampton, 1995) to the way US presidents have handled the press (Gower, 2007).

With such conceptual breadth and variety of sources comes the increased danger of disproportionately focusing on one area, while missing another one altogether. At the same time, due to the exploratory nature of this study, it is hard to predefine the exact topical scope of the literature. Consequently, I adopted a less ambitious, yet attainable goal: to paint a general picture of existing research on the subject by starting from exploring the most likely places, systematically expanding the list of potentially relevant sources, while maintaining a balance between the opposing perspectives which define the subject and the controversy surrounding it.

**Procedure and Results**

The procedure included the following specific steps:

1. An exhaustive search of select key publications from journalism studies and public relations, most likely to contain relevant material. For the field of journalism studies this selection included Journalism and Journalism & Mass Communication Quarterly (previous title: Journalism Quarterly). For public relations, the selection included the

Journal of Public Relations Research, Public Relations Journal, and Public Relations Review.

2. A search of other publications from both fields using broad search terms (such as *public relations*, *news coverage*, *press/news release*)
3. Search of communication databases using narrow search terms (such as *press/news release* and *information subsidies*).

Table 1. *Reviewed literature by publication.*

| Publication | Frequency | Percentage |
|---|---|---|
| Public Relations Review | 9 | 20% |
| ICA Conference Proceedings | 6 | 13% |
| Journalism & Mass Communication Quarterly (Journalism Quarterly) | 5 | 11% |
| Public Relations Journal | 5 | 11% |
| Journal of Public Relations Research | 4 | 9% |
| Journal of Business Communication | 2 | 4% |
| Pragmatics | 2 | 4% |
| Communication Research Reports | 1 | 2% |
| Harvard International Journal of Press/Politics | 1 | 2% |
| International Journal of Strategic Communication | 1 | 2% |
| Journal of Communication | 1 | 2% |
| Journal of Pragmatics | 1 | 2% |
| Journal of Sociolinguistics | 1 | 2% |
| Journalism | 1 | 2% |
| Journalism Practice | 1 | 2% |
| Journalism Studies | 1 | 2% |
| PCTS Proceedings (Professional Communication & Translation Studies) | 1 | 2% |
| Public Relations Quarterly | 1 | 2% |
| Revista Latina de Comunicación Social | 1 | 2% |
| TOTAL | 45 | 100% |

In all cases, possible relevancy was initially determined by article title and abstract. Later, through a closer examination of each article, the initially collected data set of about 250 articles was trimmed down to 45 relevant papers, published between 1978 and 2012, which constitute the set of literature used for this review and are summarized in Table 1.

**Limitations**

This review of existing research, although by no means exhaustive, is representative of the literature which can be located through searching the *Communication & Mass Media Complete* online database using the outlined procedure.

The review was limited to papers dealing with print and online news media. Research on broadcast media and other forms of information subsidies, such as video news releases, as well as research conducted outside of the social sciences, although potentially relevant, pose their own unique problematics and call for very different research methods and technologies; therefore, were considered to be beyond the scope of this study.

I did not include the sources reviewed in Chapter One – i.e., *American Journalism Review*, *Columbia Journalism Review* and *Public Relations Quarterly* – because the articles published in those sources are mostly opinion and professional advice pieces. The few articles which examine data are not research papers per se: that data is mostly anecdotal and the method is not scientific, so the conclusions made by the authors may be valid only in the context of the examined data.

## Press Release Research: Common Features

Research on press releases is conducted across multiple fields, including but not limited to journalism, mass communication, business communication, sociolinguistics and language studies. In most cases, such research originates in the field of public relations.

The most common type of research on press releases is a comparison of one or more press releases to the news media coverage they are assumed to have generated. Most of these papers were quantitative and shared a similar methodology, although they differed in their degree of generalizability – i.e., one case study of a single event versus a set of press releases issued by a given company or received by a given publication, versus trend analysis across industry and/or publications. Papers also varied in sample size, approaches to measurement, and data analysis techniques, although the majority were similar in the type of research questions they addressed, as well as their overall methodology.

### Sampling

Sampling procedures varied across studies. Press releases were usually obtained directly from the organization (e.g., Hale, 1978; Martin & Singletary, 1981; Morton & Warren, 1992, etc.), from the organization's website (e.g., Gilpin, 2007), or by searching PR Newswire and Business Wire (e.g., Connolly-Ahern, Ahern, & Bortree, 2009; Murphy, 2010). In most cases, the entire population of press releases representing the period or organization was analyzed, although there were exceptions where random sampling was used (e.g., Warren & Morton, 1991). The number of press releases ranged

from 11 to 1,211 (*Mdn* = 154); most studies analyzed roughly between 50 and 200 press releases, although there were outliers: Anderson (2001) analyzed 11 press releases, while Donsbach, Jandura and Jandura (2005) had a set of 1,015 press releases, and Kiousis, Popescu and Mitrook (2007) had 1,211.

News coverage was usually collected through a purposeful sample by sampling specific publications. Reported sample sizes ranged from 62 to 6,699 (*Mdn* = 654) collected from two to 282 newspapers (*Mdn* = 5). Newspaper selections were made for various reasons: some studies used a set of dailies and weeklies (e.g., Martin & Singletary, 1981; Morton & Warren, 1992), others sampled from local papers (e.g., Hale, 1978; Morton & Warren, 1992) or compared elite papers to popular papers (e.g., Lehman-Wilzig & Seletzky, 2012). Most studies sampled articles from a set of large papers such as the New York Times or the Wall Street Journal "chosen because they wrote the largest number of articles" (Anderson, 2001), in some cases limiting the sample to the New York Times (e.g., Hong, 2008).

**Variables and Measurement**

The measured variables represented a broad set of criteria by which press releases and news articles were evaluated. The number of independent variables ranged from a few to 71 (Seletzky & Lehman-Wilzig, 2010). Press releases were classified by subject (Martin & Singletary, 1981); by their usage of passive voice versus active voice, length of paragraphs, length of sentences, and length of individual words (Warren & Morton, 1991); date sent to newspapers, company generating the releases, and point of view (Ohl et al., 1995); as well as news importance, novelty and usefulness (Lehman-Wilzig &

Seletzky, 2012). The newspapers containing the articles were characterized by type (daily, nondaily, wire services) and region (local, national, or on the basis of geographical location (Martin & Singletary, 1981).

News coverage (as well as its implied impact) was measured in column inches (Hale, 1978), length, immediacy and the existence of an accompanying photo, black-and-white or color (Seletzky & Lehman-Wilzig, 2010), etc. The content of the articles was manually classified based on its point of view (Ohl et al., 1995), its overall tone as positive, negative or both (Martin & Singletary, 1981); as well as the numerous variables specific to each study. The measurement of qualitative variables, such as point of view, was often described as determining whether "the majority of paragraphs presented the same point of view as the originating [news release], presented arguments to that point of view ... or presented a balance in point of view (neutral)" (Ohl et al., 1995).

One of the more detailed approaches to measuring the impact of press releases on the content of the news articles was offered by Martin and Singletary (1981), who measured the verbatimness of the article – i.e., the degree to which the text of the article corresponded to the text of the press release. The authors "counted the number of complete sentences appearing in the newspaper article that were identical to the news release, and divided the identical (verbatim) sentences by the total number of sentences in the article" (p. 94). Thus, if an article consisted of 10 sentences, five of which appeared in a press release, the verbatimness of this article would equal .5, or 50%. This could be also described as the proportion of text borrowed from a press release. The authors used a simplified scale categorizing the articles as "verbatim, minor change, intermediate

change, major change or complete change, depending upon the percentage of verbatimness" (p. 94).

**Data Analysis**

In most cases, quantitative data analysis was limited to counting press releases and news articles with only a few exceptions: Hale (1978) calculated correlations (a) to determine the strength and direction of the association between the press release variables and the amount of news coverage, and (b) "to determine if any characteristics were emphasized in the newspapers significantly more or less than in the [press releases]." (p. 699). Levin (2002) used logistic regression to measure the same type of associations; Schultz, Kleinnijenhuis et al. (2012) used basic computational content analysis; Gilpin (2007) conducted centering resonance analysis, which is a method which applies basic concepts from network theory to "measure centrality and influence of lexical elements" (p. 12) – which, according to the author, represents key themes in a text.

**Experimental Research Design**

There were a few exceptions among the quantitative studies in regards to methodology; two of them employed an experimental research design. Morton and Warren (1992) experimented with "localizing" the content of the press release: 174 press releases out of the 197 sent out to newspapers were "localized by being partially rewritten for each newspaper to which they were sent" (p. 1025). The goal of the authors was to determine if such "localization" affected the chances of the press release to be accepted. Bressers and Gordon (2010) conducted an experimental study, which tested whether

geographical relevance was a good predictor of press release acceptance by sending "Kansas newspapers news releases over a four-month period addressing four children's health issues" (p. 1). The authors "manipulated the degree to which information was localized in print news releases and measured the impact on selection, placement, and retention of key message characteristics in newspaper content" (p. 1). The authors utilized "commercially available database management software and minimal staff resources" to develop "a system that electronically created news releases ... each customized to every county in Kansas," which were then sent to 242 newspapers. "The same news releases were generated with more general state-level data – typical of those produced by social service agencies" – and sent to the other half of the newspapers. Both studies found that locality of news was an important factor determining the acceptance of a press release for publication (p. 3).

There were a few other studies, both qualitative and quantitative (or, rather, a mixed methods study) which do not fall under the same category with the rest. These studies will be discussed towards the end of this chapter.

## Research Goals and Perspectives

In terms of research goals and perspectives, the studies reviewed in this section can be grouped into two broad categories: studies primarily concerned with addressing the practical needs of the public relations industry – and those that don't.

Papers in the first have three common themes. One is the examination of the impact of press releases; this is accomplished, in most cases, through describing the amount of news coverage based on a selection of press releases, or conducting content

analysis, both quantitative and qualitative, to explore the framing of topics in press releases and in corresponding news coverage, and examine possible agenda-building effects. The second theme deals with improving the effectiveness of the press release as a tool for generating news coverage. This theme is manifested through examining specific criteria which may have led to the acceptance of press releases by the media. A third theme is closely related, yet is different enough to warrant a separate category: it deals with developing "formulas of success" – i.e., combinations of criteria which may be used to tweak a press release to improve its chances to make it into the news.

**The Impact of the Press Release**

The majority of press release studies underscore its importance as a tool for "getting into the news," which, according to Palser (2006) "is, of course, the Holy Grail" (p. 90) for the public relations industry. Kiousis, Mitrook, Wu, & Seltzer observed that press releases have been shown to have an impact on how the media portray political candidates (2006, as cited in Waters, Tindall, & Morton, 2010). Furthermore, according to Berger, media coverage that stemmed from information subsidies was even found to influence policy issues that were not salient on either the media or the public's agenda (as cited in Palser, 2006). Comrie (1997, as cited in Waters et al., 2010) found a moderately strong positive correlation between proactive media relations efforts and the amount and tone of resulting media coverage.

Another example of the significance of press releases is their direct impact on the company's financial health. Henry (2008) conducted a study of earnings press releases. The author explains that such press releases are issued by firms "after the end of each

quarter to announce their results for the period [and] though voluntary, are an important means by which firms communicate to investors about their financial performance" (p. 363). The author conducted an in-depth rhetorical analysis of the genre of the earnings press release, followed by a quantitative analysis exploring the relation between the stock market reaction to such press releases and measures of their stylistic attributes, including tone of text, length, numerical intensity, complexity, and various subtle promotional techniques, such as the amount of emphasis placed on particular items of information (which has been described as a framing technique by Entman (1993)), which can be achieved through placement in the text, through repetition, or both. Henry concluded that as a result of the broad discretion concerning earnings press releases, firms make many choices about the way their earnings press releases are written, and the choices of stylistic attributes influenced investors' reactions - i.e., framing the company's financial performance in positive terms had a direct impact on investors' perceptions of the company's performance.

One notable exception is a study by British scholars, publicized in two papers (Lewis, Williams, & Franklin, 2008a, 2008b). The study examined the relationship between press release content and the news media . The authors analyzed 2,207 news reports taken from UK "quality and mid-market" newspapers and found "extensive use of copy provided by public relations sources and news agencies" (Lewis et al., 2008a, p. 27). Unfortunately, neither of the two papers provides a detailed description of the method and, specifically, the procedure the authors used to identify instances of matching content. This shortcoming will be addressed in more detail in Chapter Four.

**Press releases and the amount of coverage.** Studies aiming to characterize the relationship between press releases and news coverage usually describe the amount of such coverage based on a selection of press releases in quantitative terms and, in most cases, are limited to a case study involving a single event, organization, or publication. Hale (1978) compared the coverage in newspapers and press releases of one year's decisions of the California Supreme Court. Through this study, the author tried to "establish the influence of press releases on the content or kind of information that [was] published" (p. 696). However, it is noteworthy that the specific research question - "how is the subject matter of press releases correlated with the subject matter of news accounts?" (p. 696) – does not imply causality, for correlation never implies any form of causality or influence - which creates a disconnect with the research goal of establishing the influence. Martin and Singletary (1981) examined how Pennsylvania newspapers treated press releases issued by the Department of the Auditor General, which is a state agency. The research questions in this case were purely descriptive and did not attempt to establish any causality: "to what extent were the news releases reproduced verbatim..., ... how many newspaper articles resulted from the news releases, ... was publication of a news release related to geographical distribution [or subject, or critical tone, etc.]" (p. 93). Another example of a descriptive paper is a paper by Zhang (2004) who examined how major U.S. newspapers used Saudi Arabia's press releases. Similar to the study by Martin and Singletary (1981), the research questions in Zhang's paper dealt with the number of releases, the number of resulting articles, the comparison of the quantity of articles in the New York Times to other newspapers, as well as a the usage of quotes and their location in the article compared across different newspapers. Other similar studies

included a paper by Gilpin (2007) who examined organizational press releases and mainstream media coverage of Wal-Mart in 2006, and Sweetser and Brown (2008) who examined the impact of press releases on media coverage during the July 2006 Israel - Lebanon conflict.

**Content analysis of press releases and news coverage.** Content/topic analysis can be conducted qualitatively and quantitatively. Qualitative analysis usually compares the framing of topics in press releases to their framing in news articles, often assumed to be based on these press releases. Anderson (2001) conducted such framing analysis to determine the relationship between two competing companies' press releases and the corresponding news coverage. van Hoof, Hermans, and van Gorp (2008) tried to detect the influence of press releases on the use of different frames in the 2006 Dutch Election Coverage by analyzing specific frames and the tone used in combination with those frames. Holody (2009) compared frames of press releases from the Death with Dignity National Center with the frames in the resulting newspaper coverage. Hyejoon, Byung-Gu, and Ji Won (2009) conducted a similar study on obesity-related issues.

Quantitative studies typically examine the amount of news coverage of specific topics and compare those numbers to the number of press releases or the relative amount of information on each topic in those press releases. The goal of such studies is to find evidence of agenda building or agenda setting effects. A notable example is a study by Kiousis et al.(2006) who compared press releases, media content and public opinion on political issues and candidate images during the 2002 Florida gubernatorial campaign. The paper investigated whether the salience of issues in press releases was positively related to the salience of issues in media coverage, and whether the salience of issues in

media coverage was be positively related to the perceived salience of issues in public opinion – thus, collecting evidence for agenda setting effects of press releases. Similar studies were conducted by Donsbach et al. (2005) on the influence of political public relations on media coverage in Germany; Kiousis, Popescu and Mitrook (2007), comparing companies and their financial performance; Kiousis, Soo-Yeon, McDewitt and Ostrowksi (2009) on agenda building effects in election campaigns; and Schultz et al. (2012) on agenda building effects during corporate crisis.

## Improving the Effectiveness of the Press Release

The fact that journalists treat press releases with a considerable amount of skepticism has not escaped public relations scholars who often note that "editors express a negative attitude toward press releases" (Baxter, Aronoff, Honaker, as cited in Warren & Morton, 1991, p. 1). However, whereas journalism scholars are concerned with the underlying nature of information subsidies, seeing the press release as the manifestation of this practice, public relations scholars are more concerned with the reasons for such a negative attitude displayed by the news media, and conduct their research to uncover such reasons, looking mostly at the characteristics of the press release itself. Results vary across studies, but the discovered reasons can be summarized as characteristics dealing with press release format, style, and content.

**Readability.** Baxter (1981) noted that "most editors [complain] that news releases are much too general or contain too much 'advertising puffery'" (p. 1). Other authors mentioned lack of brevity, clarity and directness, syntactical structure (Walters et al., 1994 as cited in Bollinger, 2001), poor writing quality and overall poor execution.

Warren and Morton (1991) suggested that the main problem was grounded in stylistic differences: "there may be differences in writing style between journalists and public relations practitioners" (p. 113). To support this claim, the authors compared readability levels of press releases used and not used by newspapers to determine if they differed in reading difficulty. The study used a sample of 181 press releases sent out to news media from three public universities. It used a formula for calculating readability suggested by Flesh in 1947 (as cited in Warren & Morton, 1991), as well as a set of other metrics including usage of passive voice versus active voice, length of paragraphs, length of sentences, and length of individual words. Based on the study's results, the authors concluded that "editors, when evaluating releases ... either consciously or subconsciously, consider readability to be a significant factor" (p. 117), which - according to the authors - supported the assumption that "the style and quality of the writing [were] primary reasons for accepting or rejecting releases" (p. 117). In other words, the authors suggested that it was not a matter of substance, or even quality of writing; instead the problem was explained by different approaches to writing style.

**Style.** Another examination of writing style to see whether it affected press release acceptance by news media was conducted by Moody (2009) who surveyed newspaper editors from U.S. dailies and weeklies "to determine their preference for press releases written in either the narrative style or the inverted-pyramid style" (p. 4). Moody's study received "mixed results"; however, in a later study, the author indicated that "writing style was seen as having an unquestionable link to an editor's assessment of certain press release characteristics, such as whether a release was found to be more

interesting and enjoyable, more informative, clearer and more understandable and more credible" (p. 1).

**Newsworthiness.** The actual news content of the press release, specifically the newsworthiness of the information, was found to be another important criterion. Turk (1986, as cited in Bollinger, 2001) and several other authors (Morton 1986, Minnis and Pratt 1995, as cited in Bollinger, 2001) mention several types of newsworthiness criteria, including location of the source - i.e., the local angle, timeliness of the information, as well as its impact as being the decisive factors determining publication of the press release. Bollinger (2001) describes newsworthiness (or its absence) as "the news elements' impact" and "irrelevance of content to community" as some of the reasons for which press releases are not accepted by editors.

## Searching for the Formula of Success

**What determines success.** So how does one increase the chances of a press release to be used by the media? According to research, this can be done by making the content newsworthy, improving the writing quality and changing the writing style (as well as following the numerous recommendations discussed in Chapter 1 of this dissertation). Many papers examine the criteria which may have determined a press release's acceptance by news media by measuring and analyzing the various features of a press release.

Ohl, Pincus, Rimmer and Harrison (1995) examined such criteria in the context of a corporate takeover. A similar study was conducted by Anderson (2001) where the author's goal was to determine which company conducted more effective public relations.

Morton and Warren (1992) compared the proximity of the public relations release source to the localization of the facts in the story to determine which characteristic resulted in more use of the press release. Hong (2008) examined the relationship between the newsworthiness of press releases and their publication. DiStaso (2012) studied annual earnings press releases and the corresponding news coverage to determine whether the tone, salience, length, as well as other characteristics were predictors of a press release being published.

**Designing a formula.** In some cases, authors use the results of their data analysis to predict the effectiveness of a press release (with *effectiveness* defined in terms of resulting media coverage), thus, designing a "formula of success" for press releases. "What about if practitioners could write a press release and then statistically measure it for impact?" (p. 31) – that's the motivation for Bollinger (2001), who developed a scoring method to measure the impact of press releases. The formula, having dynamic social impact theory as its theoretical foundation, takes into account multiple criteria, including style of writing (active versus passive voice), a subset of newsworthiness criteria, as well as the number of sources. The formula was successfully tested on a sample of 722 press releases generated in 1997 by the Media Relations Department of a major university.

Although newsworthiness has been found to be an important criteria in determining whether a press release will be published, there was at least one study which deliberately excluded content from its measurements and focused on form (Levin, 2002). The author designed a model for press releases using multiple variables and logistic regression to test whether knowledge of the "assumed necessary attributes" of a press release were good predictors of its success *regardless* of its content. This study is

noteworthy, for it uses scientific method to test the type of "criteria for success" which has been offered as advice in numerous opinion articles written by public relations practitioners and scholars alike, and reviewed in Chapter One.

Nevertheless, in most cases, it is the content-related features of a press release that are measured and analyzed. Seletzky and Lehman-Wilzig (2010) conducted a study and developed a formula based on various public relations content-related elements and tools and their effect on the success of press releases. Later, Lehman-Wilzig and Seletzky (2012) examined whether elite and popular newspapers were influenced differently by public relations through the analysis of detailed criteria based on the text of press releases.

Yet another similar approach was proposed by Superceanu (2011), who suggested a model of analysis based on a specific measurable criteria – informativity, a concept defined by Beaugrande and Dressler (1994, as cited in Superceanu, 2011) as "the extent to which a presentation is new or unexpected for the receiver" (p. 23). Of course, this definition is a simplified description of entropy - a concept from information theory which quantifies the information contained in a message and is sometimes explained as the degree of surprise caused by new information, and was developed by Cloude Shannon in his famous "Mathematical Theory of Communication" (1948), a concept adapted for the field of communication and known within the context of the Shannon-Weaver model of communication (Shannon & Weaver, 1963).

## Other Research Perspectives

Examining the impact of press releases and finding ways of getting a press release accepted by the media is not the only research angle applied to the subject of press releases: there are a few notable exceptions.

**Quotations in a press release.** Johnson Avery and Kim (2008) studied the use of direct quotations in press releases "to reveal the nature of quotes and use of sources" (p. 1). The study found that the use of quotations improved credibility, and, therefore – as the authors conclude – "research must delve deeper into the management of credibility and uncertainty through direct statements from sources both internal and external to the organization" (p. 12).

However, another study, conducted from a different perspective and outside the context of press release effectiveness, found that direct quotations are not necessarily proof of credibility (Sleurs, Jacobs, & Van Waes, 2003). This in-depth qualitative study examined the process of press release writing using discourse analysis in combination with ethnography and methods from cognitive psychology. In particular, this exploratory study analyzed the construction of "so-called pseudo-quotations" often used in press releases and, according to the authors, "easily copied by journalists into their own news reporting" (p. 1). The study's results indicated that quotations were, indeed, "preformulated" – i.e., drafted by the public relations agency or staff member as opposed to being genuine statements made by the executes to whom they were attributed, and were included in press releases because they were likely to be reproduced by journalists, or because the organization producing the press release assumed so.

**Transforming the press release.** Another example of research on press releases conducted outside the field of public relations are several studies by Dutch researchers (Maat, 2007, 2008; Maat & de Jong, 2012; Van Hout, Pander Maat, & De Preter, 2011). These studies compare the text of corporate press releases to the text of news articles based on these press releases. However, instead of determining "criteria for success," the authors aimed to identify how the language of a press release – especially evaluative and promotional language – is neutralized or transformed by a journalist.

The authors' primary argument is, in part, based on the concept of journalistic objectivity. According to Tuchman (1972), journalists tend to emphasize objectivity in their reports by "[gathering and structuring] 'facts' in a detached, unbiased, impersonal manner" (p. 664). Tuchman suggested several strategic procedures for ensuring objectivity, including the "presentation of conflicting possibilities" and "presentation of supporting evidence" (p. 667). Schudson (2001) later observed that a "journalist's job [consisted] of reporting something called 'news' without commenting on it, slanting it, or shaping its formulation in any way" (p. 150).

Van Hout et al. (2011) found that "reliance on ready-made source texts prompt news frames which enable reporters to write fast and efficiently while also forcing them to introduce new frames which balance the story, establish authority and maximize news value" (p. 1876). In particular, Maat (2007, 2008) and his colleagues have found evidence that journalists may, indeed, transform the language of press releases, neutralizing its promotional elements and making their report more balanced and objective. According to Maat and de Jong (2012), the transformations journalists apply to a press release result in news articles which (a) are more factual – through providing facts instead of subjective

evaluations; (b) are more impartial – through reducing positive evaluations and adding negative information; and (c) show no "bias of selection" of content – through offering "contextual information about events outside the company" (Maat & de Jong, 2012).

## The Changes Brought by the Internet …Or Not

### The Changing Media Environment

It needs to be noted that some of the reviewed studies predate the Internet (or, at least, its active usage in the last decade or two). More recent studies sometimes note that the Internet has changed the nature of interaction between news media and public relations: "The Internet makes it easy for public relations people to reach out directly to the audience and bypass the press, via websites and blogs, social media and videos on YouTube, and targeted e-mail" (Sullivan, 2011, p. 37). At the same time, Sullivan cautions that the development of online communication and news dissemination, the significance of the source of the story may have increased: "In the modern world, news does not stay in one place for long. Stories may begin on a newspaper blog or a TV website, but they soon ripple across the Internet like a splash in a pond ... The ripple effect makes the original story that hits the web – and the source of information it is based on – even more important" (p. 37).

Other researchers have observed that press releases may be no longer written for journalists alone. In fact, Waters, Tindall and Morton (2010) claim to have found evidence of a reverse of roles: "No longer are journalists passively receiving news releases and media kits from practitioners wanting to get publicity for their organization. Instead, journalists are throwing their own needs at practitioners through social media

outlets" (p. 260). To support this statement, the authors examine two online services which enabled journalists, through the use of email lists, to "ask for very specific content for stories they were working on while reaching large numbers of public relations practitioners" (p. 249). Thus, the authors argue, "rather than pitching stories to journalists and competing for printed space or airtime, practitioners are now attempting to catch media placements for their organizations by responding to journalist inquiries" (p. 249). Regardless of the validity of this study's methodology, the numbers, questionable as they might seem, speak for themselves: one of the two examined services reported having 80,000 sources and 30,000 journalists while issuing 3,000 queries per month.

**The Press Release Has Not Changed**

The changing media environment has unquestionably changed the nature of modern public relations, which has been examined in numerous studies involving public relations, the Internet and social media in particular (Bajkiewicz, Kraus, & Hong, 2011; Verhoeven, Tench, Zerfass, Moreno, & Verčič, 2012; Wright & Hinson, 2009). However, other studies show that the changing media environment *does not* mean that the nature of the press release has changed. Strobbe and Jacobs (2005) concluded that although the use and language of press releases has been rapidly changing, the "preformulation" of quotations – i.e., the pseudo-quotation examined by the authors in an earlier paper (Sleurs et al., 2003), "seems to survive more or less unharmed" (Strobbe & Jacobs, 2005, p. 291). The authors concluded that online press releases "are quite a complex sort of text that deserves further research, both in the way they are being written by public relations

professionals and the way they are subsequently being used (including both the screening by the news agency's editorial staff and the rewriting by journalists)" (p. 291).

In a more recent study, Catenaccio (2008) conducted an analysis of the press release genre, specifically, the online press release. Examining the interplay of informative and self-promotional elements of a press release, the author found that the press release is very similar to a typical news article. The author concluded that if it were not for the "explicit declaration that it is a press release, the company logo, company description, contact details … [which] crucially contribute to the identification of press releases as such, and alert the reader – be it journalists or the general public – to the fact that they originate in the company or institution, and may, therefore, be biased," a press release could be easily taken for an article: "without these, a 'perfect' press release may well be interpreted as a genuine news story, and its evaluative components as the result of external judgment" (p. 27). Thus, the concern about journalists using press releases without attribution, so passionately expressed by many journalism scholars (cited in Chapter 1), finds strong support through empirical research: "distinctions clear to the editors and staff of the nation's largest daily may be far less clear to its readers," (Ambrosio, 1980, p. 35) states an opinion article in the *Columbia Journalism Review*; Catenaccio's in-depth study confirms this sentiment through empirical research.

## Summary of the Reviewed Research

This chapter has demonstrated that although research on press releases is conducted in multiple fields, most of it comes from the field of public relations and typically focuses on the practical needs of the public relations profession.

The most common type of research has been found to be a comparison between press releases and the news coverage they are assumed to have generated. Most of the papers had similar research goals and methodology, and varied in sample size, measurement, and data analysis.

In terms of research goals and perspectives, the studies had several common themes. One theme is the impact of press releases with a focus on analyzing the amount of news coverage based on a selection of press releases, or on a comparison of how similar topics were framed in press releases and in corresponding news coverage. Other themes dealt with improving the effectiveness of the press release, including the design of formulas for calculating such effectiveness.

There were a few alternative research angles. One study examined the press release writing process and found that quotations used in a press release (and in subsequent news articles) were mostly "preformulated" – i.e., drafted by PR staff as opposed to being genuine statements by the executives to whom they were attributed. Other studies examined how the language of press releases is transformed by journalists.

Whereas some scholars have argued that the changing media environment has affected the relationship between news media and public relations, others have found that the nature of press releases remains unchanged.

Overall, the chapter has demonstrated that research on press releases is plentiful, comes from more than one field and addresses a variety of research questions. Nevertheless, existing research has considerable shortcomings which prevent it from adequately addressing the problem discussed in this dissertation. I will discuss these shortcomings in the next chapter.

# CHAPTER 4.
# THE GAP IN RESEARCH AND GOALS OF THIS STUDY

## What Is Missing in Current Research

At the beginning of this dissertation I reviewed the various concerns journalism scholars and press critics have expressed over the years in regards to journalists using press releases as a news source with little or no edits. These scholars say this practice may lead to news media serving the needs of the public relations industry and its clients instead of serving the public. These concerns were, for the most part, opinions backed by mostly anecdotal evidence, and could have been explained, at least in part, by the century-old antagonism between the two professions. However, a review of articles on press releases written by public relations practitioners over a similar period of time suggested that most of those opinions had considerable merit to them.

Furthermore, through an examination of relevant mass communication theory, such as agenda-setting, as well as the broader concepts of agenda building and framing (conceptualized as second-level agenda-setting, or frame setting), a theoretical framework emerged explaining the process of information exchange between the two fields, and the goals of each side. As a result, the concerns voiced by journalism scholars and, to a certain extent, confirmed by public relations authors, were given theoretical support and revealed a broader issue: publishing press releases as news may lead to organizations represented by the public relations industry having disproportionate favorable coverage in the news, which may affect, to some extent, the way the media's audience thinks of these organizations. The core of the problem is that news media are expected to be the watchdog watching over these very organizations and offering the audience an impartial account. Essentially, if the theoretical reasoning is correct, this very

role of news media in a democratic society may be compromised by excessive reliance on public relations outlets for news.

Therefore, an investigation of the extent of such practice – i.e., of the extent of press release use by news media, especially with little or no edits – is critically important for mass communication scholarship, as well as for preserving the media's ability to produce independent news accounts. Furthermore, before any steps towards resolving this troubling issue can be planned, substantial evidence is needed demonstrating the scope of the problem and the specific details pertaining to how such practice occurs while remaining virtually unnoticed by the general public.

Such evidence could be supplied by studies examining press releases together with news articles known to have used these press releases as a source, with a sample large enough and representative of the overall population, so that valid generalizations can be made. However, a comprehensive review of research dealing with press releases suggests that these kind of studies are in short supply.

**Shortcomings of Previous Research**

A review of previous research on press releases has revealed a number of troubling shortcomings. Most importantly, the majority of such research originated in the field of public relations, and therefore, was focused on solving the practical needs of the public relations industry – i.e., findings ways to improve the chances of a press release being published by news media. Overall, the following shortcomings can be listed:

1. Studies addressing the practical needs of the public relations industry do not contribute to a broader understanding of the problem.
2. Most of the studies deal with small sample sizes; therefore, their findings cannot be easily generalized.
3. There were a few studies which used relatively large samples; however, those samples were not representative: researchers usually focused on several large newspapers because of the accessibility of that data. Other studies focused on media from a particular geographic location. None of the studies boasted a sample large enough and representative enough of both media and public relations sources, to be definitive.
4. The several studies which examined how journalists transformed press release content before using it in articles originated in the Netherlands and, thus, were based on Dutch media. The one study which found extensive evidence of press release content used in news media was based on UK media. None of such studies were based on US media or public relations.
5. Only one study directly addressed the issue of press release content being used verbatim; however, that study was from 1981, which makes it quite dated: it's more than 30 years old.
6. There were no studies on attribution – i.e., news media acknowledging the press release as the source of news.

**The Missing Connection Between Article and Press Release**

However, the most significant shortcoming is that only a few of the studies comparing press release content to news media content reliably demonstrate that the articles are, indeed, based on the press releases in question. While some authors obtained their samples (or part of them) from clipping services (e.g., Warren & Morton, 1991), thus relying on unspecified procedures used by the clipping service for identifying relevant press coverage, most studies employed keyword search to locate articles and/or press releases of similar content in databases like Factiva and LexisNexis, as well as company websites (e.g., Brechman, Lee, & Cappella, 2011; Gilpin, 2007; Holody, 2009; Zhang, 2004). Keyword search locates documents with relevancy defined in terms of content similarity, which is established based on the provided keywords.

Content similarity, regardless of how similar the press release and the news article are, does not imply causality. Similarity alone is not enough to reliably establish a directional link between two texts, implying that one was used to create the other – i.e., that the text of the press release has been actually used as a source for the news article. A newsworthy event can be communicated through means other than a press release – such as an interview, a press conference, a website, word of mouth, or even personal communication – not to mention the numerous tools and communication venues provided by social media and, undoubtedly, used by organizations to communicate their messages to the public. In other words, there are numerous ways in which a journalist can obtain information on a topic that also happens to be announced through a press release. Thus, comparing topics is simply not enough; a better, more reliable method of establishing a relationship between a press release and a news article is needed.

However, topic similarity remains a common way to establish a definitive relationship between a press release and the news coverage it is assumed to have caused. Furthermore, most of the studies comparing press releases to news coverage do not provide much details about the procedure they used to determine whether an article was, indeed, based on a press release – i.e., whether *the actual text* of the press release was used by the journalist who wrote the article. With rare exception, this connection between article and press release is established through a process which is described only in general terms. Zhang (2004) explains that "stories were read to see whether they used the press releases" (p. 10); Seletzky and Lehman-Wilzig (2010) only says that the "news items [were] checked" (p. 253). One of the more detailed descriptions of the procedure, offered by Brechman et al. (2011), stated that they examined press releases and news articles covering scientific research: the authors made "additional sweeps … to eliminate irrelevant articles" (p. 501) and limited the sample to those texts which "contained traceable reference information to published research" (p. 502), after which they conducted the same steps for press releases.

Many studies do not even state their procedure for identifying a definitive relationship between a press release and an article; the authors, apparently, assume that since the article is retrieved from an electronic database using the same keywords as the press release, or even is on the same topic, it is, therefore, based on the press release or is affected by the content of that press release. Thus, it is not entirely clear what ties the press releases to the news articles in such studies.

Yet, the findings which stem from the assumption of such a definitive relationship are formidable. Lehman-Wilzig and Seletzky (2012) and Seletzky and Lehman-Wilzig

(2010) use their results to discuss and quantify the successful publication of press releases, Zhang (2004), Sweetser and Brown (2008) and Ohl et al. (1995) speak of the agenda-building role of news releases, Kiousis et al. (2006) and Kiousis et al. (2009) discuss the agenda building and agenda setting effects of information subsidies by exploring linkages between press releases, media coverage and public opinion, Anderson (2001) discusses the influence of press releases on media coverage but not the news content, whereas Hyejoon et al. (2009) and van Hoof et al. (2008) discuss the influence of press releases specifically on news content, etc.

Such assertions of press release influence are found in most of the reviewed studies – and that is not surprising: in most cases the whole point of examining press release content alongside news media content is to compare them, to make a meaningful conclusion about their relationship, and to suggest their broader significance. Consider the conclusion offered by Lewis et al. (2008a):

> When news coverage was analyzed, findings revealed that 30 per cent of published items were wholly dependent on agency copy with a further 19 per cent strongly derivative from agency materials. In a further 13 per cent of stories agency copy was evident along with information from other sources, while 8 per cent of items used mainly other information … Newspapers make little acknowledgement of this reliance on agency copy even when they publish such materials in more or less verbatim form (p 29-30).

**Establishing a Connection Between Article and Press Release**

There are a few exceptions – i.e., studies that go into the trouble of specifying the method for establishing a definitive relationship between the press releases and news articles in their data sample. Unfortunately, such studies have to rely on small samples which are not representative of the overall population and, therefore, cannot be easily generalized – a shortcoming justified by a major methodological issue which I address in

this dissertation. Nevertheless, by establishing a causal relationship between press release and news article, these studies setup a reliable conceptual foundation, which is essential for any subsequent arguments and conclusions drawn with regard to the relationship between press release content and news coverage.

Alcoceba-Hernando's (2010) study is one of these exceptions: the author uses explicit attribution to the press release as evidence of a relationship between the article and the press release. The author explains that "from this selection of news [he] only took into account the news that appeared in the press and made reference to the press releases" (p. 4). Unfortunately, while this procedure guarantees the link between the press release and the article, it does not take into account articles which do not explicitly mention the press release as a source – which, according to many journalism scholars, *may be the rule rather than the exception* in media practice (see Chapter One).

A few studies established a relationship between press release and news article by conducting in-depth textual analysis, comparing individual words and phrases from the two texts side-by-side (e.g., Maat, 2007, 2008; Maat & de Jong, 2012). These studies provide a detailed analysis of how press release content is used and transformed by journalists in the case of a given press release or company, or a type of corporate announcement, such as a product launch press release.

### Pros and Cons of Using Verbatim Text Matches to Detect Relationships

There have been at least two studies which employed a method which, compared to other studies, appears to be one of the most reliable ways to establish a causal

relationship between a news article and a specific press release: identifying verbatim text matches between the two texts.

Martin and Singletary measured the degree of "verbatimness" – i.e., the proportion of press release content used in news articles without change (1981). The authors took a rather basic approach and measured overlapping content using complete sentences as unit of analysis, discarding any partially matched sentences. Morton and Warren (1992) offered a similar approach to establishing a reliable connection between a press release and a news article. Like Martin and Singletary, the authors based their argument on verbatim matches between press releases and news articles: "a press release was considered used if it was published substantially in its original form" (p. 1026).

In both of these cases, a verbatim text match between a press release and an article enabled researchers to tie a specific press release to a specific article beyond reasonable doubt. Certainly, this approach has considerable limitations. Martin and Singletary (1981) considered only complete sentences., which means that any partial sentence match, even if it was highly indicative of a relationship between the two texts, was not taken into consideration. It is reasonable to expect that a journalist copying the text of a press release verbatim into a news article he or she is writing, at the very least, would attempt to make the similarity of the texts less evident – and breaking up sentences is the obvious, simplest way to do that. Thus, chances are, the authors missed a considerable amount of connections between press releases and news articles. Morton and Warren (1992) do not specify the exact method of establishing a verbatim match. However, regardless of what procedure they followed, their method – like the one used by Martin and Singletary – was bound to have considerable limitations as well.

The problem of verbatim matches is that it is a very strict requirement. As I will demonstrate in the following chapter, there are various ways of identifying a verbatim match, yet none of them offer a perfect balance between a definitive clue pointing to a connection between two texts, and the flexibility of a partial match suggesting the possibility of such a connection. A long enough sequence of words, such as a sentence, may point to a connection between a press release and a news article in one case, but it fails to uncover the same connection in another case, where the sentence is broken up by the insertion of one single word. Thus, no matter how this method is designed, it is bound to miss a connection between a press release and a news article if the text borrowed from the press release is completely paraphrased.

One solution could have been to account for paraphrased text. However, without a verbatim match, it may be harder to argue that the text of the news article is, in fact, a paraphrase of the press release and not some other source. As I have mentioned earlier in this chapter, there are numerous ways in which content created by a corporation may reach a journalist, and a press release is but one possibility. Thus, paraphrased text only suggests a similar (or even very similar) topic – a topic that might have originated from a press release, or an interview, or a press conference, or a social media website – the scope of possible sources is simply too broad for a definitive conclusion. Therefore, paraphrased text can be used to answer broader questions about public relations content in general (or even still broader: about all content generated by a corporation). However, using it to establish a definitive connection between a specific press release and a news article for the purpose of examining the use of press release content – not PR content in general – appears to be less reliable than using verbatim matching.

**Why Finding Verbatim Text Matches is a Difficult Task**

What would be a reasonable way to establish that an article is, indeed, based on the text of a press release – i.e., that the journalist actually *used* the press release in some way when writing the article? Based on the discussion in this chapter, two conditions come to mind: (a) have the article appear soon after the press release; and (b) have a verbatim match – i.e., a long enough sequence of words (not a sentence though) appear in both the press release and the article matching each other verbatim. These two conditions could be used as sufficient evidence that the latter is based on the former.

The first condition is taken into consideration by most researchers. The second is more of a problem: Brechman et al. (2011) started with an initial sample of 5,876 articles (retrieved from LexisNexis through keyword search) and went through a thorough procedure eliminating irrelevant texts, that reduced their sample to only 20 pairs of matching articles and press releases. Lewis et al. (2008a), on the other hand, do not describe the process they used to identify matches between 2,207 news articles and an unspecified number of corresponding press releases. However, considering their very specific claim about finding that "30 per cent of published items were wholly dependent on agency copy" (p. 29), it appears that they have located definitive evidence of press release content in the texts of approximately 728 news articles (which is 30% of their total sample of 2,207).

What's more, Lewis et al. (2008a) acknowledge that at least some of their data demonstrated that there were verbatim or almost verbatim matches between press release and article: "Newspapers make little acknowledgement of this reliance on agency copy even when they publish such materials in more or less verbatim form" (p 29-30). The

problem I see in drawing such specific conclusions has to deal with the sample size of the study: it appears that the authors located "more or less verbatim" matches in the process of examining 2,207 news articles.

Identifying matching sequences of text is a very difficult task. Consider an article and a press release, each 500 words long. Let's assume that a matching sequence of 10 words is enough reason to establish a relationship between the article and the press release. Thus, to establish a relationship between this article and press release, one needs to locate a sequence of at least 10 words which appears both in the article and in the press release. If this matching sequence appears at the beginning of the text, or even at the beginning of a paragraph, it will be easily spotted. If it's a particularly long sequence, such as a paragraph, it will be easy to notice it. However, if it's 10 words (or even more than that but much less than a paragraph), spotting them in both texts might be quite difficult. The matching 10 words may occur in different places: it could be the first paragraph of the article and the last paragraph of the press release or vice-versa; they may both occur at the beginning, in the middle, or towards the end of both texts – the number of possibilities is quite large. A simple exercise in combinatorics will yield the precise number of such possibilities: there are 500 - 10 + 1 = 491 ways to pick a sequence of 10 words in a 500-word text; each of these ways may correspond to 491 potential matching sequences from the other text; thus, the total number is 491 x 491 = 241,081. There are 241,081 ways in which a 10-word sequence can match in two 500-word texts (see Figure 1). But Lewis et al. (2008a) examined a sample of 2,207 news articles, most of which, most likely, exceeded 500 words.

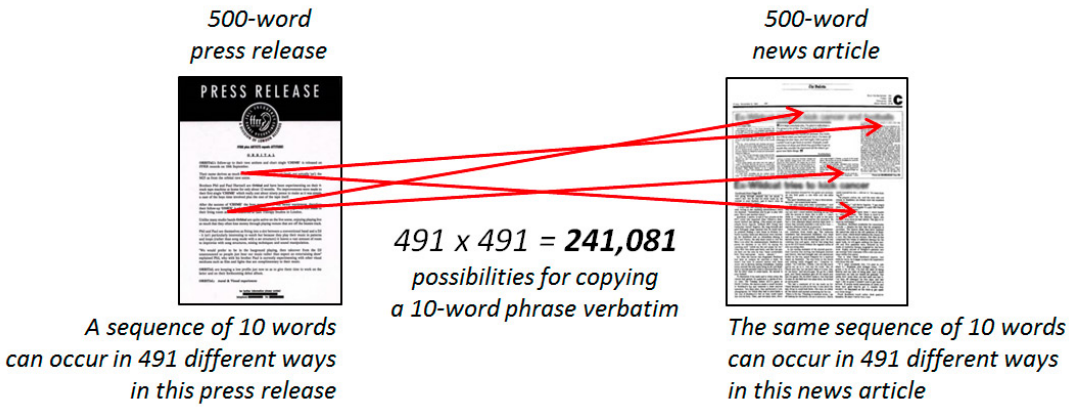



*Figure 1*. The task of finding verbatim text matches.

Certainly, when one knows what to look for (i.e., what kind of text, what topics, what keywords, etc.), it simplifies the task; besides, an experienced coder should be able to look for several words at a time – i.e., a human does not have to follow the explicit word-by-word sequence of steps which would have been followed by an algorithm. Nevertheless, the number of *potential* combinations demonstrates the potential magnitude of the problem. Even with an army of research assistants a problem of this size cannot be solved manually – and yet, Lewis et al. make no mention of computation. Thus, it is not clear, how could have the authors conducted such a large number of text comparisons between articles and press releases to reach their conclusive results? Unless, of course, they were not making any verbatim comparisons in the first place.

The topics, the keywords, and even the occasional quote in an article can, indeed, match a given press release, but even so, one cannot conclude that the news article relies on the press release without an in-depth examination of the texts in question (akin to the textual analysis conducted by Maat (2007, 2008); Maat and de Jong (2012) and Maat and

de Jong (2012). Furthermore, inferences about the field in general can be made provided the sample size is large enough for those inferences to be statistically meaningful. Finally, a sample needs to be representative of both news media and public relations sources for such inferences to be valid.

The combination of these requirements constitutes an obstacle which explains the reason why such studies have not been conducted – i.e., the reason for the gap in research: the required level of detail (which has been possible only through manual analysis of text) combined with the size of the problem (i.e., the number of text comparisons or calculations) make the task manually intractable. In this dissertation, I argue that such obstacles can be overcome with a little help from the field of computer science.

## Computer Science and Analysis of Text in Mass Communication

### Defining the Scope

I use the term "computer science" in the title of this chapter intentionally: for even though my discussion deals with computation as a method, the concept of computation refers to computing – i.e., performing calculations. Thus, strictly speaking, computation is used in most research, especially in quantitative studies (although today the term is commonly used to refer to calculations performed by a computer). The approach I am describing is not so much about computing per se – after all most studies compute something. However, whereas a typical quantitative study does not raise the question about what exactly to compute (and, of course, that computation may be very complex), this study does pose such a question; in fact, the reason for the study itself is the question of whether or not a particular problem in journalism studies, which is too large to be

solved manually, can be addressed computationally. Thus, this study, at least in part, is about the application of "computational thinking" as a problem solving approach – a foundational concept in the field of computer science, which, in the words of Jeannette M. Wing, is "the study of computation – what can be computed and how to compute it" (Wing, 2006, p. 34).

There are numerous ways in which computation can be applied to problems in journalism studies and a dissertation cannot cover such a topic in any reasonable depth. Thus, this study is neither an attempt to provide an overview of the near-infinite cross-product of journalism research problems and computational methods, nor it is an attempt to provide a general introduction to the field of computer science. Instead, I will focus on the specific goals of this study, namely, measuring the reliance of news media on the press release; I will demonstrate the utility of computer science in carrying out a range of tasks which, due to their problem size, have been the primary, if not the only, obstacle preventing reliable investigations of this relationship on sufficiently large and representative data samples.

## Content Analysis: Methodological Context for Computation in Mass Communication

In journalism (and mass communication) research, application of computational methods has been explored in the context of content analysis, a research method which began in the early 20$^{th}$ century with quantitative newspaper studies where the volume that the text of an article occupied was calculated by counting the width of the column in inches (Krippendorff, 2012). Starting with the 1940-50s quantitative analysis of

newspapers gradually began to transform into more systematic, scientific analysis of content. This trend can be ascribed to social scientists entering the field of media studies. They began to empirically analyze news media and ask complex questions, often grounded in theory and demanding rigorous statistical analysis. This approach required precise numbers, so measuring content in inches was given up in favor of measuring the frequency of occurrences and co-occurrences of words and phrases (Krippendorff, 2012).

A formal definition of the method of content analysis was formulated first by Berelson and Lazarsfeld in 1948 (Berelson & Lazarsfeld, 1948) and later by Berelson in 1952 (Berelson, 1971), who defined it as "a research technique for objective, systematic, and quantitative description of the manifest content of communication" (p. 18). This definition lists four key characteristics of content analysis:

1) Objective. Objectivity implies replicability: the results should depend on the procedure, not on the analyst.
2) Systematic. The procedure, defined as a set of explicit rules, is applied to all the content being analyzed, with all content being treated in the same manner.
3) Quantitative. The source data is described through frequency counts, which are considered to be an accurate representation of the content. Quantification allows statistical data analysis and making statistical inferences about the overall population.
4) Manifest content: – i.e., apparent content; content as it appears in the text, as opposed to how that content may or may not have been intended to appear.

The objective and systematic characteristic of content analysis have withstood the test of time and are included in various definitions of the method. The quantitative and

manifest content criteria, on the other hand, have been repeatedly questioned over the decades. In the following sections, I provide a brief summary of the main points of the debate over each of these two criteria, for these debates may have been one of the reasons for which the method of content analysis is fragmented across disciplines with different fields having their own view of the method and their own set of tools for its application (For a detailed account on content analysis, see Budd, Thorp, & Donohew, 1967; Carney, 1972; Krippendorff, 2012; Neuendorf, 2002).

**Quantitative versus qualitative.** Qualitative method can be described as a means for exploring and understanding the meanings individuals ascribe to social phenomena. It "seeks to preserve and analyze the situated form, content, and experience of social action, rather than subject it to mathematical or other formal transformations" (Lindlof & Taylor, 2002, p. 18). Compared to quantitative method, qualitative method is capable of producing data with more depth, representing a broad contextual picture of the phenomena. Quantitative research, on the other hand, is based on quantifiable measurements and is focused on objectivity and replicability and the positivist approach to social phenomena (Bryman, 1984). In quantitative method, the researcher distances himself or herself from the observed phenomena to minimize the possibility of bias and subjective interpretation.

The difference between these methods can be viewed as a disagreement based not so much on data gathering techniques, as on treatment of social reality. Qualitative approach assumes there is no observable objective reality in social scientific studies due to the complexities of measuring social phenomena; that social reality is "constructed" rather than scientifically discovered; and, therefore, interpretation is key to understanding

data. Quantitative research is based in the assumption that there exists an objective reality which can be examined in terms of quantifiable indicators.

Berelson (1971) argued for quantification being a requirement in content analysis. Later, Holsti offered a definition which had no restriction on the quantitative (or qualitative) description of manifest content: content analysis was defined as "any technique for making inferences by objectively and systematically identifying specified characteristics of messages" (Holsti, 1969, p. 14). Krippendorff (2012), however not only questioned the distinction between quantitative and qualitative approach in content analysis, but considered all text analysis to be qualitative. Krippendorff argued that since such analysis is conducted through reading text, it is qualitative by definition – since "all reading of texts is qualitative" (p. 22).

It needs to be noted, however, that Berelson indicated in his book, published six decades prior to Krippendorff's argument, that reading and content analysis are very different activities in the first place (Berelson, 1971). Besides, even if text were qualitative data, like most other data in qualitative analysis, such data can be quantified. Holliday (2007) cautions: "Social research is a complex area, and attempts to divide it into hard categories will always suffer from oversimplification. Qualitative research will always involve quantitative elements and vice versa" (p. 2).

**Manifest content versus latent content.** The other objection to Berelson's definition of content analysis deals with manifest content. Berelson (1971) described content, or communication content, as a body of meanings expressed through symbols. He explained that in the context of media research described by Harold Lasswell as "who says what to whom with what effect," content analysis represented the *what* (p. 13).

Berelson limited the scope of content analysis to syntactic and semantic dimensions of language: "content analysis proceeds in terms of what-is-said, and not in terms of why-the-content-is-like-that … or how-people-react" (p. 16). In other words, the study of manifest content was meaningful in its own right, according to Berelson.

George (1973) in his Propaganda Analysis, first published in 1959, was one of the first to question the validity of content analysis limited to manifest content. Soon a new perspective on content analysis began to develop, according to which the message was not "frozen" into the text of the message (Krippendorff, 2012). Today, Krippendorff argues that the message is not a container for one meaning; that describing a message in terms of what was said, how, and to whom fails to acknowledge the analyst's own concept of what constitutes the appropriate reading of the text. Messages could have different emotional connotations for different readers; they could change over time, and they could be ambiguous altogether (what does the word "democracy" really mean?) (Krippendorff, 2012). Krippendorff even compares the content analyst with a psychiatrist who interprets clients' stories. However, he notes that the analyst must state whose norms or attitudes they are using for interpretation – then such technique would be replicable and, therefore, valid – which, according to Krippendorff are essential for content analysis, for he defines the method as a research technique for "making replicable and valid inferences from data to their context" (p. 24).

And again if we consult the foundations of the field, we find an argument counterbalancing Krippendorff's statement. Berelson (1971) admitted – as early as in 1952 – that one message could hold multiple meanings. However, he noted that there

could be different levels, or types, of communication content with analysis of manifest content being applicable to some and not applicable to others:

> If one imagines a continuum along which various communications are placed depending upon the degree to which different members of the intended audience get the same understandings from them, one might place a simple news story on a train wreck on one end (since it is likely that every reader will get the same meanings from the content) and an obscure modern poem at the other (since it is likely that no two readers will get identical meanings from the content) (pp. 19-20).

Berelson (1971) also suggested that there was point on this continuum "beyond which the 'latency' of the content (i.e., the diversity of its understanding in the relevant audience) is too great for reliable analysis" (p. 20).

Back in 1952, Berelson observed that the problem of inference in content analysis – which encompasses the debates over manifest versus latent content and quantitative versus qualitative approach – was no different from the same problem in social science in general (Berelson, 1971). I argue that this sort of debate which has been carried on throughout decades, continues to affect the methodological development of computational content analysis in mass communication and journalism research, preventing it from actively adopting solutions which have been conceived, developed and successfully implemented in other fields years ago.

## Computer-Assisted Content Analysis in Mass Communication

In mass communication, as well as journalism studies in particular, the usage of computation in content analysis is usually referred to as computerized content analysis (Stempel, Weaver, & Wilhoit, 2003) or computer-assisted content analysis (CATA) (Krippendorff, 2012). Krippendorff traces the beginning of such practice to the late 1950s

which saw "considerable interest among researchers in mechanical translation, mechanical abstracting, and information retrieval system" (p. 19). Some of the milestones, according to Krippendorff, where the first reported instance of computer-aided content analysis by Sebeok and Zeps in 1958 which applied information retrieval to analyze folktales; a 1960 paper by Hays which discussed a computer system for analyzing political documents; and the initial version of the General Inquirer system designed by Stone and Bales, and publicized as a "groundbreaking book" in 1966 (Krippendorff, 2012).

Today there are numerous applications of computer-assisted content analysis in mass communication. Popping (2000) identifies three main approaches to such analysis. One is thematic text analysis, where texts are quantified as counts of words and phrases that were classified according to a set of content categories. Thematic text analysis, according to Popping, allows the investigator to determine what concepts occur in texts. The second approach is semantic text analysis, which involves not only the identification of concepts but also the relationships among them. In semantic text analysis an encoding process is needed to acquire a semantic grammar that specifies the relations among themes; the encoded data is then used for making inferences from the texts. The third approach is network text analysis, which is derived from the semantic links among concepts and involves constructing networks of semantically linked concepts (Popping, 2000).

Krippendorff (2012) offers a different kind of classification of computer-assisted methodologies: text analysis, computational analysis, webgraph analysis, and support for qualitative text analysis. According to Krippendorff, text analysis deals with analyzing

raw text with the purpose of deriving quantitative data about the text – i.e., frequencies of occurrences or co-occurrences of words and phrases. Computational content analysis describes a process based on a theory of meaning which is then computed. Webgraph analysis is what is commonly referred to as network analysis, in this case – analysis of linked text. Finally, support for qualitative text analysis are, essentially, programs which simplify the labeling or other kinds of organizing of text (Krippendorff, 2012).

However, despite elaborate classification typologies of the method, the development of computer-assisted methodology in mass communication research may have been hindered by the conceptual debates over the definition of content analysis and its limitations in regard to the debated requirement of quantification of data and the even more debated focus on latent as opposed to manifest content. It is widely recognized within the mass communication community that computers can process large volumes of text in a relatively short period of time, that they can offer features for organizing and searching text, that they can provide more uniformity, and can reduce the time and cost of analyzing text – provided the analysis is to be *quantitative* and focused on *manifest content*. However, if the goal of such analysis is interpretation of meaning, the utility of automating word counts will be downplayed, except for the cases where meaning is statistically modeled based on the language features of the content – which is not a typical methodology in journalism or mass communication research. Thus, although the application of computation is not a new approach in content analysis conducted in the context of journalism or mass communication research, its development has considerable shortcomings.

A minor, yet rather common problem in mass communication literature on content analysis is the careless use of terminology from computer science and related fields. There is no such thing as "neuronal learning nets" in computational text analysis (Krippendorff, 2012, p. 247) – there are neural networks. The "so-called" (Krippendorff, 2012) support vector machines are a common machine learning technology in computational text analysis and due to their excellent performance have been used more often than neural networks.

A much bigger problem in such literature is that the authors, who are highly respected authorities on the method in the field of mass communication, seem to have fragmented knowledge of what computers are really capable of doing, and how their capabilities are used in other fields when dealing with processing text. As a result, alongside deep and thoughtful insights into the method of content analysis, as well as mentions of non-trivial computational techniques (both neural networks and support vector machines are used in machine learning – an area of artificial intelligence, which is a branch of computer science), this literature offers some recommendations which are misleading. For example, Popping (2000) offers detailed instructions on how to how to manually develop a text vocabulary and count all the occurrences of a word identifying and searching for all its variations. From Krippendorff (2012) we learn that that a text often needs to be manually prepared before being processed by a computer. Specific preparation steps, according to Krippendorff, depend on the specific "computational content analysis program" the researcher might use. Such manual "preediting" steps include correcting misspellings, "replacing foreign characters which characters [the system] recognizes" (Krippendorff refers to ASCII – which is a character-encoding

scheme and has nothing to do with a distinction between "foreign" and "domestic" characters), "introducing special markers to indicate syntactic distinctions" (adding double periods after sentences, carriage controls after paragraphs, etc.), "replacing pronouns with proper names and indirect references with direct ones," "marking the syntactical functions of words" (i.e., assigning part-of-speech tags, but manually), decomposing longer sentences into smaller units of text" – the list is quite long (p. 238). Krippendorff is correct by assuming that a computer, by itself, won't know that all the different forms of a word are supposed to be counted as the same word (i.e., in most cases, the words *write*, *wrote*, *written* will be expected to be processed as the verb *to write*). However, what the author fails to see is that identifying all forms of a word is one of the many tasks a computer is meant to do. Needless to say, the "data cleaning" procedures suggested by Krippendorff negate the whole premise of using a computer in the first place! Most of these tasks have been automated in other fields, specifically in computational linguistics (or natural language processing) – a field Krippendorff mentions in his book, yet fails to give it the treatment it deserves (for a description of such techniques, see Jurafsky & Martin, 2008; Manning & Schutze, 1999).

**Computational Text Analysis in Other Fields**

Automating some parts of content analysis is, certainly, not the only task which can be assigned to computers in regards to mass communication research. Researchers from various fields have been experimenting with computational text analysis over the years: some have tried applying data clustering algorithms to framing analysis (Murphy, 2001), others have applied centering resonance analysis (Gilpin, 2007) and network

theory (Tremayne, 2004; Tremayne, Nan, Jae Kook, & Jaekwan, 2006) to problems in mass communication. Computational linguists combine linguistic data with computer science to develop rule-based and statistical models of natural language used in a broad scope of applications. Yet another direction of research related to journalism studies is the analysis of the flow of news and information between mainstream media and social media (Leskovec, Backstrom, & Kleinberg, 2009; Tremayne, Weiss, & Alves, 2007).

Overall, the combination of computer science, linguistics, and statistics, coupled with the body of knowledge of a particular academic field has been successfully applied in various fields, including economics (e.g., Gentzkow & Shapiro, 2007), political science (e.g., Monroe, Colaresi, & Quinn, 2008) , business (e.g., Kogan, Levin, Routledge, Sagi, & Smith, 2009) and psychology (e.g., Tausczik & Pennebaker, 2009). There have been individual attempts to utilize such approaches in journalism studies and mass communication research in general, but there have been very few so far.

In a text written for a broad audience, Oard (2009) provides a classification of the types of computational text analysis which is a convenient way to envision "computer-assisted" analysis of text beyond the field of mass communication. Oard's classification includes paraphrase, extraction, classification, clustering and information retrieval:

1) Paraphrase: automating the conversion of part of a text, which is a set of ideas – to another expression of the same ideas – used in the context of summarization and machine translation. Summarization is done to express the main ideas of the text more succinctly (for example, summaries of search results returned by a search engine). Machine translation is, essentially, translating – i.e., generating a phrase in another language corresponding to the initial phrase.

2) Extraction: extracting spans of text important for some purpose (for example, locating named entities in a text or finding structured data in web pages – such as items and prices, or even identifying context relevant to a given word or set of words).
3) Classification is automatically assigning some text to a class from a known set of classes. Oard points out that the combination of extraction and classification can be used for a new task altogether: automatic coding for themes in a text.
4) Clustering is the same as extraction and classification, except that it is done without "human supervision" and determines the classes, or categories, on its own. The downside of this approach is that it is often difficult to generate meaningful clusters and their descriptions (summarization can help only to some extent).
5) Finally, Oard lists information retrieval as a field which deals with storing and retrieving very large amounts of information (pp. 36-37).

## Computational Methods and The Goals of this Study

### The Building Blocks of Computational Text Processing

The typology offered by Oard (2009) in the previous section is, certainly, not the only way of describing all the different groups of computational text analysis; still, it is convenient as a generic description of how computation can assist in text analysis. However, the important part about all these various applications is that all of them use a common foundation – a foundation essential to any application of computation to research in mass communication, as well as other fields.

The "building blocks" of any computational system that deals with processing text are, essentially, the same and originate in computer science. These "building blocks" do not constitute a specific computational method; they are far more general than a given method, or even the type of methodology we refer to as computational text analysis. Instead, they constitute the foundation which makes computational text analysis possible.

Consider, for example, a study which deals with text obtained from a website. Regardless of the specific academic field which constitutes the research context of this given study, regardless of what kind of data analysis we are to conduct on this text, there are a number of steps which precede the data analytical stage of our study. Such steps may include discovering a potentially relevant web page, programmatically accessing that web page and collecting its source code, stripping out any markup tags and extracting the actual text. The same, or very similar tools will be utilized, and the same approaches will be considered. The extracted text, most likely, will be brought to a common encoding, it may be split into sentences, individual phrases, words or, maybe, punctuation; the words might be stemmed, a vocabulary (or concordance) may be generated, etc. These kind of steps are common to any text-processing system. Combined together, they produce tools for collecting, processing, and analyzing text computationally, which may be custom-tailored to a specific research task.

Comparing a press release to a news article with the purpose of finding matching text is one such task. Comparing a large set of press releases to a large set of news articles is an extension of this task, which is impossible to complete manually within a reasonable period of time, but becomes quite realistic when executed by a computer. A detailed description of the "building blocks" which enable each computational step

involved in such a task is, to be sure, beyond the scope of this dissertation. Furthermore, such a description belongs in a computer science textbook – not a scholarly paper. However, by demonstrating such basic steps in the context of this study, I intend to provide an example of their application to a research problem in mass communication; such a demonstration, to the best of my knowledge, has not been done in this context.

**Computing Verbatim Matches to Identify a Relevance Sample**

Although a demonstration of the basic steps which serve as prerequisites to any computational text analysis is an important part of this dissertation, the primary goal of computation in this study is the generation of a relevance sample.

A relevance sample, according to Krippendorff (2012), is a sample reduced to the data most relevant in the context of a given research question. Krippendorff has noted that a relevance sample is especially appropriate for content analysis, where it makes sense, especially considering "very large electronic text databases and the internet," to reduce the population from all documents to relevant documents only (Krippendorff, 2012, pp. 120-121).

It appears that this sampling procedure is especially suitable for this dissertation, where the task is to address the methodological shortcomings of previous research which have often prevented scholars from examining the problem of press release content in news media in sufficient detail.

Earlier in this chapter I explained why establishing a definitive connection between a specific press release and a news article for the purpose of examining the use of press release content is best achieved through finding verbatim text matches in press

releases and news articles. I then discussed the methodological issues preventing researchers from using this approach on adequate data samples due to its complexity. However, with computation, using this approach becomes possible.

In the next chapter, I will describe how computation can be used to process a very large data set and detect relationships between press releases and news articles through locating verbatim text matches of sufficient length. Thus, the significance of this computational approach lies primarily in the reduction of a potentially very large data set consisting of press releases and news articles exhibiting a similarity in topics – which has been shown to be not enough to draw any conclusions about a causal relationship between two texts – to a much smaller sample, consisting of press releases and news articles, explicitly tied to each other by sequences of matching text – which makes this sample much more reliable in comparison with the initial data set.

Furthermore, the relevance sample is expected to be sufficiently small to be suitable for both computational and manual textual analysis. This is especially promising considering that a relevance sample contains the data most relevant to a specific research question. In the case of this study, computation applied to locating verbatim text matches between press releases and news articles generates a sample, potentially containing the most interesting and the most significant data; data which is most likely to contain instances of paraphrased text which warrants the researcher's close attention.

Thus, although a method relying on verbatim text matches cannot measure the degree of press release influence on news content, it puts very large initial data sets within the grasp of qualitative mass communications researchers by pointing them to the most critical data, which demands careful in-depth textual analysis. In subsequent

chapters, I will show that such an approach indeed pays-off: in addition to addressing the specific research questions outlined in the next section, I will show how my investigation lead to a discovery of a striking example of PR influence in the form of a corporation "manufacturing" statements, getting elected officials to repeat them, and the media reporting them as a regular news story.

**The Two Goals of This Study**

This study has two distinct goals. One is to address the concerns discussed in earlier chapters, as well as the shortcomings identified in previous research, thus, contributing to a broader understanding of the nature and scope of news media's use of the press release. The other goal is to explore the possibilities offered by applying computational methods to the problem of press release content in news media, and to demonstrate the utility of this approach for journalism and mass communication research in general.

The first goal will be addressed through formulating a set of research questions which will cover specific issues highlighted in previous chapters and will be investigated through the application of a combination of quantitative and qualitative methodology.

The second goal will be addressed in a different manner. Rather than formulating a "pilot" research question of the type "can a system be built to handle a given problem 'X'" which begs the answer "yes it can," I will describe the process of selecting, designing, implementing, and running the main computational components I use for my solution. I will explain each step in reasonable detail and, in the end, will evaluate its

usefulness and applicability to similar problems in journalism and mass communication research.

The main practical outcome of applying computation in this study is the generation of a relevance sample – a concept described in the previous section – which is instrumental in addressing the specific research questions outlined at the end of this chapter. Nevertheless, the ultimate goal of using computation in this dissertation is to introduce my colleagues to computational approaches to problem solving in journalism and mass communication research.

**Issues to Investigate**

The concerns voiced by journalism and mass communication scholars over press releases being used as news sources by the media have been shown to be plentiful, have been indirectly confirmed by public relations literature, and have been supported by theoretical argument. These issues can be summarized in the following three groups:

1. The use of press release content with little or no change, with little or no contextual information added. Scholars give many examples of such practice (e.g., Bagdikian, 1974; Lewis et al., 2008a, 2008b; Martin & Singletary, 1981; Smith, 1977; Sullivan, 2011) and agree that in most cases this results in self-serving statements published on behalf of the organization (Jones, 1975; Smith, 1977).
2. When press release content is used, the resulting news report will end up having traits typical of a press release: its language will favor the sponsoring organization, there will be a strong prevalence of positive content over negative

content (Bagdikian, 1974; "A press release and a news story," 1975; Sullivan, 2011), although some researchers have observed that the promotional, overly evaluative language of press releases is neutralized or transformed by the journalist (Maat, 2007, 2008; Maat & de Jong, 2012);

3. Finally, scholars have observed that news articles using press release content make no attribution to the press release as a source (Ambrosio, 1980; Bagdikian, 1974; Jones, 1975; Smith, 1977), which is most troublesome, for "the audience deserves to know if what they are reading/seeing is a handout from a commercial or a government source" (Potter, 2004, p. 68). In fact, this problem is even more serious: Catenaccio (2008)found that a modern press release is so similar to a news article that "a 'perfect' press release may well be interpreted as a genuine news story, and its evaluative components as the result of external judgment" (p. 27).

A comprehensive review of previous research on press releases has revealed several critical shortcomings which have been discussed in detail in the first section of this chapter. Taking these shortcomings into consideration, I address the summarized issues by conducting a study using an adequate sample, representative of US news media and public relations sources alike; computing verbatim text matches, thus reducing the initial data set to a relevance sample containing the most critical data, consisting of pairs of press releases and news articles which exhibit a direct relationship with each other; and investigating this data with regard to language, content, and proper attribution of news to public relations sources.

The primary limitation of my approach is reliance on verbatim text matches, which would have prevented me from measuring the degree of press release influence on news content. However, that would have been a much broader research question which should be addressed in a separate study. The goal of this preliminary study, on the other hand, is to demonstrate how a very large initial data set can be put within the grasp of a qualitative mass communications researcher – thus, exploring the possibilities offered by applying computational methods to mass communication and journalism research, while contributing to a broader understanding of the problem of press release content in news media.

## Research Questions

All research questions are based on the constructed relevance sample consisting of pairs of press releases and news articles, which will be identified in the data set construction phase of the study.

*RQ1.* Given a press release, which is used as a source for a news article, what is the proportion of the press release text used without any change? In other words, how much of the press release text is used verbatim?

*RQ2.* Given a news article, which uses a press release as a source, what is the proportion of the article's text not copied without any change from the press release? In other words, how much of the article's text is not copied verbatim from the press release?

*RQ3.* How does press release content compare to the content of a news article, which uses that press release as a source, in regards to the evaluative, or subjective, language they use? I hypothesize that a news article will use less evaluative, or subjective, language compared to the press release it uses as a source.

*RQ4.* How does press release content compare to the content of a news article, which uses that press release as a source, in regards to the polarity (positive versus negative) of the evaluative, or subjective, language they use? I hypothesize that the language of the news article will be less positive compared to the language of the press release it uses as a source.

*RQ5.* Do news articles, which use press releases as a source, provide attribution to the press release? In other words, do such articles mention the press release as the source of their content?

# CHAPTER 5.
# RESEARCH DESIGN

Due to the exploratory, interdisciplinary nature of this study, I provide a more detailed description of methodology compared to a typical quantitative or qualitative study in journalism or mass communication. The purpose of this detailed description is to demonstrate how methods from computer science and mass communication research can be successfully combined to tackle a specific problem in journalism studies – an approach not typical for journalism or mass communication scholarship.

The overall methodology is grounded in two fields: mass communication and computer science. Initial data collection is handled through a combination of manual and automated procedures. Subsequent data processing, including the construction of the final data set of matching press releases and news articles is mostly automated, with minimal manual supervision, employed at the final stage to filter out irrelevant data not detected by the automated procedure. Data analysis is handled through a combination of quantitative analysis, typical in mass communication research, and a computational approach, often used in the initial stages of subjectivity analysis – an area in computational linguistics dealing with the "recognition of opinion-oriented language in order to distinguish it from objective language" (Pang & Lee, 2008, p. 5).

This chapter will cover all the main steps that constitute the method I developed for this study:

- sampling strategy for building a text corpus consisting of press releases issued by a number of corporations ($N = 40$), and news articles covering the same corporations;

- data collection procedures, including processing and storage of the collected press releases ($N$ = 6,171) and news articles ($N$ = 48,664);
- data analysis procedures employed for addressing specific research questions outlined in the previous chapter.

Figure 2 illustrates these steps. The step involving finding dependencies between news articles and press releases – i.e., locating verbatim matches between texts – will be described in detail in the next chapter.

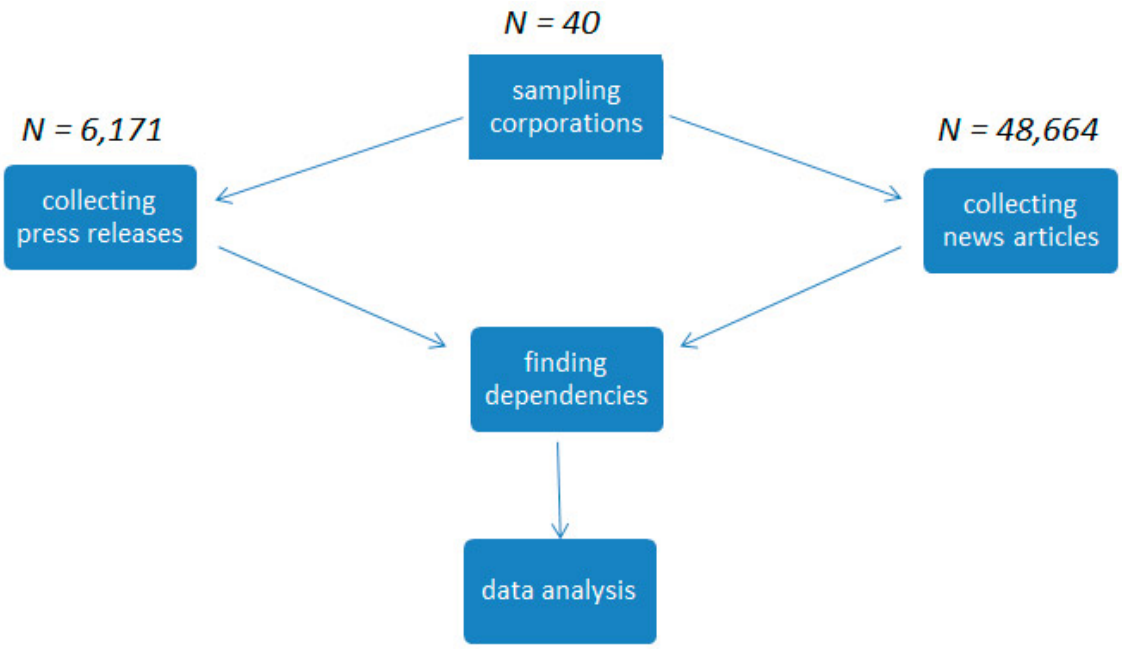



*Figure 2*. Method summary.

## Sampling

Since this study deals with investigating the relationship between press release content and news media content, the data represents two entities: organizations producing press releases on the one hand, and news media which may be using the content of these press releases in their articles on the other.

### Justification of Sampling Strategy

A review of previous research has identified a number of overall methodological shortcomings, which have been common to most studies. Among them, several sampling issues have been identified: (a) inadequate sample size; (b) non-representative sample; and (c) lack of direct relationship between the press releases and the news articles in the sample, when such a relationship is implied or explicitly stated. In this section, I describe my sampling strategy, which addresses each of these issues. I start with discussing the issues of sample size and representativeness, which come into play when building the initial text corpus, which is the very first stage of the data preparation process.

**Adequate sample size and representativeness.** Although there is no formal minimum required sample size for making statistical inferences (the *t* test was initially demonstrated on a sample of size four (Student, 1908)), considerations about having sufficient power when comparing groups – which is the case in this study – suggest a sample of at least $N = 30$ (e.g., J. Cohen, 1990, pp. 1-2). Certainly, many of the studies mentioned in this dissertation had sample sizes exceeding this requirement: most studies used between 50 and 200 press releases ($Mdn = 154$), with outliers including Donsbach et al. (2005) who had a sample of $N = 1{,}015$, and Kiousis et al. (2007) with $N = 1{,}211$.

Reported sample sizes for news articles were even more impressive and ranged from 62 to 6,699 (*Mdn* = 654). Thus, sample size alone does not seem to have been a problem in previous research.

However, the requirement of adequate sample size goes hand-in-hand with the need of having that sample be representative of the population. That is where previous research, arguably, runs into a problem. When we discuss the relationship between news media and public relations sources in general – not a particular newspaper or a type of publication; not a given corporation, university, or political party – we need a sample reflecting the population about which we intend to draw conclusions. Yet, most of the reviewed studies have failed to draw a sample representing the population they analyzed: their samples of news media were typically limited to a few newspapers (e.g., Anderson, 2001; Hong, 2008), a particular type of newspapers (e.g., Martin & Singletary, 1981; Morton & Warren, 1992), and their samples of press releases were limited to a particular industry (e.g., Maat, 2007), a particular type of press releases (e.g., Maat & de Jong, 2012), or one corporation (e.g., Gilpin, 2007). Needless to say, the way the media treat press releases provided by Wal-Mart (Gilpin, 2007) is not necessarily the same as the way the media deal with press releases from Apple (Van Hout et al., 2011); and the way the *New York Times* uses press releases (Hong, 2008) will, most likely, differ significantly from the way they are treated by small local papers (Hale, 1978; Morton & Warren, 1992).

Therefore, if our study were on elite media, we may have drawn our sample exclusively from the *New York Times* – for it is commonly accepted that the *New York Times* represents elite media (Weiss, 1974, as cited in Cooley & Besova, 2009).

However, to draw conclusions about news media *in general*, we need to sample from a variety of publications. If we use the number 30 as the minimum sample size, then we need to draw our sample from at least 30 *different* publications, with, ideally, at least 30 articles from each publication – which results in a sample size of at least 900 articles (30 x 30 = 900).

**Sample representativeness in the context of agenda setting.** An argument for a larger sample can be also made in connection with agenda-setting theory, discussed in Chapter Two and used as a part of the theoretical grounding of this study.

It has been argued than information subsidies, such as press releases, play a role in setting the public agenda. The whole premise about the ability of the organization providing the press release to affect what the public thinks (or what the public thinks about) is what justifies this study. However, a typical study of agenda-setting effects does not deal with a broad media sample. Instead, such studies have employed elite media, such as the *New York Times*, to represent the media agenda. Various reasons have been cited for this choice of media: some mention their availability, high circulation and assumed effect on public opinion (Goya-Martinez, 2009; Hou & Ma, 2009); others explain it by "the important position" such media occupy in the media landscape (Weiss, 1974, as cited in Cooley & Besova, 2009). Chomsky (1997) explained that the audience of elite media is mostly privileged people: "The people who read the *New York Times* – people who are wealthy or part of what is sometimes called the political class [who are] involved in organizing the way people think and look at things" (p. 1). Therefore, following Chomsky's argument about elite media "set[ting] the framework in which everyone else operates, [where] what the *New York Times* tells us is what you're supposed

to care about tomorrow" (p. 1), I might have suggested that considering the context of agenda-setting theory, a sample of elite media, such as the *New York Times*, might have been sufficient in this study.

However, today the media landscape has changed dramatically. The wide availability of newspapers like the *New York Times* due to the size of their circulation is no longer a determining factor in what the public has access to in terms of news sources. According to a Pew study, the Internet is "the third most-popular news platform, behind local and national television news and ahead of national print newspapers, local print newspapers and radio" (Pew Internet, 2010). In addition, the study reports that 65% of online users do not have a favorite news site, and visit multiple news sites daily. The study also found that 56% of online users get their news from portal websites like Google News (Pew Internet, 2010). These numbers become especially significant when considered together with the *number of news sources* available through such portals. A study conducted by scientists at Cornell University reported that Google News indexes more than 20,000 news sources, excluding blogs (Leskovec et al., 2009). This dramatic change in news media availability has led scholars to question the role of elite media in today's media landscape and agenda-setting effects overall (e.g., Althaus & Tewksbury, 2002; Meraz, 2009; Shehata & Strömbäck, 2013; Takeshita, 2006). And although this study is not an exploration of agenda-setting; it seems to be reasonable – for the sake of staying consistent with the theoretical arguments made in Chapter 2 – to investigate the relationship between news media and press release content by drawing a sample of news media as large and as diverse as possible.

## Sampling Strategy for Public Relations Sources

**Limiting the data to business.** In this study, organizations producing press releases are represented by a sample drawn from the top 100 corporations on the Fortune-500 list (CNN Money, 2013) – "an annual list compiled and published by Fortune magazine that ranks the top 500 U.S. closely held and public corporations as ranked by their gross revenue" (Wikipedia, 2013).

Limiting the data to corporations, certainly, limits the scope of conclusions I can draw from such data. With such data, I cannot discuss the relationship between news media and public relations sources in general – i.e., educational institutions, government, political parties, nonprofits – these are examples of entities which fall outside the scope of this study. Without any doubt, the problem of press release content finding its way into news coverage is no less important in regards to a university than it is in regards to a corporation. However, as I have mentioned in Chapter One, public relations is a broad field and its different components – such as public affairs (aka government public relations), political public relations, or corporate communications – constitute a field too broad and diverse to be examined in sufficient depth in the context of one study. It is not unreasonable to expect that government press releases will be printed verbatim much more often than press releases from a nonprofit; and that a press release from a public university will be treated differently than a press release from a corporation. If a combined data set containing press releases and news coverage dealing with such different types of entities were to be analyzed, the *differences* between these types of entities would have been the logical focus of such a study – which is not the focus of this

dissertation. Therefore, to investigate the central problem of this dissertation in sufficient depth, I limit the scope of my study to business as the source of press releases.

**Limiting the data to the largest corporations.** The other obvious limitation of this approach is that the sample is skewed towards the largest companies, which are said to include "some of the most powerful and influential companies in the world today, [which] wield enormous power and influence government policy on a regular basis" (US Pages, 2013). Furthermore, as DiStaso (2012) noted in regard to using a similar list (Fortune's America's Most Admired companies), "one may expect that simply because of their size and popularity, these companies may be considered more newsworthy than other companies; therefore, their intermedia agenda-setting effects may be more predominant" (p. 130).

However, this limitation also serves as a considerable advantage. My task is to construct a data set consisting of press releases and news articles which have a direct relationship with each other. Building a text corpus consisting of press releases issued by a number of organizations and news articles which cover these organizations (or topics similar to those covered in the selected press releases) is the first step in constructing such a data set: as I have explained in the previous chapter, similarity of topics or the usage of the same search terms and phrases in querying a database *does not* guarantee, or even imply, that the press releases were used as a source for the retrieved news articles. Nevertheless, the absolute majority of previous studies which have implied or explicitly stated such a relationship, employed such a text corpus for their analysis. I intend to reduce such a corpus to a data set of press releases and news articles which, indeed, are related to one another. It is reasonable to expect that such relationships will be evident

only among a small subset of press releases and articles in the text corpus. Thus, by selecting the largest and, arguably, some of the most newsworthy corporations, which, therefore, are covered by news media more frequently, I avoid the need to increase my sample of corporations, which would have required more time and computational resources, making this project impossible to complete within a reasonable timeframe.

Furthermore, by selecting a sufficient number of press releases and articles per corporation, I statistically control for the individual differences in that content, thus improving the reliability of my method. For example, if I were to sample from small companies, with some of them being covered in one news article only, whatever conclusions I might have made based on that data would not be generalizable beyond that one company announcement which led to one article – i.e., the data would have been useful for a case study of that specific event covered by the media. However, with a larger company, which offers multiple press releases and which is actively covered by the media, the effects of an individual event, announced by a press release and covered by the news, will be controlled for by selecting a large enough sample of such events, as randomly as possible (including, of course, an exhaustive sample – i.e., a census of the entire population, which will be shown to be possible with selecting press releases).

Therefore, by focusing on large corporations, I am able to collect an adequate sample, which allows me to draw reliable generalizations about the issues under investigation at the cost of the sample being not representative of business in general, but, instead, being limited to its biggest representatives, which I consider to be a reasonable tradeoff.

**Selecting a Sample of Corporations.** I used a systematic sampling approach to draw a sample of 40 corporations ($N = 40$) from the top 100 corporations on the Fortune-500 list. My data sample consists of *all* the press releases issued by these corporations during 2012 (1/1/2012 – 12/31/2012) and available on their respective websites. Following is a detailed description of the steps I followed in drawing the sample.

***Initial selection.*** I started with selecting the top 100 corporations on the Fortune-500 list. Appendix B contains a list of URLs of corporate websites containing content for news media, including, in most cases, press releases (Table B1).

***Availability of a press release archive.*** I examined each corporate website trying to locate an archive of press releases. I was able to locate press release archives for all companies but for one: Google. Google offers a "News from Google" section on its corporate site (Google, 2013a), which does contain traditional press releases – i.e., the news are posted by different staff members, they allow commenting, and they don't follow any obvious formal pattern typical of a press release (and discussed in Chapters One and Three). Google also provides a traditional press release archive (Google, 2013b); however, the press releases are part of a section on investor relations and, therefore, are assumed to contain specialized information relevant to this community only.

The specialized nature of this type of press release is not enough to exclude a company from the sample; for example, Costco (Costco, 2013) was left in the sample despite providing press releases as part of it investor relations website. However, in the case of Costco, the press releases listed on the investor relations site were *the only* news formally offered by Costco as a source for media. There were, certainly, numerous examples of companies offering on their web sites a variety of information designated as

news; however, in all those cases, the companies also provided an archive of formal press releases, not limited to one subject area, like investor relations. This was clearly not the case with Google, where formal news were presented in a much less formal way; thus, while examining such content certainly is a worthwhile direction for research, it would have considerably skewed the data set in the context of this particular study which aims to examine formal press releases.

***Availability of press releases for 2012.*** My goal was to retrieve all press releases for a specified period: the year 2012. For the sake of consistency of sampling, I excluded eight companies which did not provide access to their press releases for the entire period. Those companies included Best Buy, Chevron, Coca-Cola, Humana, Intel, Oracle, Verizon Communications, and Walgreens.

***Sufficient number of press releases***. As I have previously mentioned, I required companies which provided a sufficient number of press releases – to be able to control for the effects of individual events. Therefore, I excluded another eight companies which offered less than 30 press releases. Those companies included AmerisourceBergen, Berkshire Hathaway, Murphy Oil, Sysco, TIAA-CREF, Tyson Foods, Valero Energy, and World Fuel Services.

***News media sampling issues.*** I further reduced the sample by removing six companies due to a variety of technical problems encountered while sampling corresponding news articles from LexisNexis and described later in this chapter. Those companies included AT&T, DuPont, Freddie Mac, General Motors, Medco Health Solutions, and Nationwide. I also removed six companies for which I was able to retrieve less than 100 potentially relevant articles using LexisNexis, which I determined to be an

insufficient sample with regard to my goals and comparing to the amount of coverage of other companies remaining on the list. Those companies included INTL FCStone, CHS, Enterprise Products Partners, Honeywell International, Ingram Micro, and Plains All American Pipeline.

***Final list.*** The remaining list consisted of 71 companies. To get an even distribution of companies with regard to their Fortune-100 rank, I used a systematic sampling approach. I ordered the list of companies by their rank (which represents their revenue amount) and picked every other company (starting with the first on the list) ; I then randomly selected another four companies, achieving a sample size of $N = 40$. This final sample is listed together with corresponding industry types (as identified by Fortune) in Table 2. Table A1 in Appendix A provides the list of all Fortune-100 corporations together with the reasons for each company's removal from the sample.

Table 2. *Sample of corporations.*

| Rank | Company | Industry |
|---|---|---|
| 1 | Exxon Mobil | Petroleum Refining |
| 8 | Fannie Mae | Diversified Financials |
| 9 | Ford Motor | Motor Vehicles and Parts |
| 13 | Bank of America Corp. | Commercial Banks |
| 14 | McKesson | Wholesalers: Health Care |
| 16 | J.P. Morgan Chase & Co. | Commercial Banks |
| 17 | Apple | Computers, Office Equipment |
| 18 | CVS Caremark | Food and Drug Stores |
| 19 | International Business Machines | Information Technology Services |
| 20 | Citigroup | Commercial Banks |
| 23 | Kroger | Food and Drug Stores |
| 26 | Wells Fargo | Commercial Banks |
| 28 | Archer Daniels Midland | Food Production |
| 34 | MetLife | Insurance: Life, Health (stock) |
| 35 | Home Depot | Specialty Retailers: Other |
| 37 | Microsoft | Computer Software |
| 38 | Target | General Merchandisers |
| 39 | Boeing | Aerospace and Defense |
| 41 | PepsiCo | Food Consumer Products |
| 43 | State Farm Insurance Cos. | Insurance: Property and Casualty (mutual) |
| 45 | WellPoint | Health Care: Insurance and Managed Care |
| 49 | Comcast | Telecommunications |
| 56 | Amazon.com | Internet Services and Retailing |
| 57 | Merck | Pharmaceuticals |
| 58 | Lockheed Martin | Aerospace and Defense |
| 61 | Sunoco | Petroleum Refining |
| 63 | Safeway | Food and Drug Stores |
| 67 | Johnson Controls | Motor Vehicles and Parts |
| 68 | Morgan Stanley | Commercial Banks |
| 70 | FedEx | Mail, Package, and Freight Delivery |
| 71 | Abbott Laboratories | Pharmaceuticals |
| 76 | United Continental Holdings | Airlines |
| 83 | Delta Air Lines | Airlines |
| 84 | Liberty Mutual Insurance Group | Insurance: Property and Casualty (stock) |
| 86 | New York Life Insurance | Insurance: Life, Health (mutual) |
| 89 | Aetna | Health Care: Insurance and Managed Care |
| 90 | Sprint Nextel | Telecommunications |
| 93 | Allstate | Insurance: Property and Casualty (stock) |
| 95 | American Express | Commercial Banks |
| 97 | Deere | Construction and Farm Machinery |

## Sampling Strategy for News Media

**Limiting the data to US newspapers.** Like with public relations sources, this study does not try to encompass all news media, for that would have been too broad for a sufficiently deep and meaningful analysis. Considering the focus of this study's methodology on dealing with textual data, it is only natural to limit the media sample to print media. Certainly, a focus on news in the form of text might suggest different types of such data, including magazines, trade publications and, most importantly, the vast selection of online news media which may include anything from a news aggregator, like *Google News*, to a blog, like *Huffington Post*. However, for the sake of keeping this study within manageable limits, the scope of news media will be limited to newspapers only. Since one of the shortcomings of past research was identified in Chapters Three and Four as the lack of studies of press release content with regard to its usage by US media, the scope of the news media sample will be further limited to US newspapers.

**Limiting the data to news sources indexed by LexisNexis.** The scope of the data is further limited to US newspapers indexed by the Academic version of LexisNexis. Although LexisNexis Academic provides a list of 298 indexed news sources identified as newspapers, it is impossible to tell the exact number: 24 of the 298 sources are identified as gateway and aggregate sources; further examination of the components of these sources show listings of news sources which are not limited to newspapers and, most importantly, are not limited to the United States. We do not know how LexisNexis implements its information retrieval algorithms, so we cannot tell what proportion of US newspapers is searchable through LexisNexis. Considering that, according to USA Today (2008), as of May 23, 2008, there were 1,422 dailies and 6,253 weeklies in the US, with

the full list available on Wikipedia (Wikipedia, 2013b), and at least some of them, such as the *Cincinnati Enquirer*, it is reasonable to assume that by using LexisNexis Academic we are imposing a considerable limitation on the scope of our data. Nevertheless, to my knowledge, this was the best source available to me: Spinn3r, a media monitoring resource used by Leskovec et al. (2009) indexes only 600 mainstream news sites, which, clearly, doesn't cover the scope of US newspapers; Factiva, an alternative to LexisNexis, may have more sources indexed, but its academic version – which is the only version available to me – is quite limited.

However, most importantly, regardless of such limitations, I consider LexisNexis Academic to be good enough as a database of news media for the purposes of my study.

**Identifying relevant articles.** In Chapter 3, I reviewed several approaches to sampling news media employed in previous research where the goal was to compare press release content to news media content with some common criteria – be it a similar topic or an assumed direct relationship. However, regardless of the specific procedure, the overall goal is to locate content which is relevant to the same topics covered in the press release, or covers the same organization which issued the press release.

With a small sample size this task is relatively trivial: one may simply use the company name as a search term on LexisNexis Academic or a similar resource, and manually filter out the irrelevant results. If a larger sample is required, things get difficult very soon. One obvious problem lies in the ambiguity of the name of a company: searching for coverage of a company with a name like Kroger will result in less false positives than searching for coverage of Apple. A less obvious, but – as preliminary testing has shown – much more significant problem is the relevancy of the coverage: an

article can be primarily about the company (i.e., which is relevant), or it could be on a subject completely irrelevant in the context of the study, yet mention the company's name in regards to one of its products or services (e.g., FedEx, Ford, Google, etc.). Filtering out such irrelevant articles manually is possible only with a relatively small sample size.

Certainly, a search term can be augmented with more specific criteria describing a particular press release. However, if the sample of press releases is not limited to a few, but instead, like in this study, contains several thousand, such an approach is, clearly, not feasible.

There are ways to automatically determine whether a text is relevant to a particular topic, such as building statistical models of topics based on language features. In this case, a very large sample of media coverage could have been automatically processed, selecting the articles which had the highest probability of relevancy. Yet another approach is to use crowdsourcing – i.e., distribute the text classification task to a very large number of human coders (e.g. Amazon Mechanical Turk, 2013), in which case filtering out irrelevant articles would be possible for a relatively large data set. However, both of these approaches, to the best of my knowledge, have not been applied in journalism or mass communication research. Therefore, using them would require at least one pilot study demonstrating the applicability of the method to mass communication research.

**SmartIndexing by LexisNexis.** Luckily, LexisNexis Academic provides its own system of indexing documents with a set of tags, including names of companies. Although we do not have access to a detailed description of this proprietary technology, judging by the summary offered by LexisNexis, it is a combination of human and

automated indexing., which seems to be solve, at least in part, the problem of identifying relevant articles:

> SmartIndexing Technology … applies controlled vocabulary terms for several different taxonomies to all LexisNexis news ... The taxonomies are built and maintained by the LexisNexis Taxonomies & Indexing team, comprised of information professionals, lawyers, subject matter experts and analysts. Index terms are assigned using a unique approach that combines the best features of human and automated indexing practices. Index term rules are developed and tested by the Taxonomies & Indexing team. These rules "read" incoming documents and assign relevant index terms automatically to documents and sources (LexisNexis, 2013).

After evaluating the accuracy of this indexing using several company names from my sample, I concluded that, despite certain shortcomings, the technology was good enough for the purposes of this study.

## Building a Text Corpus

This study addresses two distinct goals: one deals with answering specific research questions outlined at the end of Chapter Four, whereas the other goal is to demonstrate the application of computer science to constructing the data set which is required to answer those research questions. Thus, my data set construction process consists of two steps: (a) collecting the initial data consisting of press releases and news article, and (b) discovering relationships between items in the collected data and, through that, building a data set consisting of pairs of items which exhibit such relationships. The second step, being one of the goals of this study, is described in detail in the next chapter, which discusses the results of the study. For the sake of clarity, I will refer to the data collected in the first step as a text corpus, a term used to describe "a large and structured

set of texts" (Wikipedia, 2013f). I will refer to the data identified in the second step as the final data set.

The process of building the text corpus used in this study consists of two parts: collecting the actual texts, and preparing those texts for further analysis. With a smaller sample which is analyzed manually the data preparation step is optional: it doesn't really matter what format the data is in as long as it is human-readable. However, in this study I use computational methods for several tasks, for which transforming the collected texts into machine-readable form is a prerequisite. Therefore, beside data collection, I describe data processing (or preparation for analysis) as a separate stage.

**Collecting News Articles**

I used LexisNexis Academic to collect a sample of $N = 48{,}664$ news articles related to the 40 corporations in the sample of public relations sources. This section describes the details of the search procedure I followed.

My goal was to collect as many relevant news articles per company as possible. I developed a search procedure by trial and error, trying to maximize the amount of coverage per company while staying within manageable limits and keeping the data reasonably relevant. The procedure was systematic, yet it varied depending on the amount of coverage available for each company: for example, if there were more than 4,000 search results for the year (e.g., *Bank of America*, *Fannie Mae*, *Apple*), I used additional criteria to narrow down the search. As a result, the data I collected cannot be used to compare companies with regard to the *amount* of news coverage.

Following is a detailed description of the procedure I used for collecting news articles relevant to each of the 40 corporations I had selected for my sample of public relations sources. I provide these detailed steps, which I worked out through trial and error, as a guide for researchers who may wish to replicate my results.

**Search procedure.** I click "News", then "Newspapers & Wires"; I then select two sources: "US Newspapers" & "Wires and Small Town Papers (US)." In Advanced Search, I add index terms to my search query, which represent the company name. LexisNexis offers a list of terms to choose from, which often include alternative company name spellings (e.g., *Apple* versus *Apple Inc.* versus *Apple INC*), company names from different periods (*Apple* versus *Apple Computer*) , and subsidiaries – both local and international (*Apple Sales International*, *Apple Canada Inc.*). If there were more than 20 such indexing terms for a company (which was often the case), I selected the primary terms only (i.e., excluding subsidiaries).

Narrowing down the results in the case when a company had more than 4,000 search results for the year was accomplished by marking a "relevancy" checkbox which ensured that only articles where the company name appeared among the major terms would be returned (after some experimentation, I found that a term is considered to be a major term if its relevancy is greater than 75%, based on a LexisNexis proprietary scoring scale).

Finally, after selecting an appropriate date range, I select the "Newspapers" category to view the search results. The format of the imported results was text, current category only (i.e., newspapers), with the body and byline as the selected fields.

LexisNexis allows importing up to 200 documents at a time; therefore, I repeated the same procedure for each chunk of 200 search results.

In the process of sampling I encountered a few problems, which were caused, most likely, by indexing errors: *General Motors* had no categories available at the time of sampling (the issue has been resolved), using *AT&T* as an index term resulted in no results at all when the relevancy checkbox was checked, or in more than 3,000 results for any time period with the checkbox unchecked. However, such errors were rare and resulted in the exclusion of only a few companies from my initial selection.

**Preparing news articles for analysis.** The imported articles are initially stored as text files, each containing up to 200 articles. Prior to any analysis, the text in these files must be turned into machine-readable data – i.e., text data which can be processed by a computer program. For this, I wrote a program which read in the text from the files, identified all the relevant structural elements of each article – such as date, name of newspaper, title, byline, and body – and stored them in structured form, ready for further analysis. (Data was stored in text files and a database; implementation details are available in Appendix D.)

## Collecting Press Releases

A large sample of press releases can be obtained through LexisNexis Academic which provides access to press release distribution services like BusinessWire and Newswire. However, my goal was not to collect as many press releases as possible; instead, I needed press releases issued by *specific* companies. Considering this requirement, a better way to collect such data is to obtain such press releases directly

from these companies – since most companies provide unrestricted access to their press release archives on their websites. As I have noted earlier in this chapter, only the companies which provide such access to the press releases were included in the sample.

An examination of the press release archives provided by the companies in the data sample demonstrated that there were more than 6,000 press releases available. One possible approach would have been to randomly sample from each company. Assuming I had maintained the minimum sample requirement per company to statistically control for differences between individual press releases, this approach would have resulted in a data set of 1,200 press releases (30 press releases per company) – which is still a very large number to be collected manually. Consider the following: we need to record the company, the date, the title, and the body of each press release in some structured form; and we do that by accessing each press release through our Internet browser – by clicking a link, copying and pasting the text of each field (date, title, body) into a text file, formatting each entry – 1,200 times. At the rate of one press release per minute that would take 20 hours of repetitive, highly error-prone manual data-entry.

**Web page scraping.** Instead, I opted for an automated approach, known in the computer science community as web page scraping. When a large amount of data needs to be collected from a website, or websites, with the target web pages displaying some sort of pattern in the way they present the data we are after, rather than manually copying the data, it is much more efficient to write a program which will access each page automatically and "scrape" the page extracting the relevant data and store it in a structured format for further use. Figure 3 demonstrates the repetitive elements in data

presentation on several web pages, which can be recognized by a program and, therefore, processed automatically.


*Figure 3*. Repetitive elements on web pages.

The data is extracted from the source code of the page. Any web page is nothing but a file containing text; programs can be written to retrieve such a file and "read" its contents, saving the relevant data. If the relevant data appears in the same form in multiple files, we can "instruct" out program to recognize, extract and save this data. For example, if we wanted to extract the title of a web page (i.e., the text appearing at the top of the browser window), we would simply "instruct" our program to extract the text

appearing between these two elements of the web page source code: <*title*> and <*/title*>. Thus, by giving our program a list of such pages, we can extract many titles in seconds.

My program took a list of URLs as input, which were addresses of lists of press releases for each company, accessed each page, extracted the relevant links to press releases, then accessed each press release, extracting its date, title and body, storing all this data in structured form for further analysis. A diagram displaying part of this process is displayed in Figure 4. The top screenshot is that of a web page source code, containing both markup code and the web page text content. The bottom screenshot on the left displays the extracted body of the press release. The bottom screenshot on the right displays a structured list of extracted data: press release code (a unique identifier), date, title, and the URL from which it was extracted. As a result of this process, I was able to collect thousands of press releases within an hour; and, when needed, I repeated the process with a single keystroke. (Implementation details are available in Appendix D.)

Thus, I collected *all* the press releases issued by each company in my data sample for 2012 and made available on their respective websites. The total number of collected press releases was $N = 6{,}171$.

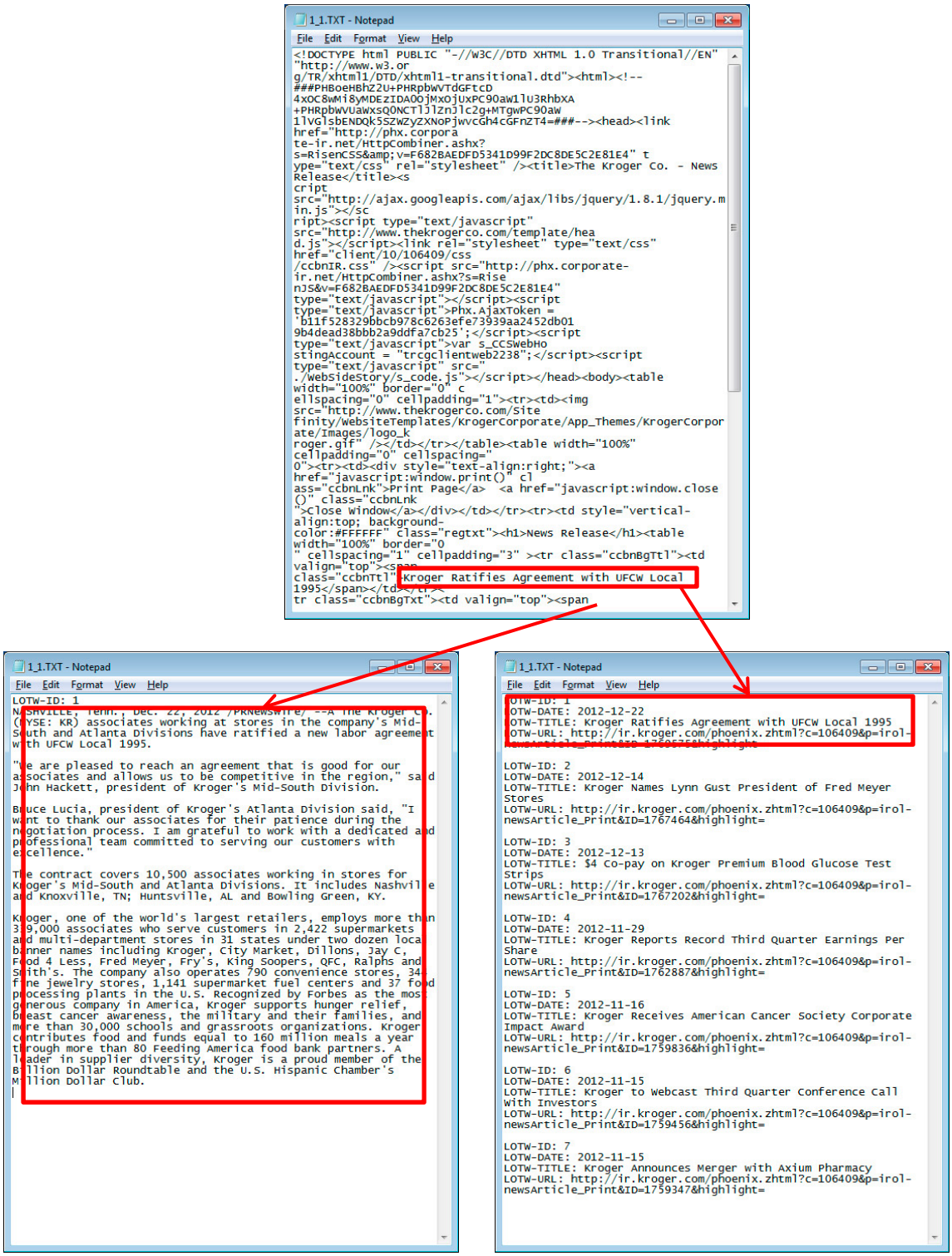



*Figure 4*. Text extraction from web pages.

## The Constructed Text Corpus

The constructed text corpus consisted of 6,171 press releases issued by 40 corporations, and 48,664 news articles with their topics being related to these 40 corporations. Table 3 list the following data for each company: the number of press releases, the number of news articles, and the cross-product of press releases and news articles, which represents the number of possible matching pairs between article and press release. The cross-product, although not used directly, is representative of the number of possible combinations for a matching press release and news article for each company and, as such, is relevant to next chapter's discussion.

Table 3. *Text corpus.*

| Corporation | News Articles | Press Releases | Articles * Releases |
|---|---|---|---|
| Ford Motor | 4,162 | 529 | 2,201,698 |
| Boeing | 3,709 | 359 | 1,331,531 |
| Fannie Mae | 3,036 | 311 | 944,196 |
| Bank of America Corp. | 2,873 | 298 | 856,154 |
| Wells Fargo | 2,707 | 184 | 498,088 |
| Lockheed Martin | 1,071 | 393 | 420,903 |
| Amazon.com | 2,293 | 162 | 371,466 |
| Comcast | 3,937 | 90 | 354,330 |
| Microsoft | 895 | 253 | 226,435 |
| Apple | 3,979 | 55 | 218,845 |
| American Express | 1,702 | 127 | 216,154 |
| Delta Air Lines | 1,518 | 138 | 209,484 |
| PepsiCo | 877 | 197 | 172,769 |
| Sprint Nextel | 488 | 321 | 156,648 |
| Target | 619 | 235 | 145,465 |
| International Business Machines | 388 | 357 | 138,516 |
| Citigroup | 1,821 | 73 | 132,933 |
| FedEx | 1,142 | 114 | 130,188 |
| Aetna | 567 | 212 | 120,204 |
| Allstate | 1,100 | 104 | 114,400 |
| Exxon Mobil | 1,399 | 67 | 93,733 |
| J.P. Morgan Chase & Co. | 551 | 170 | 93,670 |
| Home Depot | 1,954 | 30 | 58,620 |
| United Continental Holdings | 467 | 121 | 56,507 |
| CVS Caremark | 399 | 134 | 53,466 |
| Deere | 722 | 72 | 51,984 |
| Merck | 405 | 108 | 43,740 |
| MetLife | 442 | 97 | 42,874 |
| Abbott Laboratories | 365 | 105 | 38,325 |
| State Farm Insurance Cos. | 777 | 43 | 33,411 |
| WellPoint | 391 | 72 | 28,152 |
| Johnson Controls | 243 | 89 | 21,627 |
| Kroger | 328 | 60 | 19,680 |
| Morgan Stanley | 237 | 61 | 14,457 |

(continued)

| | | | |
|---|---|---|---|
| McKesson | 100 | 137 | 13,700 |
| Safeway | 198 | 64 | 12,672 |
| Archer Daniels Midland | 170 | 74 | 12,580 |
| Sunoco | 283 | 42 | 11,886 |
| New York Life Insurance | 188 | 63 | 11,844 |
| Liberty Mutual Insurance Group | 161 | 50 | 8,050 |
| TOTAL | 48,664 | 6,171 | 9,681,385 |

## Data Analysis

I conduct my data analysis using the final data set which I construct on the basis of the text corpus I described in the previous section. Since the construction of the final data set is one of the two goals of this study, it is one of the anticipated results in its own right and, therefore, is covered in detail in Chapter Six, which discusses the results of this study. In this section, I describe the methods I use to answer specific research questions which were formulated in Chapter 4, with the understanding that the data set required for this analysis has been constructed.

### Selecting the Unit of Analysis

Most of the data analysis in this study, as well as the entire process of discovering relationships between press releases and news articles is determined by one key consideration: the unit of analysis.

In fact, selecting the unit of analysis is one of the key considerations in any kind of data analysis. This step is especially critical in the context of this study due to its extensive use of computational methods, where an explicitly stated unit of analysis is a prerequisite. When designing a computational solution, one starts, essentially, with an algorithm – "a step-by-step procedure for calculations" (Wikipedia, 2013d). The one

thing to keep in mind is that every step needs to be stated explicitly – i.e., a statement like "compare text A to text B" implies that the computer will know what we mean by "text." Thus, the first thing a researcher has to do is determine how texts are to be compared – i.e., one needs to select the smallest piece of data under consideration – i.e., the unit of analysis.

**Characters.** Generally speaking, a text span can be defined in terms of characters, meaningful sets of characters (e.g., words), or sets of sets of characters delimited by punctuation (sentences). We can go further and consider paragraphs, or sections of text consisting of multiple paragraphs, but that will not serve the purposes of this study. Selecting a character as the unit of analysis for a computational approach is a tempting possibility: with this approach, the algorithmic solution is a trivial exercise in string matching (for a description of this type of computational problems, see NIST, 2013, or Wikipedia, 2013c).

However, it turns out that matching characters may be not the best approach. On the one hand we do not care if the text in the press release differs from the text in the article in terms of punctuation, capitalization or white space. We also don't care how many of the commas used in a quote in a press release made their way into the news article or whether they have been replaced by semicolons. On the other hand, we *do* care that only complete words match – i.e., with character comparison, we would have a match of 3 characters between " Betelgeuse" and "username" with "use" being the match. Needless to say, this match is meaningless for our purposes. Thus, using a character as a unit of analysis is, clearly, not an option.

**Sentences.** Our next choice might be sentences. However, a preliminary examination of the data showed that complete sentences are rarely used verbatim; besides, we would want to detect instances when a sentence was used partially, provided the sequence of matching words would have been long enough to be meaningful. A typical example of matching text between a press release and a news article is demonstrated in Table 4: although it is obvious to the reader that the entire quote has been used, our sentence-matching algorithm would pick up only the second sentence.

| Press Release | News Article |
|---|---|
| "During the past two years, we have successfully introduced a more centralized organization to our Upstream; BP's largest organizational change for two decades. I believe it is now timely and appropriate to appoint a fully dedicated chief executive to this, our largest business." | "During the past two years, we have successfully introduced a more centralized organization to our upstream, BP's largest organizational change for two decades**,"** **Mr. Dudley said in a statement. "**I believe it is now timely and appropriate to appoint a fully dedicated chief executive to this, our largest business." |

Table 4. *Partial sentence match; added text is boldfaced.*

**Words versus tokens.** The obvious choice for a unit of analysis appears to be a word. We may choose to ignore punctuation, which will eliminate problems like the ones demonstrated in Table 4. However, words may cause issues as well. Consider matching "New York" and "New Hampshire" – do we really want to match partial names of states? Furthermore, if we were to use the naïve approach of telling the computer that a word is a sequence of characters separated by white space or punctuation, we would run into all

kinds of problems: consider, for example, the following sentence: "Isn't Mr. O'Neil a New Yorker?" Using white space and punctuation as a delimiter wouldn't do us any good: we would end up with "words" like *isn*, *t*, *mr*, *o*, *neil*, *a*, *new*, *yorker* (a total of eight words), which are, most likely, not the words we were hoping to identify: *isn't*, *mr.*, *o'neil*, *a*, *new yorker* (a total of five words).

Instead of dealing with words, we use tokens. A token "is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing" (Manning, Raghavan, & Schütze, 2008). A tokenizer is a program which splits text into tokens (usually we first split a text into sentences with a sentence tokenizer, after which we split each sentence into tokens). This approach transforms our text into a sequence of meaningful tokens which can be then automatically compared with another sequence of tokens with the goal of finding matching sequences of a given length. There are several tokenizers available; besides, one can be built from scratch, although this would be the least efficient approach: like most tasks in natural language processing, tokenization, is not a trivial problem. In my solution, I used the NLTK library (Natural Language Toolkit, 2013). Implementation details are available in Appendix D.

**Computing Proportions of Matching Text**

The first two research questions (RQ1 and RQ2) are addressed by calculating proportions of matching text, which is quite straightforward. The text is measured in tokens.

**RQ1. Given a press release, which is used as a source for a news article, what is the proportion of the press release text used without any change? In other words, how much of the press release text is used verbatim?** To address this question, I compute the proportion of news article content copied verbatim from a press release to the entire content of that press release.

**RQ2. Given a news article, which uses a press release as a source, what is the proportion of the article's text not copied without any change from the press release? In other words, how much of the article's text is not copied verbatim from the press release?** To address this question, I compute the proportion of news article content not copied verbatim from a press release to the entire content of that news article.

### Analysis of Evaluative Language

To analyze the usage of evaluative language and examine whether or not journalists tend to use language which is less positive and more neutral and negative in tone and, overall less evaluative, or subjective, I will use a computational approach, often used in the initial stages of subjectivity analysis. I will verify this approach by comparing its results to results of manually coding a sample of the data.

The concept of individual words being indicative, to some extent, of the tone of the text is not new in journalism and mass communication research (e.g., Maat, 2007; Martin & Singletary, 1981). However, detecting instances of such words automatically is not typical in journalism studies (and, to my best knowledge, has not been attempted so far). However, it is not uncommon in fields, such as computational linguistics;

specifically, it is often used as one of the basic steps in subjectivity analysis, a concept mentioned at the beginning of this chapter.

Thus, to address research questions 3 and 4, I use the following procedure.

**Step 1. Computational approach.** To identify evaluative language, I use a subjectivity lexicon consisting of 8,221 words, or terms, marked with their part of speech tag and measures of subjectivity and polarity. For example, the term *weak* is marked as an adjective with its polarity being negative, whereas the term *best*, occurring as any part of speech, is marked as positive. (For a description of the lexicon and a brief history of its development, see Wilson, Wiebe, & Hoffmann, 2009). I write a program which takes the text of a document (a press release or a news article) as input, runs a sentence and a word tokenizer and a part-of-speech tagger on it, after which it attempts to match each token to a term from the subjectivity lexicon. Successful matches are stored and are associated with the sentence in which they occurred. Figure 5 illustrates how this process works: the program recognizes the verb *help* as an entry in the subjectivity lexicon.

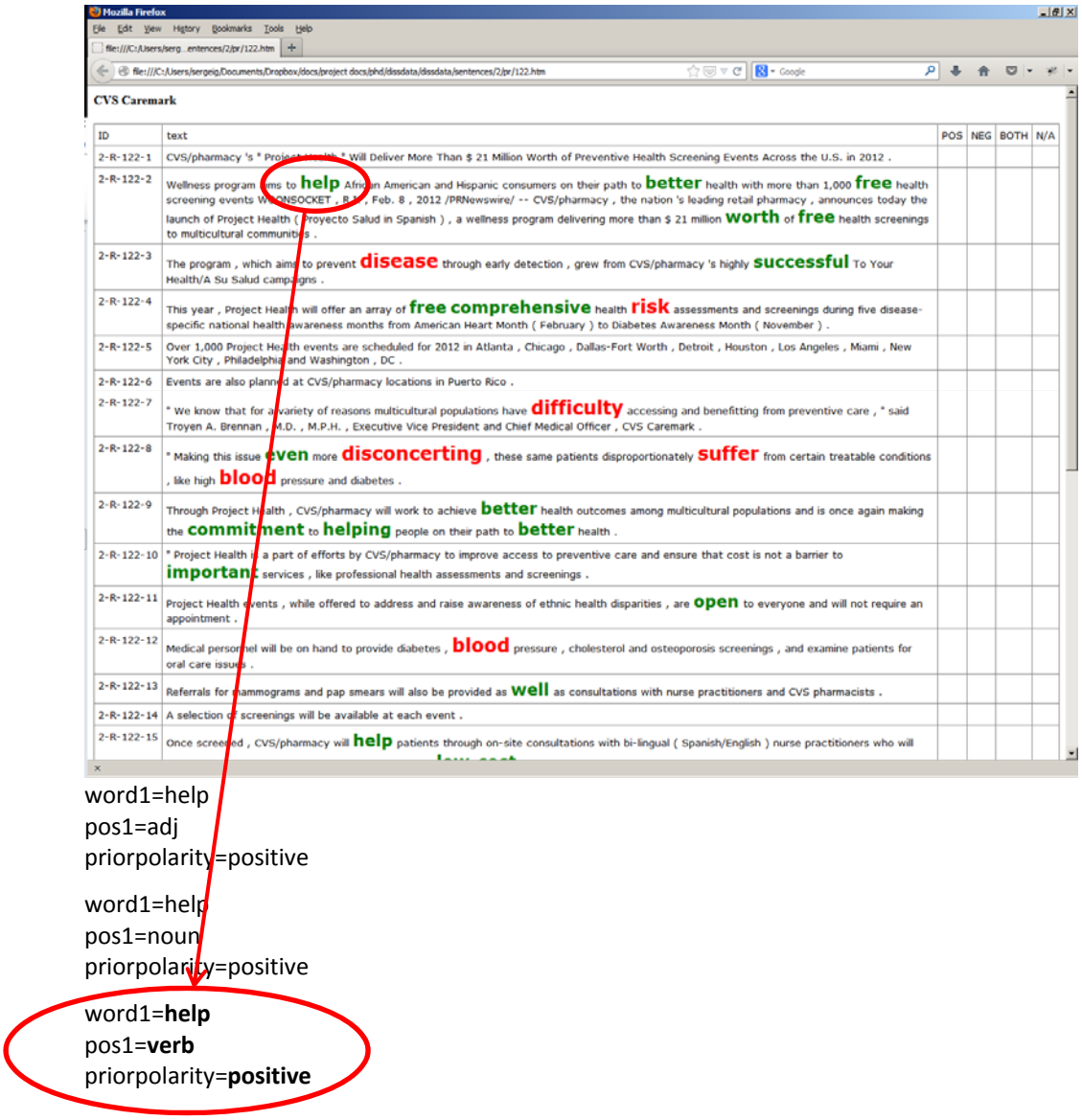



*Figure 5*. Detecting potentially subjective words.

Research has shown that multiple occurrences of a subjective term generally do not increase the evaluative/subjective orientation, or polarity, of the text (Pang, Lee, & Vaithyanathan, 2002), therefore, I mark a sentence as *positive* if it contains at least one positive term and no negative terms; I mark a sentence as *negative* if it contains at least one negative term and no positive terms; I mark a sentence as *both* if it contains both

positive and negative terms; and I mark a sentence as neutral if it contains no terms identified as positive or negative. Figure 6 illustrates how the automated scoring works. A sample result of such automatic coding is presented in Appendix E.

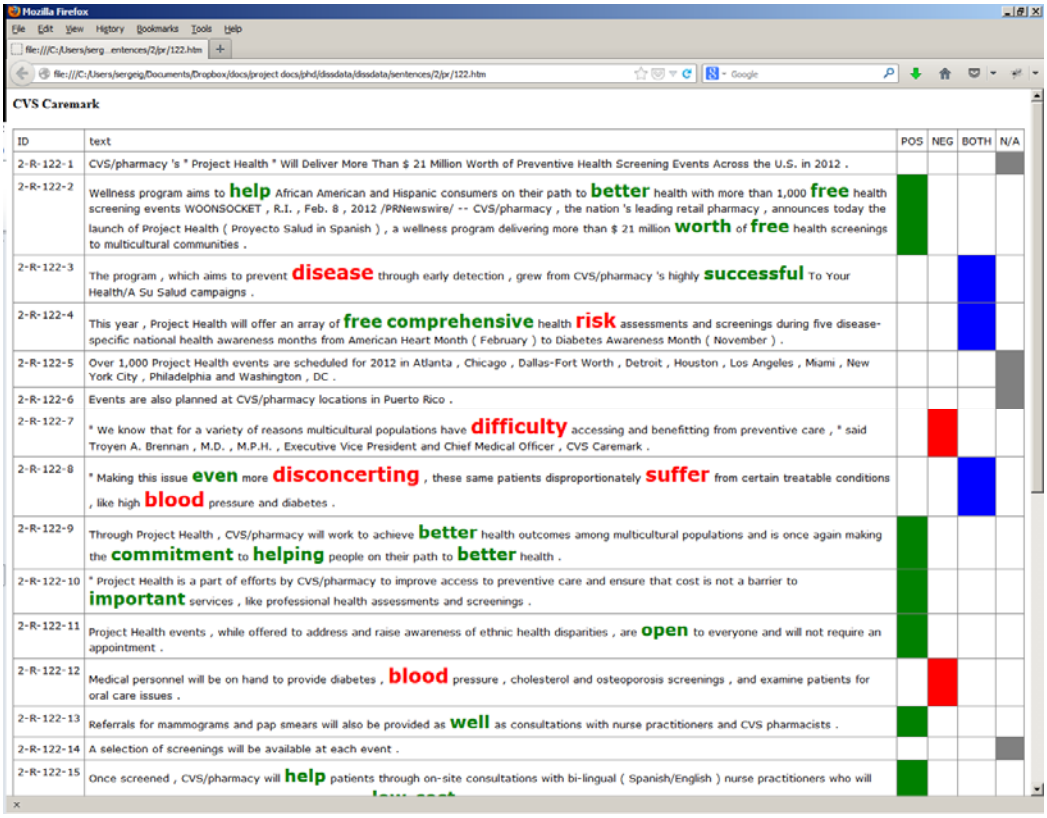



*Figure 6*. Calculating subjectivity and polarity scores.

**Step 2. Manual approach.** Research has shown that the subjectivity and polarity measures of a word does not necessarily indicate the subjectivity and polarity of the sentence. Wilson et al. (2009) referred to this as prior polarity and contextual polarity: i.e., while the prior polarity of the term *unpredictable* may be negative, its contextual polarity in a phrase like *unpredictable movie plot* will be positive – and that's but one example of numerous possibilities. Thus, it may be the case that measuring the usage of evaluative language will not represent the overall tone of the article or the press release – i.e., the occurrence of terms identified as positive in a sentence may not make that sentence positive, or even subjective at all. With this in mind, I conduct a manual analysis of a sample of the data. I select 30 press releases from the entire data set, after which I select one news article for each press release out of the set of news articles which have been shown to use that press release as a source. My selection is only partially random: I select pairs which are of comparable length (i.e., I do not select a lengthy financial earnings report and a short article of the "business briefs" type); I also do not select pairs where most of the article text is based on the press release.

Two coders were trained to code each sentence as being positive, negative, both positive and negative, or neutral with regard to its tone. Thus, a sentence describing a positive or negative fact was treated as neutral; whereas a sentence that was perceived as evaluative, or subjective, was coded accordingly. For example, a sentence containing a phrase like "we are pleased to have built upon our longstanding relationship…" (Microsoft, 2012) would have been coded as positive, whereas a sentence like "The Home Depot closes seven big box stores" (Home Depot, 2012), despite the content which

might be perceived as negative, would have been coded as neutral, since it is an impartial statement of fact.

Intercoder reliability analysis using the Kappa statistic was performed to determine consistency among coders and was found to be Kappa = 0.73 ($p < 0.001$), which demonstrates substantial agreement between coders (Landis & Koch, 1977). (SPSS output is available in Appendix I, Figure I1.)

**Step 3. Comparing approaches.** The results of manual coding are used as the "gold standard". I use the Kappa statistic to determine whether the results of the automated approach are reliable. I also calculate precision, recall and the harmonic mean for each category (positive, negative, both, and neutral).

**RQ3: How does press release content compare to the content of a news article, which uses that press release as a source, in regards to the evaluative, or subjective, language they use?** I calculate the evaluative language score as the proportion of sentences labeled as *positive/negative/both* to all sentences in each document. Scores may range from 0 to 1: when all sentences are coded as neutral, the score will be zero; when all sentences are coded as positive, negative or both, the score will be one; thus, the more subjective sentences – the closer the score will be to one and vice versa.

I use a paired samples *t* test to compare the calculated scores of press releases to news articles to see if there is a statistically significant difference between them with regard to evaluative language. I conduct this procedure for the results of both manual and automated coding.

**RQ4. How does press release content compare to the content of a news article, which uses that press release as a source, in regards to the polarity (positive versus negative) of the evaluative, or subjective, language they use?** I calculate the evaluative language polarity score as the proportion of the difference between the number of sentences labeled as *positive* and sentences labeled as *negative* to all subjective sentences in each document, if there are more than zero subjective sentences; otherwise, I assign a score of zero. Scores may range from -1 to 1: when all subjective sentences are coded as positive, the score will be one; when all subjective sentences are coded as negative, the score will be negative one; when there is an equal number of positive and negative sentences, or when all subjective sentences are coded as both, the score will be zero. Thus, the more positive sentences – the closer the score will be to one, and the more negative sentences – the closer the score will be to negative one, and the more balanced the coding is, the closer the score will be to zero.

I use a paired samples *t* test to compare the calculated scores of press releases to news articles to see if there is a statistically significant difference between them with regard to evaluative language. As with the previous research question, I conduct this procedure for the results of both manual and automated coding.

## Measuring Attribution to the Press Release

**RQ5. Do news articles, which use press releases as a source, provide attribution to the press release? In other words, do such articles mention the press release as the source of their content?** Finally, to examine whether or not journalists make explicit attributions to a press release when they use its content, I searched each

article in the data set for occurrences of the words "press release" and "news release." I tested a 10% random sample of the identified attributions for false positives and found none. However, as an additional step, to test for false negatives – i.e., instances of attribution which I may have missed due to the absence of the words "press release" and "news release" – I manually examined press release/news article pairs representing 50% of the companies, selecting companies 1 through 20, which accounted for 38% of the data ($N = 616$).

# CHAPTER 6.
# RESULTS

This chapter consists of two parts. The first part describes the results of designing a method for discovering relationships between press releases and news articles, and constructing the final data set, or relevance sample – which makes all the subsequent data analysis possible. The second part describes the results of addressing the specific research questions outlined in Chapter Four.

## Discovering Relationships Between Press Releases and News Articles

In previous chapters, I identified the most important shortcoming of past research to be the failure to establish a direct relationship between the press releases and the news articles in the sample, when such a relationship is implied or explicitly stated. In Chapter Four, I argued that one way to be reasonably certain that an article is, indeed, based on the text of a press release – i.e., that the journalist *used the text of the press release* as a source when writing the article – is to have a long enough sequence of words appearing in both the press release and the article. Naturally, the article must have been published soon after the press release was issued by the company. Identifying matching text spans has been shown to be a difficult task which cannot be accomplished through a manual approach within a reasonable period of time. Thus, a computational solution is the only reasonable way to establish this relationship using matching text spans as the definitive criteria.

There are numerous technical issues involved in designing a computational solution for this problem, most of which are beyond the scope of this dissertation. Instead

of focusing on these technical issues, I discuss the main steps and the key algorithmic choices I made in the design process, which include:

- finding matching sequences of text;
- selecting the minimum sequence length;
- eliminating matches which do not discriminate between press releases;
- limiting the timespan between press release and news article.

From this point on, for clarity's sake, I will refer to texts of press releases and news articles as documents.

**Finding Matching Sequences of Text**

In the previous chapter I explained how each document can be represented as a sequence of tokens, with a token being the optimal unit of analysis in the context of this study, including the task of discovering relationships between press releases and news articles.

The first step after selecting the unit of analysis is to design or select an existing algorithm to compare documents. As I have explained in Chapter Four, comparing texts with regard to a sequence of matching words (I used an example of a sequence of 10 words) is a very difficult task. In fact, the real task of finding matching sequences of words (or any kind of units) in two documents is much more difficult than I initially described. Consider the following two requirements:

a) We may be interested in sequences of less than 10 words. In fact, we are interested in sequences of at least n words, with $N >= n > 1$ (where $N$ is the length of the shorter document);

b) Furthermore, we are interested in discovering *all* matching sequences, satisfying condition (a) – i.e., there may be more than one matching sequence in one pair of documents.

Therefore, if there are multiple matching sequences of length n, we need an algorithm which will discover *all* such matching sequences. Consider the following basic example:

- Let *a, b, c, d* represent some individual words
- Let *X* represent any sequence of words, excluding *a, b, c, d*
- Let document *D1* be represented as "*ababcabcd*"
- Let document *D2* be represented as "*XabXabcdXabcX*"

One approach to finding all matching sequences of length n, starting with the longest would be to match the longest sequence *abcd* in document *D2*, and after that continue searching on the left and right of that sequence – thus, finding *ab* and *abc*. There are algorithms which do just that. Many of them are implemented and made available through programming language libraries, one of which is the Python *difflib* module, which is the implementation I used for my solution. This algorithm finds "the longest contiguous matching subsequence … The same idea is then applied recursively to the pieces of the sequences to the left and to the right of the matching subsequence" (Python Standard Library, 2013). A more detailed discussion of this algorithm is beyond the scope of this study.

**Selecting Minimum Sequence Length**

Before running the selected algorithm to find matching sequences in two documents, one must decide what is the minimum length for a sequence to be

meaningful. After all, the goal is to find a connection between two documents where a span of matching text is assumed to indicate that one document was used for writing the other. A sequence of two words is, clearly, a bad indicator of any kind of relationship. One approach is to use basic information retrieval methods, such as calculating TF/IDF frequencies – i.e., term frequencies/inverse document frequencies (Wikipedia, 2013e), to determine which words are better discriminators between documents and, therefore, are better indicators of text being copied from one document to another, as opposed to being used in identical form due to chance. However, this approach would have steered the dissertation away from its primary focus. Given the fact that using a minimum sequence length as a cut-off would be a good enough solution within the context of this study, I opted for this solution. After some experimentation, I set the minimum sequence length to be seven (tokens), excluding any punctuation which was not part of a token.

I immediately ran into problems. There were sequences of seven or more tokens which, clearly, did not indicate a relationship between documents. Such sequences included company names, names of executives with their titles (with the titles being quite descriptive), or combinations of such sequences. For example, the sequence "said Dave Melton, a driving safety expert with Liberty Mutual Insurance and managing director of global safety" (Liberty Mutual, 2012) consists of 17 words – which is a considerable length for matching text, yet it is common to many press releases of the company in question, and even if it were unique to one press release, it is very generic and is not a good indicator that the article was based on the press release (i.e., whatever "Dave Melton said", the journalist, most likely, would use this very phrase to wrap the quote.

Another common problem was that an article could be covering the same topic as the press release, but not be based on the press release, which was evident from the text, but only through manual examination. The most common example – product release news: product information is usually included both in the press release and in the article, but also in product descriptions found in its packaging, in its advertising messages, or in numerous interviews. The problem is that such content may span much more than seven words – and yet tell us nothing about the relationship between the press release and the article.

The only reliable solution I found was to combine automation with manual filtering: after finding all matching sequences of length seven or greater and running a number of automated filters to remove the some irrelevant data, I manually examined every single matching case – which resulted in the elimination of roughly 20% of the cases.

**Eliminating Bad Discriminators**

One major issue surfaced after the first run of the matching algorithm. My task was to use each matching sequence to pair a specific press release to one or more articles. However, it turned out that there were matches which appeared in more than one press release. I expected this to happen with press release boilerplates – i.e., the same text which is tagged on to almost every press release describing the organization issuing the release. However, it turned out that identical text could occur anywhere in the body of a press release.

I wrote a program which analyzed the discovered pairs of articles and press releases identifying all instances of matching sequences of text appearing in more than one press release. As expected, the identified instances were mostly boilerplate text (all text is lowercase, punctuation-free):

> …kroger the nation 's largest traditional grocery retailer… (Kroger, 2012)
>
> …stores in all 50 states the district of columbia puerto rico u.s. virgin islands guam 10 canadian provinces mexico and china… (Home Depot, 2012)
>
> …is one of the nation 's leading providers of entertainment information and communications products and services… (Comcast, 2012)

Allstate appeared to be standing out in terms of the amount of such reusable text which often included descriptions of promotional programs or events organized by Allstate, as well as organizations which appeared to be partners in such programs:

> …is the nation 's largest publicly held personal lines insurer… (Allstate, 2012a)
>
> …largest publicly held personal lines insurer serving approximately 16 million households through its allstate encompass esurance and answer financial brand names widely known by its slogan… (Allstate, 2012a)
>
> …commitment to strengthen local communities the allstate foundation allstate employees agency owners and the corporation provided 28 million in 2011 to thousands of nonprofit organizations and important causes across the united states… (Allstate, 2012c)
>
> …about the survey this survey of americans age 18 and over was conducted by phone december 3-6 2011 among a nationally representative sample of 1,000 american adults the margin of error for the national sample of residents is 3.1 percent the survey was conducted by fti consulting inc. fti for allstate…(Allstate, 2012b)

Another very common type of repeating elements was text of the form *said [name of person], [title of person]* – which, certainly, is to be expected:

> …said sara nelson editorial director of books and kindle… (Amazon, 2012)
>
> …said samuel r. allen chairman and chief executive officer… (Deere, 2012)

As a result of this analysis, if a non-unique sequence was the only connection between a press release and a news article, I discarded the pair: for if the same sequence appeared in another press release, there was no reliable way to establish the relationship between the news article and the "right" press release. Thus, the final data set used to answer the research questions in this study does not contain instances of relationships where the article may be related to more than one press release.

**Limiting the Timespan**

One more consideration was in regards to the timespan between the press release and news article. Among the identified potential pairs of documents there were press releases and news articles with many months separating them. In all of these cases the matching text was short and generic. Besides, considering the amount of press releases supplied to news media (discussed in Chapters One, Two and Three), it is very unlikely that a journalist will use a press release as a source for an article which is several months old. Therefore, I established an additional limitation for my procedure; I only considered press releases and news articles which were separated by one month or less.

There were numerous other considerations in regards to establishing a relationship between a press release and an article; however, all of them were technical in nature, and discussing them is beyond the scope of this study. However, key implementation details are available in Appendix D.

**The Final Data Set**

Overall, each of the 6,171 press releases was compared to each of the 48,664 news articles, which resulted in more than 300,000,000 document comparisons. The constructed data set consisted of 797 press releases, used as sources in 1,522 news articles, which constituted 1,643 relationships between press release and news article, based on 3,493 instances of matching text sequences with a length of seven tokens or greater. The data set is summarized in Table 5. The updated method summary is displayed in Figure 7. Appendix F contains a screenshot of a generated webpage displaying a press release and a news article side-by-side with matching text highlighted.

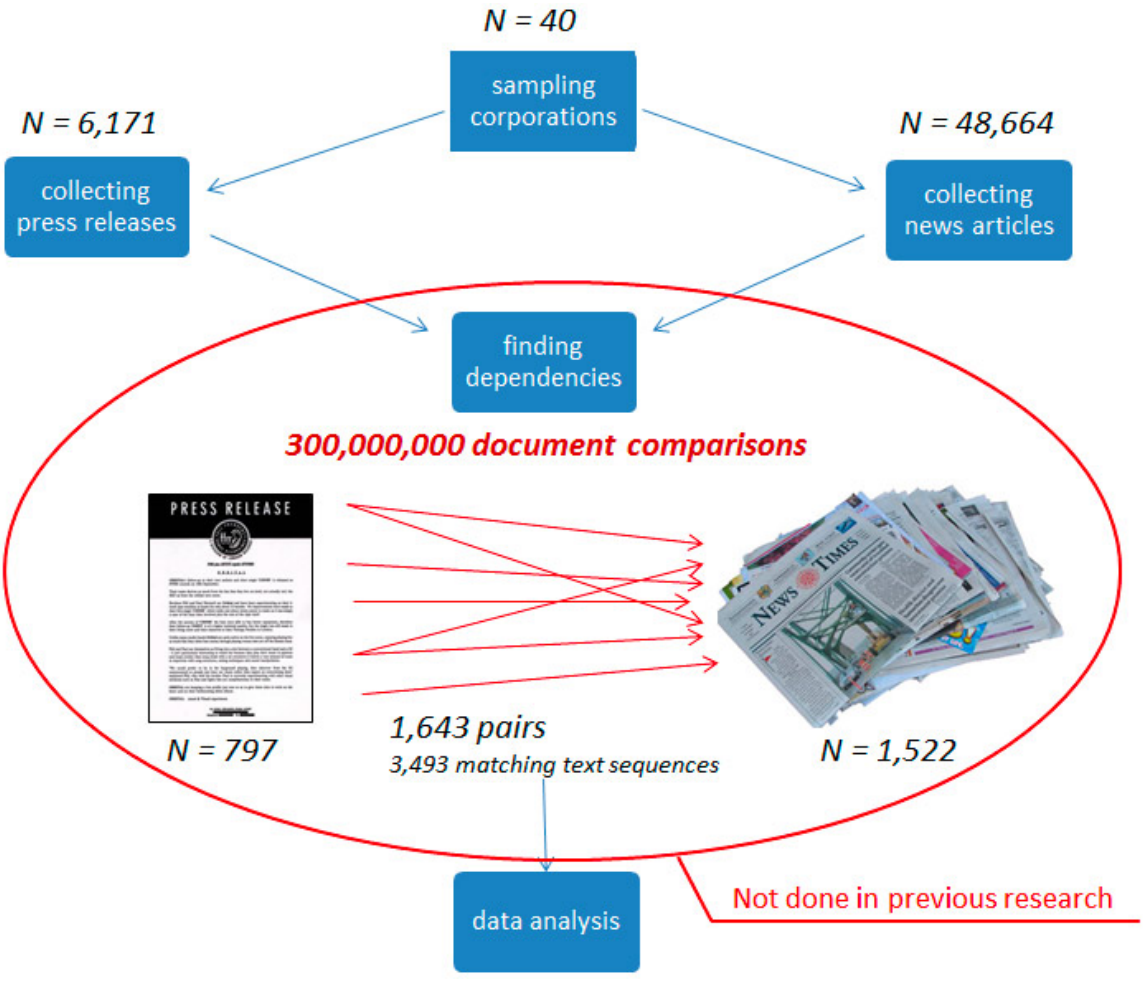



*Figure 7*. Updated method summary.

Table 5. *Final data set.*

| | All Data | | Matches (Block Length >= 7) | | | |
|---|---|---|---|---|---|---|
| Company | Releases[a] | Articles[b] | Releases[c] | Articles[d] | Pairs[e] | Blocks[f] |
| Ford Motor | 529 | 4,162 | 150 | 277 | 336 | 543 |
| Apple | 55 | 3,979 | 32 | 176 | 193 | 343 |
| Lockheed Martin | 393 | 1,071 | 46 | 63 | 71 | 192 |
| Wells Fargo | 184 | 2,707 | 45 | 67 | 71 | 219 |
| Microsoft | 253 | 895 | 25 | 66 | 66 | 127 |
| Bank of America Corp. | 298 | 2,873 | 40 | 65 | 66 | 175 |
| Boeing | 359 | 3,709 | 32 | 57 | 60 | 113 |
| Target | 235 | 619 | 27 | 49 | 51 | 125 |
| Delta Air Lines | 138 | 1,518 | 24 | 45 | 46 | 76 |
| Citigroup | 73 | 1,821 | 13 | 44 | 44 | 82 |
| American Express | 127 | 1,702 | 26 | 44 | 44 | 125 |
| Comcast | 90 | 3,937 | 22 | 44 | 44 | 113 |
| Amazon.com | 162 | 2,293 | 23 | 42 | 42 | 83 |
| Exxon Mobil | 67 | 1,399 | 13 | 36 | 38 | 62 |
| Fannie Mae | 311 | 3,036 | 16 | 32 | 34 | 90 |
| Abbott Laboratories | 105 | 365 | 16 | 26 | 29 | 63 |
| CVS Caremark | 134 | 399 | 19 | 21 | 28 | 68 |
| Aetna | 212 | 567 | 18 | 26 | 28 | 65 |
| United Continental Holdings | 121 | 467 | 19 | 25 | 27 | 63 |
| Allstate | 104 | 1,100 | 18 | 26 | 27 | 95 |
| MetLife | 97 | 442 | 12 | 26 | 26 | 70 |
| International Business Machines | 357 | 388 | 21 | 25 | 25 | 91 |
| WellPoint | 72 | 391 | 8 | 24 | 24 | 35 |
| PepsiCo | 197 | 877 | 16 | 23 | 24 | 49 |
| State Farm Insurance Cos. | 43 | 777 | 8 | 23 | 23 | 53 |
| Sprint Nextel | 321 | 488 | 6 | 23 | 23 | 42 |
| FedEx | 114 | 1,142 | 15 | 19 | 21 | 42 |
| Merck | 108 | 405 | 13 | 17 | 17 | 23 |
| Home Depot | 30 | 1,954 | 12 | 13 | 14 | 21 |
| Liberty Mutual Insurance Group | 50 | 161 | 9 | 13 | 13 | 41 |
| Sunoco | 42 | 283 | 3 | 12 | 12 | 29 |
| Archer Daniels Midland | 74 | 170 | 6 | 11 | 11 | 25 |
| Johnson Controls | 89 | 243 | 7 | 11 | 11 | 31 |

(continued)

| | | | | | | |
|---|---|---|---|---|---|---|
| New York Life Insurance | 63 | 188 | 6 | 9 | 9 | 15 |
| Kroger | 60 | 328 | 6 | 9 | 9 | 23 |
| Safeway | 64 | 198 | 6 | 8 | 8 | 12 |
| Morgan Stanley | 61 | 237 | 5 | 6 | 8 | 18 |
| Deere | 72 | 722 | 6 | 7 | 8 | 32 |
| McKesson | 137 | 100 | 2 | 6 | 6 | 8 |
| J.P. Morgan Chase & Co. | 170 | 551 | 6 | 6 | 6 | 11 |
| TOTAL | 6,171 | 48,664 | 797 | 1,522 | 1,643 | 3,493 |

[a]Number of press releases in the text corpus. [b]Number of news articles in the text corpus. [c]Number of press releases identified as sources for news articles. [d]Number of news articles which used press releases as a source. [e]Number of relationships between press releases and news articles. [f]Number of matching sequences of text, or blocks, in relationships between press releases and news articles.

## Research Question Findings

### Computing Proportions of Matching Text

**RQ1:** Research Question One stated: Given a press release, which is used as a source for a news article, what is the proportion of the press release text used without any change? In other words, how much of the press release text is used verbatim? To address this question, I computed the proportion of press release content appearing in the news article verbatim to the entire content of that press release.

The results demonstrated that, news articles which contain text copied verbatim from a press release, copy on the average six percent of that press release ($M$ = .06, $Mdn$ = .03, $s$ = .0021). The remaining 94% is not used verbatim. Figure 8 demonstrates a frequency distribution of these values. Thus, in the majority of cases, news articles copy only a small part of the press release. In particular, only in one percent of the cases, news articles copied *at least 41%* of the press release; in 84% of the cases, news articles copied *less than 10%* of the press release; and in 24% of the cases, news articles copied *less than*

*one percent* of the press release. These numbers, however, do not imply that this is the extent to which the press release content is used: they only report verbatim usage. Further research is needed to measure and analyze potential instances of paraphrased text.

Appendix G demonstrates examples of different levels of press release content used verbatim in news articles.

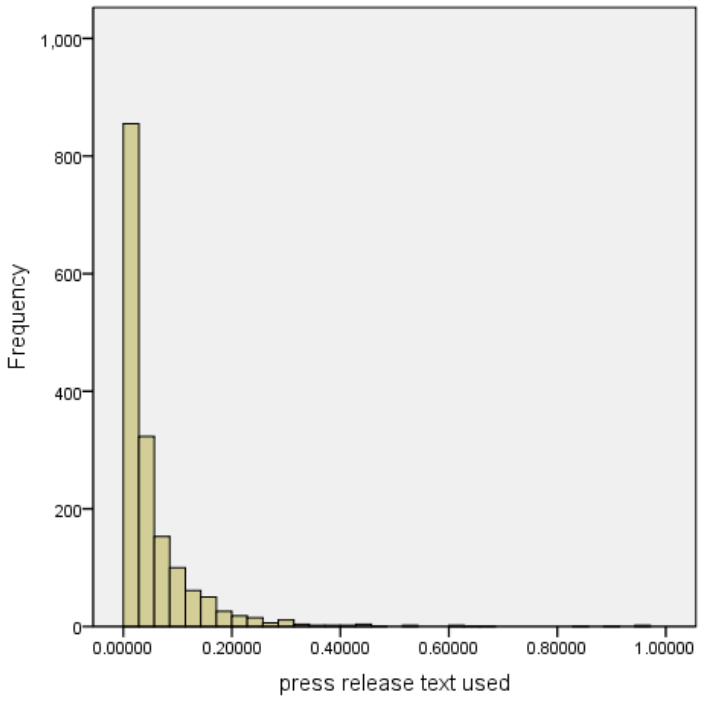



*Figure 8.* Proportions of press release verbatim usage in news articles.

Data aggregated by company, displayed in Table 6, shows that average press release verbatim usage ranges from 1.2% to 10.9%, with companies like Allstate and Lockheed Martin boasting the highest proportion of their press release verbatim usage by the publications in the data set.

Table 6. *Press release content used in articles aggregated by company.*

| Company | Used | Cases |
|---|---|---|
| Allstate | 10.85% | 27 |
| Lockheed Martin | 10.79% | 71 |
| Archer Daniels Midland | 9.61% | 11 |
| Kroger | 8.87% | 9 |
| Johnson Controls | 8.39% | 11 |
| Target | 8.26% | 51 |
| Sprint Nextel | 8.18% | 23 |
| Abbott Laboratories | 7.86% | 29 |
| Fannie Mae | 7.80% | 34 |
| Wells Fargo | 7.59% | 71 |
| International Business Machines | 7.43% | 25 |
| Apple | 7.05% | 193 |
| Liberty Mutual Insurance Group | 6.79% | 13 |
| Microsoft | 6.30% | 66 |
| CVS Caremark | 5.95% | 28 |
| Morgan Stanley | 5.79% | 8 |
| United Continental Holdings | 5.55% | 27 |
| Bank of America Corp. | 5.40% | 66 |
| Boeing | 5.24% | 60 |
| Comcast | 5.17% | 44 |
| Delta Air Lines | 4.99% | 46 |
| American Express | 4.92% | 44 |
| Aetna | 4.50% | 28 |
| Safeway | 4.48% | 8 |
| Citigroup | 4.27% | 44 |
| Deere | 4.06% | 8 |
| FedEx | 3.71% | 21 |
| State Farm Insurance Cos. | 3.63% | 23 |
| Home Depot | 3.59% | 14 |
| Ford Motor | 3.49% | 336 |
| MetLife | 3.37% | 26 |
| Exxon Mobil | 3.33% | 37 |
| Merck | 3.04% | 17 |
| New York Life Insurance | 2.91% | 9 |
| Amazon.com | 2.89% | 42 |
| WellPoint | 2.85% | 24 |
| PepsiCo | 2.77% | 24 |
| J.P. Morgan Chase & Co. | 2.51% | 6 |
| Sunoco | 1.50% | 12 |
| McKesson | 1.20% | 6 |

Data aggregated by publication is available in Appendix H. It ranges from 0.1% to 68.6%, with both extremes represented by small papers with only one case each. Large newspapers, on the other hand, demonstrate numbers which are quite similar among each other: *The Baltimore Sun* uses on the average 2.7% ($N = 11$), *USA Today* uses on the average 3.2% ($N = 50$), *The Philadelphia Inquirer* uses on the average 3.5% ($N = 28$), *The New York Times* uses on the average 3.9% ($N = 104$), *Los Angeles Times* uses on the average 4.3% ($N = 46$), *The Washington Post* uses on the average 4.5% ($N = 56$), and *The San Francisco Chronicle* uses on the average 5.4% ($N = 13$). All these numbers refer to proportions of verbatim text usage.

Aggregated data, of course, is not generalizable within the context of this study and is only considered in the context of the constructed data set. Furthermore, I abstain from making any comparisons between corporations or between publications based on such aggregate data even within the context of this data set. Any such comparison would need to, at the very least, account for the number of cases per corporation or publication, which varies considerably. With regard to corporations, the number of cases varies from six (McKesson) to 336 (Ford Motor). With regard to publications, the number of cases varies from one (represented by 26 newspapers which include small papers like the *Oroville Mercury Register*, as well as large papers like *The Christian Science Monitor*) to 104 (the *New York Times*).

What this data may suggest is that (a) some corporations are covered by news media much more than others, and (b) some publications are less interested in covering big business than others – but this would need to be tested in the context of a different study with a completely different research design.

**RQ2:** Research Question Two stated: Given a news article, which uses a press release as a source, what is the proportion of the article's text not copied without any change from the press release? In other words, how much of the article's text is not copied verbatim from the press release? To answer this question, I computed the proportion of news article content not copied verbatim from the press release to the entire content of that news article.

The results demonstrated that when a news article contains text copied verbatim from a press release, the average proportion of the rest of the article's text – i.e., the text *not* copied verbatim is 92% ($M$ = .92, $Mdn$ = .96, $s$ = .1215). The remaining eight percent of the content is copied from the press release verbatim. Figure 9 demonstrates a frequency distribution of these values. Thus, in the majority of cases, news articles add a considerable amount of text to the text copied verbatim from a press release. In particular, only in two percent of the cases, as much as half of the news article's content was copied verbatim from a press release; in 25% of the cases, verbatim press release content accounted for *less than 10%* of the news article's text; and in 92% of the cases, verbatim press release content accounted for *less than one percent* of the news article's text.

Like in the case of RQ1, these numbers do not imply that this is the extent to which the press release content is used: they only report verbatim usage. Further research is needed to measure and analyze potential instances of paraphrased text.

Appendix G demonstrates examples of different levels of original text being added to press release content used in a news article.

Data aggregated by company, displayed in Table 7, shows that the amount of added text (i.e., text not copied verbatim) ranges from 79.2% to 96.7%, with companies

like J.P. Morgan Chase & Co (*N* = 6), Safeway (*N* = 8), and Exxon Mobil (*N* = 37) compelling journalists to add the most text to their press releases, and IBM (*N* = 25) and Wells Fargo (*N* = 71) – the least.

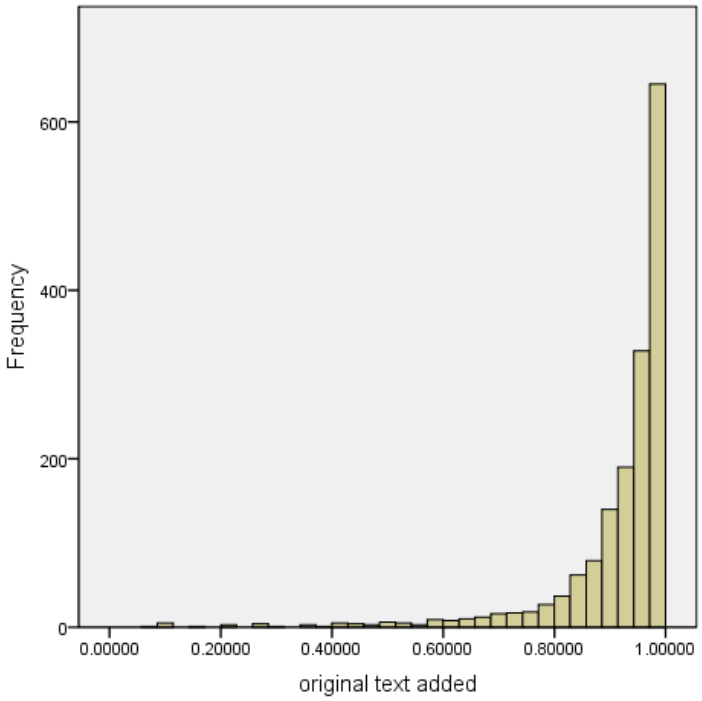



*Figure 9.* Proportions of text added to press release verbatim content.

Data aggregated by publication is available in Appendix H. It ranges from 99.6% to 47.3%, with an outlier of 9.9%. Like in the data for RQ1, extreme values are represented by small papers with only one case each. Large newspapers, as with RQ1 data, demonstrate numbers which are, again, quite similar: *The Baltimore Sun* adds on the average 96.5% (*N* = 11), *USA Today* adds on the average 95.6% (*N* = 50), *The Philadelphia Inquirer* adds on the average 95.5% (*N* = 28), *The New York Times* adds on the average 95.9% (*N* = 104), *Los Angeles Times* adds on the average 93.1% (*N* = 46), *The Washington Post* adds on the average 92.9% (*N* = 56), and *The San Francisco*

*Chronicle* adds on the average 96.0% ($N = 13$). All these numbers refer to proportions of content added to the text which was copied verbatim from a press release. Whether or not such content was borrowed from the press release and paraphrased cannot be determined through these measurements.

Table 7. *Original content added to press release content aggregated by company.*

| Company | Added | Cases |
|---|---|---|
| J.P. Morgan Chase & Co. | 96.61% | 6 |
| Safeway | 96.51% | 8 |
| Exxon Mobil | 96.25% | 37 |
| Ford Motor | 95.12% | 336 |
| Delta Air Lines | 95.11% | 46 |
| WellPoint | 95.01% | 24 |
| Home Depot | 94.99% | 14 |
| McKesson | 94.66% | 6 |
| Boeing | 94.33% | 60 |
| New York Life Insurance | 94.32% | 9 |
| Merck | 94.05% | 17 |
| State Farm Insurance Cos. | 93.97% | 23 |
| Microsoft | 93.60% | 66 |
| Apple | 93.57% | 193 |
| United Continental Holdings | 93.29% | 27 |
| Sprint Nextel | 93.01% | 23 |
| Kroger | 92.89% | 9 |
| FedEx | 92.77% | 21 |
| Citigroup | 92.64% | 44 |
| Sunoco | 92.02% | 12 |
| Amazon.com | 91.70% | 42 |
| PepsiCo | 91.13% | 24 |
| Target | 90.67% | 51 |
| Liberty Mutual Insurance Group | 90.62% | 13 |
| Abbott Laboratories | 90.35% | 29 |
| American Express | 89.51% | 44 |
| Archer Daniels Midland | 88.87% | 11 |
| Lockheed Martin | 88.72% | 71 |
| MetLife | 88.53% | 26 |
| Comcast | 88.38% | 44 |
| Fannie Mae | 88.36% | 34 |
| CVS Caremark | 87.41% | 28 |
| Johnson Controls | 87.25% | 11 |
| Bank of America Corp. | 87.03% | 66 |
| Aetna | 86.88% | 28 |
| Morgan Stanley | 85.44% | 8 |
| Allstate | 83.52% | 27 |
| Deere | 82.35% | 8 |
| Wells Fargo | 80.37% | 71 |
| International Business Machines | 79.19% | 25 |

Like in the case with RQ1, aggregated data is not generalizable and is only considered in the context of the constructed data set. Comparisons between corporations or between publications based on this data cannot be made reliably. It is tempting to conclude that IBM is more trustworthy than Ford Motor, since journalists add 95.1% to the Ford Motor press releases they use – which is more than the 79.2% they add to press releases by IBM, but such conclusions cannot be drawn: sample size alone precludes such comparison: Ford has 336 press releases used my media, whereas IBM has only 25.

## Analysis of Evaluative Language

**Manual coding.** As a first step, a press release and a news article were selected randomly and coded by both coders. The press release and news article, as well as their respective matching pairs, were removed from the data set. A total of 33 sentences were coded. The intercoder reliability for the two raters was found to be Kappa = 0.73 ($p < 0.001$), which demonstrates substantial agreement between coders (Landis & Koch, 1977). (SPSS output is available in Appendix I, Figure I1.)

The coders proceeded to coding the rest of the data (28 press releases and 28 news articles, with a total of 1,174 sentences).

**Automated coding.** As a second step, the same data coded by the two coders was coded automatically by a program which based its calculations on prior term polarity. The intercoder reliability for the manual and automated approach was found to be Kappa = 0.21 ($p < 0.001$), which is too low to consider the results of automated coding to be reliable. (SPSS output is available in Appendix I, Figure I2.)

A low measure of agreement between manual and automated method was expected since word subjectivity and polarity does not imply that the overall sentence subjectivity and polarity will be the same. For example, the adjective *bad* has negative polarity when considered out of context, but in the context of a particular sentence it may be positive: for example, "the movie was *not bad* at all!" Research has shown that word polarity, and even subjectivity in general, is a bad predictor of the polarity and subjectivity of the overall sentence (Wilson, Wiebe, & Hoffmann, 2009). Word sense ambiguity was another problem which was not addressed in this study. For example, the most common negative term in both press releases and news articles was the noun *vice*; however, in most cases it was used in combination with the word *president* – i.e., *vice president*, and, therefore, was not used in a subjective sense. (A list of the top 20 subjective terms is available in Appendix E, Table E1).

I calculated precision, recall and F-measure for all coding categories (positive, negative, both and neutral) to see what categories were most problematic. Table 8 displays the results of these calculations. The results demonstrate the following:

a) The program does very poorly when identifying a sentence as negative or both positive and negative: it is correct only in nine percent for the "negative" category and only in six percent for the "both" category.
b) The program performs better than average identifying sentences which were perceived by the coders as positive (the program correctly identified 65% of such sentences).
c) Most of the sentences were perceived to be neutral by the coders, which suggests that either press release and news media text is mostly neutral, or

the coding instructions were too general – i.e., most sentences were treated as not being subjective despite them containing subjective terms.

Since the results of the intercoder reliability test applied to human coding versus automated coding were too low to be considered reliable, I use the data coded by human coders to answer research questions RQ3 and RQ4.

Table 8. *Precision, Recall, and Harmonic Mean (RQ3 and RQ4).*

| | Human | Program | TP | TN | FP | FN | Precision | Recall | F-Measure |
|---|---|---|---|---|---|---|---|---|---|
| Positive | 159 | 393 | 103 | 771 | 290 | 56 | 0.26 | 0.65 | 0.37 |
| Negative | 38 | 96 | 9 | 1,095 | 87 | 29 | 0.09 | 0.24 | 0.13 |
| Both | 17 | 121 | 7 | 1,089 | 114 | 10 | 0.06 | 0.41 | 0.10 |
| Neutral | 1,006 | 610 | 579 | 183 | 31 | 427 | 0.95 | 0.58 | 0.72 |

*Note*. TP: true positives; TN: true negatives; FP: false positives; FN: false negatives.

**RQ3: How press releases and corresponding news articles compare in terms of subjective language.** Research Question Three stated: How does press release content compare to the content of a news article, which uses that press release as a source, in regards to the evaluative, or subjective, language they use? I hypothesized that a news article will use less evaluative, or subjective, language compared to the press release it uses as a source.

Since the results of the intercoder reliability test applied to human coding versus automated coding were too low to be considered reliable, I conducted a statistical test using the results of manual coding only.

A paired samples *t* test demonstrated a statistically significant difference between the mean subjectivity score of press releases ($M$ = .20, $s$ = .14) and news articles ($M$ = .14, $s$ = .08), $t(27) = 2.059$, $p$ = .025, $\alpha$ = .05, which supports my hypothesis. I conclude that a news article uses less evaluative language compared to the press release. SPSS output is available in Appendix J, Figure J1.

**RQ4: How press releases and corresponding news articles compare in terms of the polarity of their evaluative language.** Research Question Four stated: How does press release content compare to the content of a news article, which uses that press release as a source, in regards to the polarity (positive versus negative) of the evaluative, or subjective, language they use? I hypothesized that the language of the news article will be less positive compared to the language of the press release it uses as a source.

As in RQ3, since the results of the intercoder reliability test applied to human coding versus automated coding were too low to be considered reliable, I conducted a statistical test using the results of manual coding only.

A paired samples *t* test demonstrated a statistically significant difference between the mean polarity score of press releases ($M$ = .78, $s$ = .40) and news articles ($M$ = .53, $s$ = .68), $t(27) = 1.876$, $p$ = .036, $\alpha$ = .05, which supports my hypothesis. I conclude that a news article uses less positive language compared to the press release it is based on. SPSS output is available in Appendix J, Figure J2.

Thus, the data demonstrates that news articles are perceived to be less subjective and less positive compared to the press releases they are based on. A closer look at the data provides an abundance of specific examples supporting these results.

With regard to evaluative, or subjective language, the data makes it evident that journalists "tone down" the overly positive nature of press release content by simply ignoring most of it. A press release from PepsiCo starts with the a statement overflowing with positive sentiment:

> Doritos Dinamita rolled flavored tortilla chips, a spicier new offering debuted by PepsiCo's Frito-Lay division today, is everything fans love about Doritos tortilla chips, only rolled into a taquito-like shape (PepsiCo, 2012).

Probably, to make sure that the reader (or editor) gets the message that consumers – who, apparently, are not mere consumers, but *fans* of the product – do, indeed, love Doritos, that same press release reinforces the previous statement with a quote from its vice president of marketing:

> "The Doritos Dinamita product line truly accomplishes that, offering consumers everything they love about traditional Doritos tortilla chips, but with a spicy new twist" (PepsiCo, 2012).

The press release goes on to talk about "mak[ing] enjoyable foods and beverages that are loved throughout the world; ...find[ing] innovative ways to minimize [their] impact on the environment …; provid[ing] a great workplace …; and respect[ing], support[ing] and invest[ing] in local communities." (PepsiCo, 2012).

This is a typical press release. Other typical examples are statements which start with the words "we are pleased to have built upon our longstanding relationship with…" (Microsoft, 2012), "we are proud of the continued success of our program…" (Microsoft, 2012), "we are focused on delivering the next generation of…" (Aetna, 2012), or "we remain committed to working with…" (Aetna, 2012). However, most of this overly subjective and promotional language does not make it into news articles. As a result, the

overall tone of news articles which use press releases as their source appears to be less subjective.

Based on the results of testing RQ4, it appears that news articles tend to be not only less subjective; the subjective language of news articles appears to be less positive compared to the press releases which they use as a source.

Consider a press release by Fannie Mae (2012) and an article from the Deseret Morning News (2012) which has used at least nine percent of that press release (i.e., we do not know how much of the text was paraphrased in addition to the nine percent used verbatim). The press release contains sentences which have been coded as positive or neutral. Following are some examples of sentences perceived by the coders as positive:

- There is marked improvement in consumer sentiment regarding the direction of the economy, personal finances, and future home price expectations, " said Doug Duncan, vice president and chief economist of Fannie Mae.
- Fannie Mae exists to expand affordable housing and bring global capital to local communities in order to serve the U.S. housing market.
- Our job is to help those who house America .

The corresponding article does not contain these sentences; instead it has quite a few which were perceived as negative:

- People aren't dancing in the streets with optimism quite yet, but the results from Fannie Mae's December National Housing Survey show that Americans' attitudes on several economic issues are a tad better than in November.
- But not everybody is optimistic.
- The survey, however, did not ask if they danced in the street over the increase.

Thus, the resulting news article becomes much less positive compared to the source of the news.

A particularly curious example of "positive spin" comes in the form of interpretation of statistical data:

> Americans who say the economy is on the right track rose by 6 percentage points since November, while the percentage who say the economy is on the wrong track dropped by 6 percentage points (Fannie Mae, 2012).

The sentence contains two positive statements: (1) "Americans who say the economy is on the right track rose by 6 percentage points" and (2) "the percentage who say the economy is on the wrong track dropped by 6 percentage points." But why is this spin? Because the two statements, most likely, refer to just one positive change: six percent of those who thought economy was on the wrong track now think it is on the right track. Unless, of course, there is another group of those who are undecided, or think otherwise – however, such a group is not mentioned. But still, the news article avoids falling into this trap and uses a different quote describing this data:

> However, while December results show that more Americans think the economy is on the right track, consumer attitudes are still at depressed levels, with more than two-thirds saying that the economy is on the wrong track (Deseret Morning News, 2012).

Thus, based on this data, it appears that despite using press releases as a source, news articles appear to be less subjective and more neutral and even negative in tone, toning down the overly positive, promotional language of the press release they use as a source.

## Analysis of Attribution

**RQ5:** Research Question Five stated: Do news articles, which use press releases as a source, provide attribution? In other words, do such articles mention the press release as the source of their content?

To examine whether or not journalists make *explicit* attributions to a press release when they use its content, I searched each article in the data set for occurrences of the words "press release" and "news release." Out of 1,642 news articles based on press releases, 171 (10%) mentioned the press release explicitly (referring to it as "press release" or "news release"). I tested a 10% random sample of the identified attributions for false positives and found none. I conclude that news articles provide explicit attribution to press releases as their source. SPSS output is available in Appendix K, Figures K1 and K2.

To be sure, these results are limited to explicit attributions to the press release – i.e., if an article contained other forms of attribution, such as “In a statement, Edsel B. Ford II, said…” (*Automotive News*, May 14), such attribution was not taken into account by this method. The difference between this form of attribution and "said in a press release" (Fannie Mae, 2012) or "according to a press release" (MetLife, 2012) is that the latter leaves no doubt about the source of the article – i.e., a press release – whereas the former, although suggesting a high probability of the source being a press release, offers room for other possibilities as well: was an interview? Was it a statement made at a press conference? Maybe during a phone conversation? A stockholder meeting? Thus, even with attribution to the corporation, without an explicit mention of the communication medium, we cannot be sure what that medium was.

However, this distinction is not significant: after all, what we care about is attribution to the corporation regardless of its form. It becomes even less of an issue in the context of this particular study: a verbatim text match already points to the press release as the most likely source. Therefore, to account for the expected large percentage

of false negatives – i.e., instances of attribution to the press release which may have been missed due to the absence of the words "press release" and "news release", as an additional step, I manually examined press release/news article pairs representing 50% of the companies, selecting companies 1 through 20, which accounted for 38% of the data ($N = 616$).

In my analysis, I used the following four categories describing the different ways an article provided attribution:

1. no attribution;
2. explicit mention of press release/news release as the source;
3. attribution to the corporation or its executive;
4. explicit mention of the source outside of the body of the article.

The first category was used when there was no mention of the press release in any form, as well as no direct attribution to the corporation or a corporate executive as the source of the news. The second category was assigned if the news article included the term "press release" or "news release", using it to refer to the source of the news. The third category was the broadest in scope: it included phrases like "corporation/executive said" or "according to corporation/executive" and their variations combined with words like *confirmed, announced, reported, estimated, thinks, is predicting, forecasts*, etc. The fourth category represented the few cases when the press release (or its title) was listed as the source outside the body of the news article.

The results demonstrated that almost a quarter (22.6%) of the news articles which had been shown to use a press release as their source did not provide any attribution. Twenty-seven percent provided explicit attribution to the press release. Only in two cases

out of 616 the press release was listed as a source outside of the body of the article. Half of the articles (49.4%) provided attribution without using the words "press release" or "news release." Therefore, a more detailed analysis of the data revealed that more than three quarters (77.5%) of the news articles, that have been shown to use a press release as their source, provide attribution to that source, either explicitly mentioning the press release, or attributing the news to the corporation that issued the press release or to its executive who made the particular statement. Figure 10 demonstrates these results.

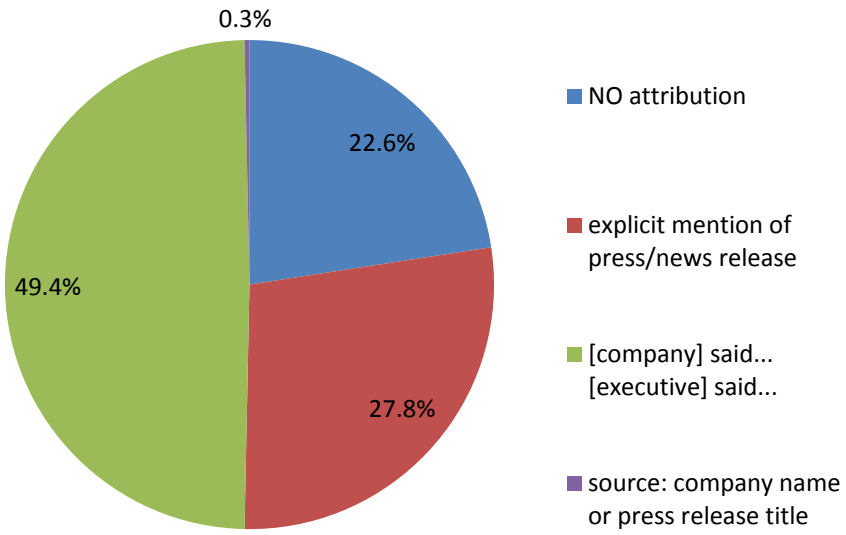



*Figure 10.* Results of manual attribution analysis.

# CHAPTER 7.
# DISCUSSION

## Restating the Problem

This dissertation explores methodological approaches to investigating the problem of news media's use of press release content. This issue of press release content has been discussed from multiple perspectives, including the point of view of journalists, press critics and journalism scholars, who voiced overall concern and distrust not only towards press releases, but public relations in general – i.e., the field that produces them. Authors argued that using content supplied by public relations, particularly press releases, as a news source with little or no edits results in mostly positive news accounts, serving the interests of the organizations supplying that content. Even worse, such news accounts, according to the authors, often provide no attribution to the press release as the source and, therefore, are assumed to be impartial, thus, misleading the audience by hiding the true agenda behind the story.

Public relations scholars offered a different perspective, suggesting that the field of public relations has gone through an evolutionary process over the past hundred years and has evolved from the unethical practice of press agentry, which was largely responsible for the reputation PR has today with journalists, to an ethical, professional field, with its primary concern being mutual understanding. Other scholars, however, have questioned such claims, arguing that this theory is not supported by historical evidence, and provided multiple counterexamples which show that the field has hardly strayed from its original objective, which still appears to be persuasion and manipulation of public opinion.

A review of articles offering the perspective of the public relations industry on press releases has demonstrated further support for the concerns expressed by journalism scholars. From their perspective, the only problem appears to be in getting the news media to publish the information supplied through the press release; any consequences of such practice for news objectivity are nonexistent in this literature. Instead of addressing the underlying problem of press release content serving the purpose of the information source, these articles look for decades-old "surface" solutions to their practical needs – such as suggesting a better typeface or layout to improve the press release and increase its odds at generating publicity for the organization behind it. The authors' ignorance of journalistic values lead to advocating better sales techniques in "negotiating" the story's publication with the editor. Finally, some authors raise considerable ethical concerns, particularly in connection with the disregard they show towards journalists in trying to "disguise" a corporate promotional message as news.

My overview of relevant mass communication theory provided a theoretical grounding for the issue. It explained the underlying goals of the organizations who issue press releases. It explored the broader implications of such content being used with little or no edits by news media, thus, leading to news coverage biased in favor of organizations which have the means for public relations.

**Methodological Issues of Previous Research**

A review of previous research on press releases has revealed a number of shortcomings, revealing a "gap in research." Most importantly, the majority of such research originates in the field of public relations and usually serves its practical needs.

Thus, numerous studies have been conducted on how to make a press release more efficient as a public relations tool and on how to identify specific criteria by which a press release is judged by news media editors.

Studies have examined the relationship between press releases and the news coverage they may have caused; however, such studies rarely provide any reliable evidence establishing the relationship between a press release and a news article. The few studies which provide such evidence by conducting in-depth textual analysis use very small samples which are not representative of the overall population and cannot be generalized.

Studies which use larger samples typically rely on a set of keywords used to retrieve news articles from an electronic database, with such keywords being the only connection between a given press release and a news article – which, however, does not indicate that a press release has been used as the source. Keyword search returns results based on content similarity alone. However, content similarity does not imply causality – i.e., similarity does not indicate that the text of a given press release has been actually used as a source for the news article.

A newsworthy event can be announced through means other than a press release – such as an interview, a press conference, or a website – not to mention the numerous communication tools provided by social media and, undoubtedly, used by organizations to communicate their messages to the public. Considering the breadth and richness of today's media environment, one cannot claim that one press release is responsible for news coverage, based on a similarity in content. Comparing topics is simply not enough;

a better, more reliable method of establishing a relationship between a press release and a news article is needed.

**Verbatim Text Matches as Indicators of Relationships**

In this dissertation, I argue that a reasonably reliable way to establish a direct relationship between a press release and a news article – i.e., to establish that the journalist actually used the press release while writing the article – is to locate text which occurs in both the article and the press release. To be sure, one may look for exact matches or paraphrased text. However, in the context of this particular problem, exact matches are more reliable indicators of a relationship compared to approximate matches.

A verbatim text match between a press release and an article, sufficiently meaningful in terms of length and content, enables a researcher to "tie" a specific press release to a specific article beyond reasonable doubt. Certainly, this approach has limitations. A sequence of words may point to a connection between a press release and a news article; however, it will fail to identify that connection if even one word is changed. It is reasonable to expect that someone copying parts of a press release into a news article they are writing would attempt to make the similarity of the texts less obvious – i.e., the person would make changes. A few changes won't matter of course: as long as there is one exact match, the connection between press release and article will be established. However, if the press release is completely paraphrased – i.e., if there is not a single sequence of matching text – the connection won't be established.

Looking for paraphrased text could be a solution to this limitation. However, there may be a price to pay – and that price is reliability of the connection between press

release and article. Without a verbatim match, it may be harder to argue that the text of the news article is, in fact, a paraphrase of the press release and not of some other source. In previous chapters I have mentioned that there are numerous ways in which content created by a corporation may reach a journalist, and a press release is but one possibility. Paraphrased text suggests a similar topic – i.e., a topic that might have originated from an interview, a press conference, or social media – the scope of possible sources is too broad to tie an article to one of them based on topic similarity.

Certainly, relaying on partial text matches is a much more flexible approach which can uncover instances of similar content which do not share any text verbatim. Thus, methods taking into account paraphrased text can be used to answer broader questions about public relations content in general. However, using such methods to establish a definitive connection between a specific press release and a specific news article for the purpose of examining the use of press release content – not PR content in general – appears to be less reliable than using verbatim text matching.

Therefore, in this dissertation, I have relied on verbatim text matching as evidence of a direct relationship between a press release and a news article.

**Computational Methods as Key to Solution**

Discovering connections between press releases and news articles has been shown to be a very difficult task, which cannot be accomplished using traditional methodological approaches typical in journalism and mass communication research – regardless of whether the matches are exact (i.e., verbatim), or partial (i.e., paraphrased). This constitutes a methodological obstacle which explains why studies examining press

releases and news articles typically do not go into the trouble of identifying direct relationships between the texts. In this dissertation, I have argued that this obstacle can be addressed and, at least, partially overcome with the application of computational methods.

I have demonstrated how computation can be used to process a very large data set, consisting of almost 50,000 news articles and more than 6,000 press releases; and detect relationships between press releases and news articles through locating verbatim text matches of sufficient length. Thus, the significance of using computation in this study lies primarily in the reduction of a very large data set consisting of press releases and news articles exhibiting a similarity in topics – which is not enough to draw any conclusions about a causal relationship between two texts – to a much smaller sample, where press releases and news articles are tied to each other by sequences of matching text.

This sample, referred to as a relevance sample, contains the data most relevant to the specific research questions raised in this study. To the best of my knowledge, this computational sampling approach has not been used within the field of journalism and mass communication research.

## Methodological Implications for Journalism Research

In this dissertation, I utilized a variety of computational methods. Specifically, I used computation to accomplish the following steps:

- Collect the source code of 6,171 web pages containing corporate press releases.
- Extract the relevant data, such as titles, dates and texts, from that source code.

- Transform the loosely structured text of 48,664 news articles into well-formatted data suitable for both computational and manual analysis.
- Reduce the collected data to a sample of most relevant items by identifying press release/news article pairs based on verbatim text matches of sufficient length by conducting more than 300 million document comparisons.
- Compute proportions of verbatim text matches in the resulting 1,643 press release/news article pairs.
- Conduct a series of text processing operations, including sentence and word tokenization, and part of speech tagging, thus demonstrating some of the basic steps used in computational text processing.

To be sure, these steps are not intended to replace in-depth textual analysis and human judgment. However, each of these steps considerably simplifies such analysis. Most importantly, by collecting, and then reducing a very large initial data set to a sample of most relevant items, despite the limitations imposed by verbatim text matching, this computational approach paves the way to constructing data sets which may prove instrumental in investigating the usage of press release content by news media – as evidenced by the striking example of PR influence presented at the end of this chapter.

Replicating the research design of this dissertation, including the construction of a similar data set requires some practical knowledge of computer science. However, this dissertation is not intended as a set of concrete implementation guidelines, although I provide an ample selection of code samples which cover the key elements of the system and should be adequate as a roadmap for an interdisciplinary research team which has a computer scientist on board.

Instead, this work is intended as a model for applying computational thinking – a problem-solving approach introduced in Chapter Four – to mass communication research. Having a program perform repetitive tasks instead of doing them manually is the core of such thinking. Through this approach – i.e., delegating any repetitive work to the computer – we free ourselves to focus on the interesting problems requiring human thought.

Writing programs to tackle repetitive tasks is a matter of common sense in computer science. Unfortunately, what's common sense in computer science is far from common knowledge in journalism and mass communication studies. In Chapter Four I showed that even renowned communication methodologists demonstrate very shallow knowledge of what computers are capable of doing, and how they can be used for research in other fields. As a result, graduate students who learn their methods from this literature go out into the field assuming that computers can do very little for their research. This is, in many cases, incorrect, as this study demonstrates. This dissertation shows how computer science can help locate, collect, process, and analyze data, using a variety of computational approaches and software tools. The introduction of basic tools used in computational linguistics research is a particularly promising gateway for future research involving large samples of text data, thus, requiring at least partial automated processing and analysis. However, the step involving the construction of the data set used for all the subsequent data analysis is the key computational ingredient: it supplies the data which is needed before one can begin addressing the research questions. This one step, through the application of computer science, deals with the methodological problem which, in my opinion, has been the key reason for the gap in research I have identified

and discussed in previous chapters. This study takes a step towards reducing this gap, replacing it with knowledge supported by empirically collected data.

Collaboration with computer scientists, computational linguists, and scientists from related fields will be instrumental in tackling larger samples and discovering new data which will enable journalism and mass communication researchers to address questions which may have been out of reach using a manual approach. Such collaboration has not been common in journalism and mass communication research, which is unfortunate. Douglas Oard (2009) offers the perfect summary of the collaboration challenge:

> We regularly hear impressive claims for what future technology – always, it seems, future technology – will be able to do for us. Why is this future perpetually just over the horizon? The reason, I argue, is simple: those who could build these marvels don't really understand what marvels we need, and we, who understand what we need all too well, don't really understand what can be built (p. 34).

This dissertation is one attempt to respond to this challenge. Its overall methodological contribution is to begin building a bridge between fields.

## The Problem of Press Release Content in News Media

The primary methodological step in this dissertation was to reduce a very large data set containing press releases and news articles exhibiting similar content to a much smaller data set – i.e., a relevance sample – which contained press releases and news articles, explicitly tied to each other by sequences of matching text. I used the constructed sample to address several specific research questions describing several aspects of the problem of press release content usage by news media. My research questions RQ1 and RQ2 addressed the extent to which press release content is used by news media verbatim.

The primary limitation of this approach is that it does not take into account paraphrased text. A closer look at the data demonstrated that while only part of the press release content is copied into the news article verbatim, other parts of the press release often appear in that article in paraphrased form. Therefore, my results for RQ1 and RQ2 only demonstrate the proportions of press release content used verbatim and cannot be applied to determining the overall scope of news media's dependence on the press release.

Nevertheless, my results provide definitive conclusions about the scope of *verbatim* usage of press release content. According to my results, news articles which contain text copied verbatim from a press release, copy on the average six percent of that press release verbatim, with the remaining 94% of the press release either not used at all, or used – partially or completely – in paraphrased form. As for the news article, the average proportion of the rest of the article's text – i.e., the text *not* copied verbatim is 92%. The remaining eight percent of the article's content is copied from the press release verbatim.

In terms of what types of content was copied verbatim more often – there were no obvious patterns: the text selected for inclusion in news articles varied considerably. In many cases, the proportion of the press release text used verbatim was quite small, but the rest of the article was, essentially, a paraphrase of the press release. Here are several examples of such paraphrase:

> *Example 1:*
> Rosneft and ExxonMobil today signed agreements to implement a long-term Strategic Cooperation Agreement concluded in August 2011 to jointly explore for and develop oil and natural gas in Russia and to share technology and expertise.

*Example 2:*
Irving, Texas, oil and gas giant Exxon Mobil Corp. said Monday, April 16, it finalized its strategic cooperation agreement with Russia's OAO Rosneft to jointly explore for and develop oil and natural gas in Russia and share technology and expertise.

Here is another example: the verbatim match is only a small part of the paragraph, whereas the rest of the paragraph is paraphrased:

*Example 3:*
MetLife Home Loans will continue to service its current mortgage customers . In addition , MetLife Home Loans will honor all contractual commitments for loans in process and expects the majority of loans to close in 90 days .

*Example 4:*
MetLife Home Loans also is continuing to service current mortgage customers and plans to " honor all contractual commitments for loans in process , " which the company expects to close by sometime in April .

Examples one and two, as well as three and four, are almost identical, and the only parts which indicate that one of each is not a press release are the references to a statement made by the company: "said Monday, April 16" and "the company expects." If it were not for these references, the articles could have been easily taken for a press release and vice versa. The companies are ExxonMobile (2012a) and MetLife (2012), and the papers are *The Deal Pipeline* (2012) and *Long Island Business* (2012).

Consider the following two examples which do not have matching text, yet are almost identical:

*Example 5:*
The two plan to start seismic and environmental programs in the Kara Sea 's Prinovozemelsky blocks later this year with drilling to begin in 2014.

*Example 6:*
In the Kara Sea , plans are under way to undertake seismic and environmental programs of East Prinovozemelsky blocks later this year in anticipation of a potential exploration well in 2014.

Which one is PR, and which one is journalistic paraphrase? It is impossible to tell! The answer is: example 4 comes from is the press release, and example 3 comes from the news article. The company is ExxonMobil (2012a), the newspaper is *The Deal Pipeline* (2012), both published on April 16, 2012.

This study did not compare publications, so I cannot make any conclusions comparing newspapers to one another in terms of their usage of press release material. Therefore, I provide the following example not as a comparison of numbers, but to balance *The Deal Pipeline* example: after all, this newspaper focuses on this type of content, so, at least, it might be not that surprising that a great deal of press release copy ends up on its pages as news. Consider the following text samples (the original formatting is preserved):

> *Example 1:*
> Scientific innovation is the driving force behind Abbott's mission to discover new ways to help patients manage their health," said John Leonard, M.D., senior vice president, Pharmaceuticals, Research and Development, Abbott. "We strive to create an environment that values invention, creativity and collaboration …
>
> *Example 2:*
> Scientific innovation is the driving force behind Abbott 's mission to discover new ways to help patients manage their health," said Dr. John Leonard, senior vice president, pharmaceuticals, research and development at Abbott.
>
> "We strive to create an environment that values invention, creativity and collaboration …

The parts identified automatically as a reliable matching sequence are underlined. Which one is the press release? It's Example 1, and the company is Abbott (2012). But how much did the journalist really change? Nothing at all, is the answer.

Here's a different example:

*Example 1:*
Prior to these changes, Mark LaNeve, senior executive vice president, agency operations and chief marketing officer, had resigned for personal reasons.

"I regret having to leave Allstate since our strategy and hard work are beginning to pay off," said LaNeve. "Allstate is a great company with an awesome brand, a winning strategy and strong leadership.

*Example 2:*
"I regret having to leave Allstate since our strategy and hard work are beginning to pay off," LaNeve said in a statement Monday. "Allstate is a great company with an awesome brand, a winning strategy and strong leadership."

Allstate Chairman, President and CEO Thomas J. Wilson said LaNeve's resignation was "a personal one."

In this case, the statements are switched, with one of them paraphrased. Example 1 is the press release from Allstate (2012d). The newspaper in both cases is *The Chicago Daily Herald* (2012b, 2012a), which is not as narrowly focused as *The Deal Pipeline* – which suggests that the type of publication is not necessarily a good predictor of the amount of press release content that publication may use.

**Comparison with Previous Studies**

This study did not set out to investigate the extent to which press release content is used by news media. However, a rough estimate can be made: out of the 6,171 press releases issued by 40 companies and 48,664 news articles assumed to be related to these companies (which was established based on indexing provided by LexisNexis) only 1,522 articles and 797 press releases were found to have instances of verbatim text matches. That's an estimated three percent of news articles. It is important to keep in mind that this estimate is based on verbatim text matches only – i.e., news articles which

use a press release in paraphrased form – with no verbatim matches – were not included in the relevance sample computed in this study. Further research is needed to investigate the extent to which press release content is used in news media in paraphrased form.

All things considered, it is hard to compare these results of this study to previous research in terms of exact numbers. On the one hand, this study does not take into consideration paraphrased text; thus news articles and press releases with no verbatim matches were not discovered with this method and, therefore, were not included in the relevance sample used to address the specific researches questions of this study. On the other hand, previous studies typically do not describe the details of the procedures they employ to establish a direct relationship between a press release and a news article. Furthermore, even though references to verbatim usage of press release content are made, they are typically not supported by exact numbers. Therefore, a direct comparison of results is not feasible.

Nevertheless, some partial comparisons can be made. While most studies discussing the amount of press release copy in news media do not specify exact numbers (e.g., Bagdikian, 1974; Jones, 1975; Lewis, Williams, & Franklin, 2008a, 2008b; Sullivan, 2011), there are some that do. Ambrosio (1980) found that 72% of the articles which she sampled from the Wall Street Journal were based almost exclusively on press releases, while 90% started with a company announcement – which is vastly different from the numbers which have been estimated through this study.

To be fair, Ambrosio's conclusions are based on a journalism review article supported by anecdotal evidence – i.e., this article does not adhere to the rigorous guidelines of formal research: the article's data comes from *one* section of *one* issue of

*one* newspaper! It is hard to imagine a less representative sample. Besides, it dates back to 1980 – so even if it reflects the pre-Internet media reality of 1980, 33 years later it may be of only historical interest.

This article gains prominence by being cited in detail; a whole paragraph is devoted to reporting the specific numbers presented as the article's findings. Furthermore, it's cited not in some obscure conference paper as one might hope; on the contrary – the paragraph appears in the latest 2012 edition of *Mass Communication* by Ralph Hanson, a textbook my University uses in the Journalism History and Structures class (Hanson, 2011, p. 303). The text appears in the chapter on public relations, in a section titled "public relations and society." It is not immediately clear from the text that the numbers cited in a textbook from 2012 come from a study more than 30 years old; yet these numbers are used to describe, if not define, the nature of news media's dependence on public relations for content. This is but one example of how a careless statement in an short opinion article may end up in a textbook, defining the state of the field; thus, educating yet another generation of journalists who will end up treating the field of public relations with considerable mistrust. The problem here is not the mistrust, which may be justified. The problem is the questionable validity of the data presented as empirical evidence supporting the author's argument.

It is tempting to blame this particular case on a careless, non-scientific approach to data sampling and measurement. However, there are similar examples in formal research (e.g. Lewis et al., 2008a, 2008b), which will be relied upon even more. After all, when a peer-reviewed paper uses a large data sample and makes broad generalizations, there is little reason to question the validity of its results. However, as I have pointed out

earlier in this paper (see Chapters Three and Four), often the claimed numbers are unlikely to be supported by reliable method due to the very large computational size of the problem – unless, of course a computational method was used. But usage of computation is uncommon in journalism and mass communication research. Even textbooks on content analysis demonstrate a lack of understanding on behalf of the authors of what computers are capable of doing. This sometimes creates misleading literature which fails to point out ways in which methodology cold be improved. As a result, studies that use reliable methods continue to tackle small samples (e.g., Maat, 2007, 2008; Maat & de Jong, 2012; Van Hout, Pander Maat, & De Preter, 2011), whereas those that use larger samples typically do not produce reliable results (e.g., Lewis, Williams, & Franklin, 2008a, 2008b).

**Analysis of Evaluative Language**

In addition to computing the proportions of press release content used in news articles verbatim, I examined the constructed relevance sample comparing news articles to press releases in terms of evaluative, or subjective, language. My goal was to estimate whether the overall tone of press releases was different from that of the corresponding news articles. Specifically, I posed two research questions (RQ3 and RQ4) and stated hypotheses for each. RQ3 was about the difference in overall subjectivity between press releases and corresponding news articles. I hypothesized that the news article will use language that is less evaluative, or subjective, than that of the source press release. RQ4 was about the difference in polarity of the subjective language used in press releases and

corresponding news articles. I hypothesized that the news article will use language that is less positive than the language of the source  press release.

I used both a manual and automated approach to code the data. My automated approach, often used in the initial stages of subjectivity analysis in computational linguistics, was rudimentary: I relied on subjectivity measures established for individual words – which proved to be insufficient. I have explained in previous chapters that the measure of subjectivity and polarity applied to an individual word, referred to as prior polarity, cannot be assumed to represent contextual polarity – i.e., the subjectivity and polarity of a word in the context of surrounding text.  This proved to be the case in this study. A closer examination of the data provided ample evidence of the limitation of measuring the subjectivity of a word out of context. Following is one such example:

> "…Boeing values its long-term partnership with Kansas, and we will continue to work with all of our stakeholders in Kansas in support of a robust aerospace industry in the state" (Boeing, 2012).

This statement is part of a press release on closing a facility in Wichita. The statement is, clearly, positive. The automated coding method recognized it as positive as well (the words *values*, *support* and *robust* were recognized as positive) – so it seems that the automated approach works. However, following is an example of how this positive statement changes its polarity when placed in context by a journalist:

> Yet Boeing's statement Wednesday that it "values its long-term partnership with Kansas, and we will continue to work ... in support of a robust aerospace industry in the state" will be no comfort to its employees as it prepares to exit the city that still calls itself the "Air Capital of the World" (Spokesman Review, 2012)

Clearly, in this context the statement becomes negative. However, the automated approach used in this study did not see any difference: it recognized the same subjective keywords –*values*, *support* and *robust* – and coded the sentence accordingly.

This example is quite typical; it demonstrates why the tone, or subjectivity of a word taken out of context is not a reliable measure of the overall tone of the sentence or the text. Not surprisingly, the intercoder reliability for the manual and automated approach was found to be too low to consider the results of automated coding to be reliable. Therefore, I used the results of manual coding to address the specific research questions posed in this study (RQ3 and RQ4).

I devised simple formulas to describe subjectivity and polarity, described in Chapter 5. The data supported my hypotheses for both research questions. Thus, in regards to RQ3, I concluded that news articles which are based on press releases use a more subdued language: a closer examination of the data provided examples showing a news article tone down the language of the press release it used as a source. In regards to RQ4, I concluded that the tone of news articles appears to be less positive than the tone of the press release they use as a source. Again, a closer examination of the data showed that journalists usually remove most of the positive subjective statements from the press release text which they use verbatim; they also typically add neutral or even negative subjective statements to their articles, thus, making the article more balanced – as evidenced in the example with Boeing.

**Attribution Analysis**

The problem of attribution is yet another concern about the nature of press release usage by news media, common among by journalists, press critics and journalism scholars, who have argued that typically articles do not mention press releases as the source of the news despite using their text, often with little or no edits.

In this study, I started by testing for explicit mentions of "press release" or "news release" in news articles which used press releases a as source. My results indicated that in 10% of the cases there was an explicitly stated attribution to the press release.

However, a closer look at the data suggested that there were many instances of proper attribution where the terms "press release" or "news release" were not explicitly mentioned. Consider the following examples:

> In a statement, Edsel B. Ford II, said: "Today, we have lost a legend in Ford Motor Company's history, and my family and I have lost a dear friend... (*Automotive News* , May 14)

> Target suggested in a statement it issued Friday that the deal doesn't do enough to fix the system." Target has no interest in surcharging guests who use credit and debit cards in order to allow Visa and MasterCard to continue charging unfair fees, "the company said. (*Star Tribune (Minneapolis, MN)*, July 24)

> On Wednesday, Microsoft said, "We take this matter very seriously and moved quickly to address this problem as soon as we became aware of it." (*The New York Times*, October 25)

Such forms of attribution seemed to be quite frequent in the data. Therefore, to account for the expected large percentage of false negatives, I took an additional step and manually examined press release/news article pairs representing 50% of the companies, which accounted for 38% of the data ($N = 616$).

The results of this more detailed analysis revealed that more than three quarters (77.5%) of the news articles provided direct attribution to a press release as their source, either explicitly mentioning the press release, or attributing the news to the corporation that issued the press release, or to its executive who made the particular statement.

Unfortunately, the studies and articles I have reviewed did not quantify the concern over the lack of attribution – i.e., authors did not mention specific numbers to describe the problem, instead using terms like *many*, *often*, or *vast majority*. Since I

cannot reliably estimate how 77.5% relates to such general statements, I do not offer any comparison between my result and general claims made in literature.

## Theoretical Implications of the Data

This study has no direct theoretical implications for the simple reason that there is no mass communication theory centered on the press release. Nevertheless, there may be indirect implications.

In Chapter Two, I discussed how a press release is conceptualized as an information subsidy – i.e., pre-packaged information offered by organizations to news media intended to offset the media's cost of producing their own information (Gandy, 1982). I discussed how the usage of such content by news media may not only lead to organizations which have the funds to have more news coverage than those who don't; but also how it would complicate the news media's task to provide a balanced and impartial account of reality – for the content prepared by an interested party is assumed to be never impartial and is biased by design.

Through a review of agenda-setting – a mass communication theory which has been thoroughly tested over the past several decades – I showed further implications of such prepackaged content being published in the news. By engaging in the process of media building, the public relations industry not only gets to obtain news coverage; through such publicity it contributes to setting the public agenda. Thus, in the words of Bernard Cohen, a corporation may be telling the public what to think about. Furthermore, an examination of the concept of framing as a second dimension of agenda-setting suggested that by publishing press releases with little or no change, the news media relay

to their audience not only the objects, or issues, included in a press release, but also the objects' attributes, which constitute the frames carefully constructed by the organizations supplying the press release. Research has shown that those frames will be favorable towards the organization, promoting its viewpoints on issues in the press release. Therefore, through a mere press release, not only can an organization "tell the news media audience what to think about" – it might also tell the audience *what to think*, or at least *how* to think about the issues the organization has contributed to the public agenda – thus creating a favorable public opinion, which has been shown to be the ultimate goal of information subsidies.

The central argument of this study was methodological; therefore, the main significance of the conducted data analysis is that it explores a new methodology. However, the results of conducting this analysis do, in fact, suggest that the problem of press release content usage by news media might be not as severe as presented in previous research. The main limitation of this study was its reliance on verbatim text matches between press releases and news articles. However, relying on verbatim matches as opposed to paraphrased text was essential to the study's design: in previous chapters I explain how such an approach helps establish a reliable connection between a specific press release and a news article. Nevertheless, this limitation prevents me from making conclusions about the proportion of press release content in news media: for my study accounted for text copied from the press release verbatim – which may be the tip of the iceberg.

Still, my results provide definitive conclusions about the scope of *verbatim* usage of press release content. Clearly, whatever bias, or whatever frames may be embedded in

a press release are guaranteed to be present in a news article that uses the press release without any change. That is why the issue of press release content being used verbatim has been raised by journalists, press critics and journalism scholars, who expressed deep concerns about "a large gray zone between the truth and a lie" which is exploited by skillful PR people to a great extent (Sullivan, 2011, p. 36). However, the data in this study have shown that instances of press release content being copied into news articles verbatim, are not as frequent as previous studies might have implied.

Furthermore, the data suggested that journalists often transform the content of the press release – even when copying parts of it verbatim – making it appear less subjective and less positive than the corresponding press release. This may be interpreted, at least to some extent, as indicating more balanced, impartial and factual news accounts that reframe the issue or dilute the PR frames.

By the same token, the data showed that more than three quarters of news articles that use press releases as a source, copying at least part of the text verbatim, provide attribution to those press releases. It is reasonable to suggest that regardless of what framing techniques had been used in the text of the press release, their effect will be considerably diminished by the attribution present in the article, since such attribution points out that the source of that particular information is not the journalist reporting the news.

It is important to bear in mind, of course, that the press release is but one part of information subsidies, and but one medium through which the public relations industry contributes to the agenda building process. My findings can by no means be construed to imply that the PR industry lacks influence. Information subsidies are only a part of the

multifaceted communication processing between the public relations industry and the public – a part that does not include the countless messages and communication that occurs directly between the organization and its numerous publics without the gatekeeping function which has been provided by news media for many decades – these techniques have become especially prevalent in the age of the Internet when gatekeepers are no longer critical to the process of mass communication.

Taking into consideration the primary limitation of this study – i.e., reliance on verbatim text matches – two different conclusions can be made. On the one hand, it may be reasonable to suggest that, as indicated by the results of computing verbatim text matches, the use of press release content in news media has diminished. On the other hand, it may be suggested that the bulk of press release usage by news media remained undetected – provided all content borrowed from press releases appeared in paraphrased form, with no verbatim matches present. This is a troubling proposition, for paraphrased text makes the source of the news less evident – which makes it all the more critical to further investigate this issue, paying considerable attention to analysis of paraphrased text.

## When Press Release Content Becomes an Issue

Measuring the extent to which news media depend on press release content, clearly, requires adequate sampling techniques, ensuring that the results of the study can be generalized beyond the sample data. However, regardless of how often or how rarely news articles use press releases as a source, regardless of how much or that text is used verbatim, even one such instance may be an issue.

However, regardless of how often or how rarely news articles use press releases as a source, regardless of how much or that text is used verbatim, even one instance may be an issue. In the process of working on my data, specifically, while writing code to catch bad discriminators (i.e., instances of press release text matching a text in a news article, but appearing in more than one press release), I came across several instances of text which would not have raised any red flags if considered separately from the rest of the data. However, data tells stories.

Among the instances of text used in multiple press releases, most were expected – such as boiler plate text or names of executives and their titles, etc. There was less formulaic text, but most of was neutral in tone and used in similar contexts – i.e., store openings or product and service descriptions:

> …although grants are unrestricted colleges and universities are encouraged to designate a portion to math and science programs… (Exxon, 2012b)

> …the clinics are staffed by nurse practitioners who provide treatment for common family illnesses and administer wellness and prevention services… (CVS, 2012)

Some of these instances were quotations by company executives, which, again, was expected: as I have mentioned in previous chapters, research has shown that press releases often include quotations from executives, which are then incorporated into news articles based on those press releases. However, research has also shown that such quotations, although appearing to be genuine statements by company executes, are often "preformulated" statements – i.e., generic statements prepared by the public relations staff. With a neutral statement conveying a fact that seems to be not much of a problem. However, I found a few cases when the text was subjective – i.e., it conveyed the

speaker's opinion or emotion related to whatever was the subject of the press release. Consider the following examples:

> ...ahead of our customers' needs by strengthening the network that links them to businesses and economies large and small," said Raj Subramaniam, senior vice president of FedEx global marketing and customer experience... (Fedex, 2012)

> …while gratified that we have been able to provide this level of liquidity to the market we recognize that the… (Fannie Mae, 2012b)

Typically, a genuine statement made by a person, containing some form of subjectivity will be less formulaic than a product description and, most likely, will not be repeated again and again, at least not word-for-word . Thus, it is surprising that Fannie Mae executives feel "gratified" exactly the same way on multiple occasions. If a reader were handed two different news articles containing identical statements expressing emotion in the same words, but within different contexts, would that reader trust the executive who made those statements? Would the reader trust the paper?

**Wells Fargo PR: Evidence of Influence**

In most cases, such preformulated statements are made by company executives. However, there are exceptions, and one of them comes from Wells Fargo's PR team. The following text appeared in a press release from September 26, 2012:

> "Though our County was hit hard by the downturn in the housing market, we are slowly rebounding," said ***Prince George's County Executive Rushern L. Baker, III. "Prince George's County*** has a great inventory of homes and attractive amenities for prospective homeowners in the ***Washington*** region. These grants will help people get over the tremendous financial hurdle of finding funds for a down payment and position them to make one of the largest investments of their lives." (Wells Fargo, 2012a)

Now consider the following text which appeared in a press release two months later, on November 27, 2012:

> "Though our county was hit hard by the downturn in the housing market, we are slowly rebounding," said ***Gilda Gonzales, CEO at The Unity Council. "Alameda and Contra Costa*** counties have a great inventory of homes and attractive amenities for prospective homeowners in the ***East Bay*** region. These grants will help people get over the tremendous financial hurdle of finding funds for a down payment and position them to make one of the largest investments of their lives." (Wells Fargo, 2012b)

The press releases are two months apart; they are very similar in structure and content. The statements are identical, except one refers to Alameda and Contra Costa counties which are located in the Bay Area, the other one refers to the Washington DC region. Clearly, this is an example of preformulated statements. However, what sets them apart from other examples is that the first one is made by the chief executive officer of the Unity Council located in Oakland, CA; whereas the other one – by Prince George's County Executive – neither of whom work for the company which issued the press releases containing both quotes. In other words, Wells Fargo's PR team wrote a generic statement for administrators who are not part of the company, yet were somehow convinced to make those statements, thus, giving these statements authenticity and impartiality.

One would think that this issue concerns the public relations industry, but not the media: after all, doesn't the media verify its sources, wouldn't a journalist check whether a statement appearing in a press release is genuine before using it in an article? Apparently not – as evidenced by at least one article, in the *San Jose Mercury New*s and the *Contra Costa Times,* published promptly in both newspapers on the same day as the corresponding press release was issued:

> "Alameda and Contra Costa counties have a great inventory of homes and attractive amenities for prospective homeowners in the East Bay, "Gilda Gonzales, chief executive officer of the Unity Council, said in a prepared release. "These grants will help people get over the tremendous financial hurdle of finding

> funds for a down payment." (*San Jose Mercury New*s, 2012; *Contra Costa Times*, 2012)

One of the many authors offering advice on public relations once suggested:

> [A press release] enables you to tell a story from your perspective, in a larger context that in all likelihood is alien to even your beat reporters. Of course, you have to disguise it as news, which refocuses on the issue of real news writing versus shameless self-promotion (Williams, 1994, p. 7).

That author could take some lessons from Wells Fargo: there's no need to disguise a corporate story as news if you convince an elected official to tell it for you.

## Concluding Remarks

### Contributions

In this dissertation, I have demonstrated how a variety of computational approaches can be applied to investigating an important problem in journalism studies. The study's contributions are two-fold: (1) it contributes to a better understanding of the nature of the complex relationship between news media and public relations; (2) it offers techniques for strengthening journalism research methodology, exploring an application of computational methods to data collection, processing, and analysis. More specifically, I use computation to automatically extract large amounts of text from corporate web sites, transform loosely structured text into well-formatted data, and reduce a very large data set to a sample of most relevant items suitable for both automated and manual textual analysis. Through conducting such analysis, I investigate the extent to which press release content is used by news media verbatim, how such content is used and whether its positive tone affects the overall tone of the news article, and whether proper attribution is made identifying the true source of the news.

My findings are limited due to the method's reliance on verbatim text matches between press releases and news articles. As a result, I cannot make definitive conclusions about the proportion of press release content in news media: despite the fact that my results, based on verbatim matches, suggest that this problem might be not as severe as presented in previous research, press release content copied into news articles verbatim might be the tip of the iceberg. Thus, I offer two possible explanations of this data: it might be that the use of press release content in news media may have, indeed, diminished; but it is also possible that most of this practice remains undetected, with all content borrowed from press releases appearing in news media in paraphrased form.

However, the central argument of this preliminary study was methodological; therefore, the main significance of the conducted data analysis is that it explores a new methodology. Most importantly, by using computation to collect, and then reduce a very large data set to a sample of most relevant items – a sample small enough to be used for in-depth textual analysis – this new approach puts such data within the grasp of anyone equipped with traditional quantitative and qualitative methods used in journalism and mass communication research.

Applying my method to this kind of data, I was able to discover a "smoking gun" – a striking example of PR influence – i.e., Wells Fargo "manufacturing" statements and getting elected officials to repeat them, with the media reporting them as a regular news story. The Wells Fargo case clearly calls for investigation: how frequently do corporate public relations put words into the mouth of public officials? Locating this particular example was made possible through the interdisciplinary methodological approach to journalism research proposed in this dissertation. And although this dissertation is not –

and cannot be – a practical guide to computer science applied to problems in journalism and mass communication research, it provides an example of how such collaboration can lead to new knowledge, which otherwise would have remained undiscovered.

**Further Research**

The methodology explored in this dissertation offers numerous possibilities for further research, extending the current study, and addressing other issues in journalism and mass communication that require locating, processing, and analyzing very large amounts of text data. The topic of this dissertation may be extended in many ways as well; the several directions I mention are but one possibility (see Figure 11).

One way to address the sampling limitations of this study is to extend the selection of news media to (a) publications not accessible through LexisNexis Academic and (b) publications other than newspapers. At the same time, it would be interesting to extend the selection of corporations beyond the top tier belonging to the Fortune-100 (or even Fortune-500) list. In addition, a study might take a closer look at the press releases listed on corporate web sites: these may or may not be the same press releases submitted by those corporations to news media.

Broadening the domain of public relations sources beyond corporations is another way to extend this study. As I have mentioned in Chapter Five, it is not unreasonable to expect that government press releases will be used by news media much more often than press releases from a nonprofit; and that press releases from a university will be treated differently than press releases from a business. With a combined data set containing press releases and news coverage representing such different types of organizations, the

differences between these types would be the focus of such a study. The same is true of news media: a comparison of different publications or types of publications with regard to their reliance on the press release would make an excellent research project.

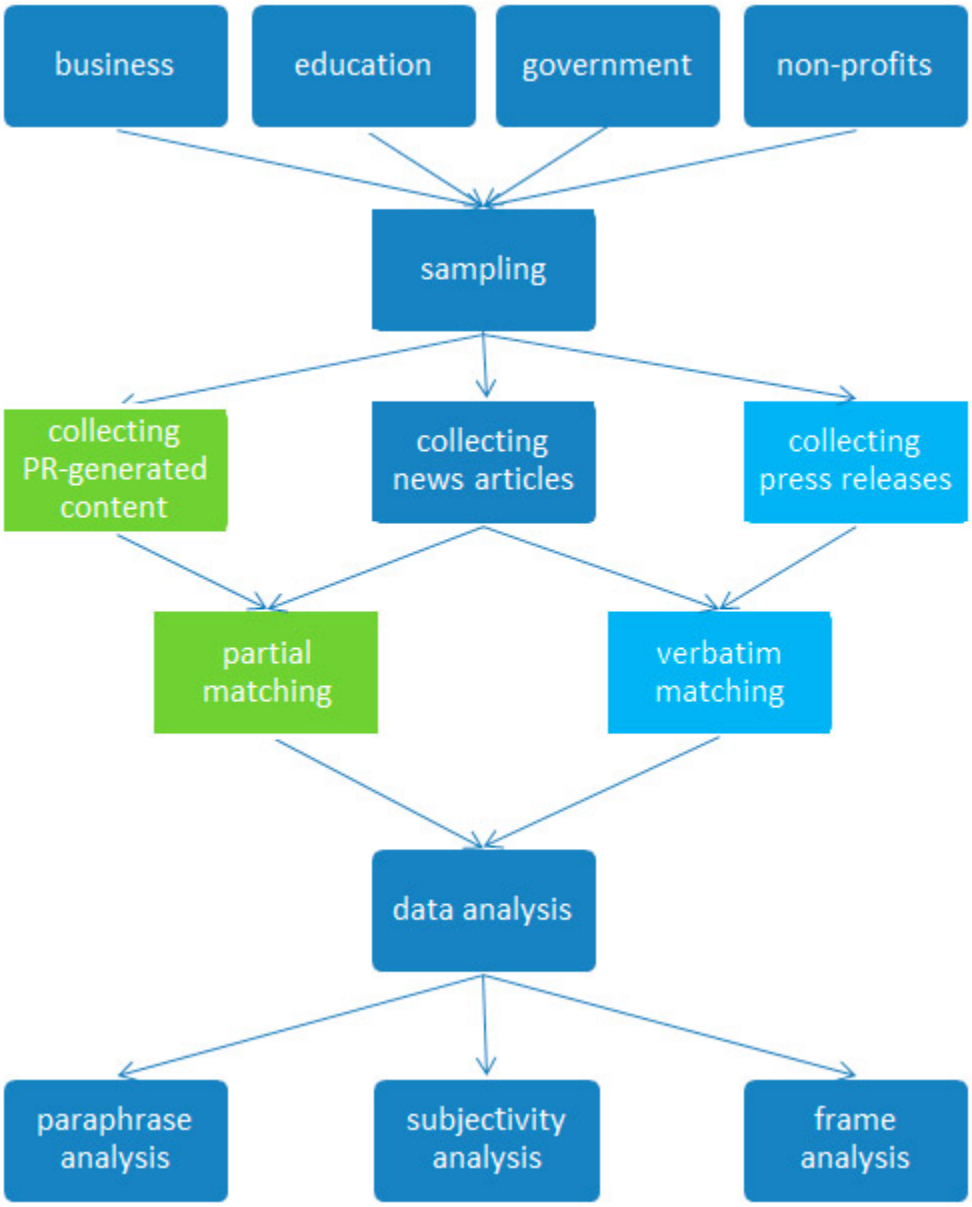



*Figure 11.* Further research.

Limiting a study to press release content is convenient. As I have argued in this dissertation, a focus on press release content justifies the limitation imposed by using verbatim text matches to tie one text to another. However, as I have also mentioned, this limits the type of questions one might ask. Investigating the degree to which press release content is used by news media is a useful exercise; however, a more important question would be about PR-generated content in general. After all, in today's media environment, a press release, clearly, is but one channel used by the public relations industry to communicate its messages to the media and the public. Thus, one might take a broader approach to the concept of information subsidies and examine other types of content offered by organizations to news media in an effort to gain favorable news coverage – such as video news releases, or the content offered through corporate websites and social media.

If the data is extended beyond press releases, limiting the method to verbatim text matching becomes no longer justified. In this case, the researcher, most likely, won't care whether the source of the message is a press release, a speech, an interview, or a corporate message posted to a social media site. Therefore, computational analysis of paraphrased text should be utilized. Researchers in computational linguistics have developed methods to automatically acquire collections of paraphrased text and then use those collections in the context of evaluating machine translation systems. One such example is TER-Plus: the system aligns words in the automatically generated translation of a given text with the words in the "gold standard" reference provided by a human "not only when [these words] are exact matches but also when the words share a stem or are synonyms." In addition, the system uses "probabilistic phrasal substitutions to align

phrases" (Snover, Madnani, Dorr, & Schwartz, 2009, p. 261). The authors distinguish between several different types of paraphrases based on paraphrase quality: "lexical paraphrases" (which differ by one word at most), "morphological variants" (phrases which differ in morphological form), "approximate phrasal paraphrases" (phrases that share partial semantic content), and "phrasal paraphrases" (i.e., true paraphrases) (pp. 264-265). Clearly, this approach, borrowed from machine translation, can be applied to aligning phrases in a PR-generated text with phrases in a news article. Furthermore, paraphrase analysis can be used to pose research questions similar to RQ1, RQ2 and RQ5, while offering a more in-depth analysis of text similarities and, through that, the potential to draw broader conclusions. (For more details on paraphrase usage in the context of machine translation, see Snover, Madnani, Dorr, & Schwartz, 2009; and Zhou, Lin, & Hovy, 2006).

Another possible extension of this study can address automated analysis of subjective language explored in the context of RQ3 and RQ4. It has been demonstrated that the automated coding approach attempted in this study cannot be used as a standalone method: the prior polarity of a word does not determine its contextual polarity. However, using a subjectivity lexicon to measure the prior polarity of words becomes a valid first step when used in conjunction with a text classifier – a program that will use the results of such coding as one of its inputs and will subsequently learn the statistical patterns in the data, with the help of training data supplied by human coders. The learned statistical patterns can reliably estimate the contextual polarity of the sentence, which can be then used to compute the overall subjectivity of the text. (For a detailed overview of subjectivity and sentiment analysis, see Liu, 2010; and Pang & Lee, 2008).

Furthermore, text classification can be used in the context of framing analysis. Robert M. Entman (1991) pointed out that news frames "can be detected by probing for particular words … that consistently appear in a narrative and convey thematically consonant meanings" (p. 7). Such "probing for particular words" can be automated, with the resulting data being used to develop statistical models of news frames. And although computation cannot substitute "careful and close reading of texts, [such methods can be] thought of as amplifying and augmenting careful reading and thoughtful analysis" (Grimmer & Stewart, 2013, p. 2): after all, unlike in traditional framing analysis, the constructed statistical models can be applied to very large data sets. (For an in-depth review of text classification used in the context of political communication, see Grimmer & Stewart, 2013).

Clearly, the possibilities of improving and extending the presented study methodologically are boundless. However, the study's methodology is not the only direction for further research. The issue of press release influence on news media can be investigated from different perspectives. One particularly interesting approach would be a comparative examination of news reliance on the press release spanning several decades. It might be expected that the influence of the press release has diminished; the question is whether it may be ascribed to the rapid development of online communication in the last decade, or whether such a decline had started earlier.

Finally, despite the attraction of advanced computational methods, one of the most exciting possibilities would be to use the proposed methodology as a foundation for developing more specific, narrowly-oriented methods with the goal of locating cases of a particular type in a large data set. This dissertation has shown that even one case,

regardless of being not statistically significant, may be of considerable value to the field of journalism studies. Wells Fargo preparing statements for elected officials was discovered by chance. Designing a study to look for such cases would require the very basic computational approaches; yet, despite its methodological simplicity, such a study could make a major contribution to our understanding of the impact of public relations on publics and journalism.

APPENDIX A.
SAMPLING FORTUNE-100 COMPANIES

Table A1. *Fortune-100 companies.*

| Rank | Code[a] | Name |
|---|---|---|
| 1 | 0 | Exxon Mobil |
| 2 | -1 | Wal-Mart Stores |
| 3 | 6 | Chevron |
| 4 | -1 | ConocoPhillips |
| 5 | 1 | General Motors |
| 6 | -1 | General Electric |
| 7 | 4 | Berkshire Hathaway |
| 8 | 0 | Fannie Mae |
| 9 | 0 | Ford Motor |
| 10 | -1 | Hewlett-Packard |
| 11 | 1 | AT&T |
| 12 | 4 | Valero Energy |
| 13 | 0 | Bank of America Corp. |
| 14 | 0 | McKesson |
| 15 | 6 | Verizon Communications |
| 16 | 0 | J.P. Morgan Chase & Co. |
| 17 | 0 | Apple |
| 18 | 0 | CVS Caremark |
| 19 | 0 | International Business Machines |
| 20 | 0 | Citigroup |
| 21 | -1 | Cardinal Health |
| 22 | -1 | UnitedHealth Group |
| 23 | 0 | Kroger |
| 24 | -1 | Costco Wholesale |
| 25 | 1 | Freddie Mac |
| 26 | 0 | Wells Fargo |
| 27 | -1 | Procter & Gamble |
| 28 | 0 | Archer Daniels Midland |
| 29 | 7 | AmerisourceBergen |
| 30 | 3 | INTL FCStone |
| 31 | -1 | Marathon Petroleum |
| 32 | 6 | Walgreen |

(continued)

| | | |
|---|---|---|
| 33 | -1 | American International Group |
| 34 | 0 | MetLife |
| 35 | 0 | Home Depot |
| 36 | 1 | Medco Health Solutions |
| 37 | 0 | Microsoft |
| 38 | 0 | Target |
| 39 | 0 | Boeing |
| 40 | -1 | Pfizer |
| 41 | 0 | PepsiCo |
| 42 | -1 | Johnson & Johnson |
| 43 | 0 | State Farm Insurance Cos. |
| 44 | -1 | Dell |
| 45 | 0 | WellPoint |
| 46 | -1 | Caterpillar |
| 47 | -1 | Dow Chemical |
| 48 | -1 | United Technologies |
| 49 | 0 | Comcast |
| 50 | -1 | Kraft Foods |
| 51 | 6 | Intel |
| 52 | -1 | United Parcel Service |
| 53 | 6 | Best Buy |
| 54 | -1 | Lowe's |
| 55 | -1 | Prudential Financial |
| 56 | 0 | Amazon.com |
| 57 | 0 | Merck |
| 58 | 0 | Lockheed Martin |
| 59 | 6 | Coca-Cola |
| 60 | -1 | Express Scripts Holding |
| 61 | 0 | Sunoco |
| 62 | 3 | Enterprise Products Partners |
| 63 | 0 | Safeway |
| 64 | -1 | Cisco Systems |
| 65 | -1 | Sears Holdings |
| 66 | -1 | Walt Disney |
| 67 | 0 | Johnson Controls |
| 68 | 0 | Morgan Stanley |

(continued)

| | | |
|---|---|---|
| 69 | 4 | Sysco |
| 70 | 0 | FedEx |
| 71 | 0 | Abbott Laboratories |
| 72 | 1 | DuPont |
| 73 | 5 | Google |
| 74 | -1 | Hess |
| 75 | -1 | Supervalu |
| 76 | 0 | United Continental Holdings |
| 77 | 3 | Honeywell International |
| 78 | 3 | CHS |
| 79 | 6 | Humana |
| 80 | -1 | Goldman Sachs Group |
| 81 | 3 | Ingram Micro |
| 82 | 6 | Oracle |
| 83 | 0 | Delta Air Lines |
| 84 | 0 | Liberty Mutual Insurance Group |
| 85 | 7 | World Fuel Services |
| 86 | 0 | New York Life Insurance |
| 87 | 3 | Plains All American Pipeline |
| 88 | 4 | TIAA-CREF |
| 89 | 0 | Aetna |
| 90 | 0 | Sprint Nextel |
| 91 | -1 | News Corp. |
| 92 | -1 | General Dynamics |
| 93 | 0 | Allstate |
| 94 | -1 | HCA Holdings |
| 95 | 0 | American Express |
| 96 | 4 | Tyson Foods |
| 97 | 0 | Deere |
| 98 | 7 | Murphy Oil |
| 99 | -1 | Philip Morris International |
| 100 | 2 | Nationwide |

[a]The codes represent reasons for excluding a company from the sample. Code *zero* marks selected companies.

-1 Not picked during final sampling
0 Used in final sample
1 LexisNexis technical issue
2 Other sampling issues
3 Less than 100 news articles
4 Less than 30 press releases
5 No traditional press releases
6 Impossible to retrieve all press releases for 2012
7 Both codes apply: 3 and 4

APPENDIX B.
FORTUNE-100 CORPORATE NEWS WEBSITES

Table B1. *Fortune-100 corporate news websites.*

1. Exxon Mobil
   http://news.exxonmobil.com/

2. Wal-Mart Stores
   http://news.walmart.com/news-archive/

3. Chevron
   http://www.chevron.com/news/press/

4. ConocoPhillips
   http://www.conocophillips.com/EN/newsroom/news_releases/2012NewsReleases/Pages/index.aspx

5. General Motors
   http://media.gm.com/media/us/en/gm/news/news_archive.html

6. General Electric
   http://www.genewscenter.com/content/default.aspx?newsareaid=2

7. Berkshire Hathaway
   http://www.berkshirehathaway.com/news/2012news.html

8. Fannie Mae
   http://www.fanniemae.com/portal/jsp/filter-media.html?year=year&month=month&keyword=Keyword&topic=all_news_categories&financial_news=all_news_subcategories

9. Ford Motor
   http://corporate.ford.com/news-center

10. Hewlett-Packard
    http://www8.hp.com/us/en/hp-news/newsroom.html

11. AT&T
    http://www.att.com/gen/press-room?pid=4800&cdvn=news&newsfunction=searchresults&beginning_month=-2&beginning_year=2013&ending_month=0&ending_year=2013

(continued)

12. Valero Energy
    http://www.valero.com/newsroom/Pages/Home.aspx

13. Bank of America Corp.
    http://newsroom.bankofamerica.com/advsearch?date_from=2012-01-01T00%3A00%3A00Z&date_to=2013-01-01T00%3A00%3A00Z&year=2012
14. McKesson
    http://www.mckesson.com/en_us/McKesson.com/About%2BUs/Newsroom/Press%2BReleases%2BArchives/Press%2BReleases%2BArchives.html

15. Verizon Communications
    http://newscenter.verizon.com/corporate/news-articles/

16. J.P. Morgan Chase & Co.
    http://investor.shareholder.com/jpmorganchase/releases.cfm?NavSection=

17. Apple
    http://www.apple.com/pr/library/2012/

18. CVS Caremark
    http://info.cvscaremark.com/newsroom/press-releases

19. International Business Machines
    http://www-03.ibm.com/press/us/en/pressreleases/finder.wss

20. Citigroup
    http://www.citigroup.com/citi/news/news_list_view.html

21. Cardinal Health
    http://cardinalhealth.mediaroom.com/index.php?year=2012&s=news_releases

22. UnitedHealth Group
    http://www.unitedhealthgroup.com/main/Newsroom.aspx

23. Kroger
    http://ir.kroger.com/phoenix.zhtml?c=106409&p=irol-news&nyo=0

24. Costco Wholesale
    http://phx.corporate-ir.net/phoenix.zhtml?c=83830&p=irol-news

(continued)

25. Freddie Mac
    http://freddiemac.mwnewsroom.com/

26. Wells Fargo
    https://www.wellsfargo.com/press/?year=2012

27. Procter & Gamble
    http://news.pg.com/news_releases

28. Archer Daniels Midland
    http://www.adm.com/en-US/news/_layouts/PressReleaseList.aspx?more=true

29. AmerisourceBergen
    http://www.amerisourcebergen.com/investor/phoenix.zhtml?c=61181&p=irol-newsarch&nyo=1

30. INTL FCStone
    http://ir.intlfcstone.com/releases.cfm

31. Marathon Petroleum
    http://www.marathonpetroleum.com/News/News_Releases/#?year=2012&month=0

32. Walgreen
    http://news.walgreens.com/section_display.cfm?section_id=1&page_flag=news

33. American International Group
    http://www.aig.com/press-releases_3171_438003.html

34. MetLife
    https://www.metlife.com/about/press-room/us-press-releases/index.html?WT.ac=GN_about_press-room_us-press-releases

35. Home Depot
    http://phx.corporate-ir.net/preview/phoenix.zhtml?c=63646&p=irol-IRHome

36. Medco Health Solutions
    http://phx.corporate-ir.net/phoenix.zhtml?c=69641&p=irol-MedcoPress&nyo=1

(continued)

37. Microsoft
    http://www.microsoft.com/en-us/news/press/NewsArchive.aspx?feedid=PressReleases

38. Target
    http://pressroom.target.com/news

39. Boeing
    http://boeing.mediaroom.com/

40. Pfizer
    http://www.pfizer.com/news/press_releases/pfizer_press_release_archive.jsp

41. PepsiCo
    http://www.pepsico.com/Media/Press-Releases.html#search

42. Johnson & Johnson
    http://www.jnj.com/connect/news/all

43. State Farm Insurance Cos.
    http://www.statefarm.com/aboutus/newsroom/pressreleases/pressreleases.asp

44. Dell
    http://content.dell.com/us/en/corp/newsroom-press-releases.aspx?c=us&l=en&s=corp&~ck=mn

45. WellPoint
    http://ir.wellpoint.com/phoenix.zhtml?c=130104&p=irol-news&nyo=1

46. Caterpillar
    http://www.cat.com/news-and-events/machine-and-engine-press-releases

47. Dow Chemical
    http://www.dow.com/news/press-releases/

48. United Technologies
    http://www.utc.com/News

49. Comcast
    http://corporate.comcast.com/news-information/news-feed

(continued)

50. Kraft Foods
    http://ir.kraftfoodsgroup.com/releases.cfm

51. Intel
    http://newsroom.intel.com/community/intel_newsroom/

52. United Parcel Service
    http://www.pressroom.ups.com/?WT.svl=Footer

53. Best Buy
    http://pr.bby.com/

54. Lowe's
    http://media.lowes.com/

55. Prudential Financial
    http://www.news.prudential.com/

56. Amazon.com
    http://phx.corporate-ir.net/phoenix.zhtml?c=176060&p=irol-mediaHome

57. Merck
    http://www.mercknewsroom.com/

58. Lockheed Martin
    http://www.lockheedmartin.com/us/news/press-releases.html

59. Coca-Cola
    http://www.coca-colacompany.com/media-center/

60. Express Scripts Holding
    http://www.express-scripts.com/pressroom/

61. Sunoco
    https://www.piersystem.com/go/doc/3433/1322919/

62. Enterprise Products Partners
    http://phx.corporate-ir.net/phoenix.zhtml?c=80547&p=irol-newsarch&nyo=1

63. Safeway
    http://www.safeway.com/ShopStores/Investors.page?#iframetop

(continued)

64. Cisco Systems
http://newsroom.cisco.com/

65. Sears Holdings
http://www.searsmedia.com/

66. Walt Disney
http://thewaltdisneycompany.com/disney-news

67. Johnson Controls
http://www.johnsoncontrols.com/content/us/en/news.html

68. Morgan Stanley
http://www.morganstanley.com/about/newsroom.html

69. Sysco
http://www.sysco.com/investor/pressreleases.html

70. FedEx
http://news.van.fedex.com/

71. Abbott Laboratories
http://www.abbott.com/news-media/news-index.htm

72. DuPont
http://www2.dupont.com/media/en-us/news-events/index.html

73. Google
https://www.google.com/intl/en/press/

74. Hess
http://phx.corporate-ir.net/phoenix.zhtml?c=101801&p=irol-news&nyo=1

75. Supervalu
http://www.supervaluinvestors.com/phoenix.zhtml?c=93272&p=irol-news&nyo=0

76. United Continental Holdings
http://www.unitedcontinentalholdings.com/index.php?section=media

(continued)

77. Honeywell International
http://honeywell.com/News/Pages/press-releases.aspx

78. CHS
http://www.chs.net/news/index.html

79. Humana
http://www.humana.com/resources/about/news/

80. Goldman Sachs Group
http://www.goldmansachs.com/media-relations/index.html

81. Ingram Micro
http://phx.corporate-ir.net/phoenix.zhtml?c=98566&p=irol-news&nyo=1

82. Oracle
http://www.oracle.com/us/corporate/press/index.html

83. Delta Air Lines
http://news.delta.com/index.php?s=43

84. Liberty Mutual Insurance Group
http://www.libertymutualgroup.com/liberty-mutual-news
http://www.libertymutualgroup.com/omapps/ContentServer?fid=1142008723439&ln=en&pagename=LMGroup%2FViews%2FLMG&ft=9&type=pr&cid=1142008723439&yr=2012

85. World Fuel Services
http://phx.corporate-ir.net/phoenix.zhtml?c=101792&p=irol-news&nyo=0

86. New York Life Insurance
http://www.newyorklife.com/about/news-room

87. Plains All American Pipeline
http://ir.paalp.com/News_Releases

88. TIAA-CREF
https://www.tiaa-cref.org/public/about-us/news-center

89. Aetna
http://newshub.aetna.com/

(continued)

90. Sprint Nextel
http://newsroom.sprint.com/

91. News Corp.
http://www.newscorp.com/news/index.html

92. General Dynamics
http://www.generaldynamics.com/news/

93. Allstate
http://www.allstatenewsroom.com/

94. HCA Holdings
http://phx.corporate-ir.net/phoenix.zhtml?c=63489&p=irol-news

95. American Express
http://about.americanexpress.com/news/?inav=about_news

96. Tyson Foods
http://www.tysonfoods.com/Media-Room.aspx

97. Deere
http://www.deere.com/wps/dcom/en_US/corporate/our_company/news_and_media/news_and_media.page?

98. Murphy Oil
http://www.murphyoilcorp.com/ir/news.aspx?Year=2012

99. Philip Morris International
http://www.pmi.com/eng/media_center/press_releases/pages/press_releases.aspx

100. Nationwide
http://www.nationwide.com/newsroom/newsroom.jsp

APPENDIX C.
PUBLICATIONS USED IN THE TEXT CORPUS

Table C1. *Publications used in the text corpus.*

1. Aberdeen American News (South Dakota)
2. Advertising Age
3. Agweek
4. AIM Jefferson (Morris, North Jersey)
5. AIM Vernon (Sussex, North Jersey)
6. AIM West Milford (Passaic, North Jersey)
7. Alamogordo Daily News (New Mexico)
8. American Banker
9. Argus (Cumberland, North Jersey)
10. Austin American-Statesman (Texas)
11. Automotive News
12. Bangor Daily News (Maine)
13. Belleville Times (Essex, North Jersey)
14. Bloomfield Life (Essex, North Jersey)
15. Bogota Bulletin (North Jersey)
16. Brattleboro Reformer (Vermont)
17. Buffalo News (New York)
18. Burnet Bulletin (Texas)
19. Business Insurance
20. Cape Gazette (Lewes, Delaware)
21. Carlsbad Current-Argus (New Mexico)
22. Chapel Hill Herald (Durham, N.C.)
23. Charleston Daily Mail (West Virginia)
24. Cheney Free Press (Washington)
25. Chicago Daily Herald
26. Chico Enterprise-Record (California)
27. CityBusiness North Shore Report (New Orleans, LA)
28. Cliffside Park Citizen (Bergen, North Jersey)
29. Clifton Journal (Passaic, North Jersey)
30. Colorado Springs Business Journal (Colorado Springs, CO)
31. Community News (Bergen, North Jersey)
32. Connecticut Post (Bridgeport)
33. Contra Costa Times (California)
34. Crain's Cleveland Business
35. Crain's Detroit Business

(continued)

36. Daily Camera (Boulder, Colorado)
37. Daily Deal/The Deal
38. Daily Journal of Commerce (Portland, OR)
39. Daily News (New York)
40. Daily Press - Victorville (California)
41. Daily Variety
42. Dayton Daily News (Ohio)
43. Deming Headlight (New Mexico)
44. Deseret Morning News (Salt Lake City)
45. Detroit Free Press (Michigan)
46. East Bernard Express (Texas)
47. Edgewater View (Bergen, North Jersey)
48. Edmonds Beacon (Washington)
49. Education Week
50. Eureka Times Standard (California)
51. Finance & Commerce (Minneapolis, MN)
52. Florida Times-Union (Jacksonville)
53. Fort Lee Suburbanite (Bergen, North Jersey)
54. Franklin Lakes-Oakland Suburban News (Bergen, North Jersey)
55. Glen Rock Gazette (Bergen, North Jersey)
56. Government Technology
57. Hackensack Chronicle (Bergen, North Jersey)
58. Hartford Courant (Connecticut)
59. Hays Free Press (Buda, Texas)
60. Herald News (Passaic County, NJ)
61. Idaho Falls Post Register (Idaho)
62. Inland Valley Daily Bulletin (Ontario, CA)
63. Intelligencer Journal/New Era (Lancaster, Pennsylvania)
64. Investment News
65. Investor's Business Daily
66. Journal of Commerce Online
67. Journal Record Legislative Report (Oklahoma City, OK)
68. Las Cruces Sun-News (New Mexico)
69. Las Vegas Review-Journal (Nevada)
70. Lawyers Weekly USA
71. Leonia Life (Bergen, North Jersey)
72. Lewiston Morning Tribune (Idaho)
73. Little Ferry Local (Bergen, North Jersey)
74. Long Island Business (Long Island, NY)
75. Los Angeles Times
76. Lowell Sun (Massachusetts)

(continued)

77. Mahwah Suburban News (Bergen, North Jersey)
78. Marin Independent Journal (California)
79. Maryland Gazette
80. Massachusetts Lawyers Weekly
81. McKenzie River Reflections (McKenzie Bridge, Oregon)
82. Messenger-Inquirer (Owensboro, Kentucky)
83. Metropolitan Corporate Counsel
84. Michigan Lawyers Weekly
85. Mississippi Business Journal (Jackson, MS)
86. Missouri Lawyers Media
87. Monroe County Appeal (Paris, Missouri)
88. Montclair Times (Essex, North Jersey)
89. Monterey County Herald (California)
90. Morning Call (Allentown, Pennsylvania)
91. Mukilteo Beacon (Washington)
92. New Orleans City Business (New Orleans, LA)
93. Newsday (New York)
94. New York Observer
95. North Carolina Lawyers Weekly
96. North County Times (Escondido, California)
97. Northern Valley Suburbanite (Bergen, North Jersey)
98. Nutley Sun (Essex, North Jersey)
99. Orange County Register (California)
100. Oroville Mercury Register (California)
101. Palm Beach Post (Florida)
102. Parsippany Life (Morris, North Jersey)
103. Pasadena Star-News (California)
104. Pascack Valley Community Life (Bergen, North Jersey)
105. Passaic Valley Today (North Jersey)
106. Pensions & Investments
107. Pittsburgh Post-Gazette
108. Pittsburgh Tribune Review
109. Plastics News
110. Providence Journal
111. Public Opinion (Chambersburg, Pennsylvania)
112. Ramsey Suburban News (Bergen, North Jersey)
113. Rhode Island Lawyers Weekly
114. Richmond Times-Dispatch (Virginia)
115. Ridgefield Park Patriot (Bergen, North Jersey)
116. Roll Call
117. Rubber & Plastics News

(continued)

118. Ruidoso News (New Mexico)
119. San Antonio Express-News
120. San Bernardino Sun (California)
121. San Gabriel Valley Tribune (California)
122. San Jose Mercury News (California)
123. Sarasota Herald Tribune (Florida)
124. Sentinel & Enterprise (Fitchburg, Massachusetts)
125. Shelton-Mason County Journal (Washington)
126. Small Business column
127. South Bend Tribune (Indiana)
128. South Bergenite (Bergen, North Jersey)
129. South Carolina Lawyers Weekly
130. South Florida Sun-Sentinel (Fort Lauderdale)
131. Spokesman Review (Spokane, WA)
132. Springfield News-Sun, Ohio
133. Star-News (Wilmington, NC)
134. Star Tribune (Minneapolis, MN)
135. St. Louis Post-Dispatch (Missouri)
136. St. Paul Pioneer Press (Minnesota)
137. Suburban News, A Publication of the Ridgewood News (Bergen, North Jersey)
138. Suburban Trends (Morris, North Jersey)
139. Sunday News (Lancaster, Pennsylvania)
140. SUNDAY TELEGRAM (Massachusetts)
141. Tahlequah Daily Press (Oklahoma)
142. Tampa Bay Times
143. Tampa Tribune (Florida)
144. Teaneck Suburbanite (Bergen, North Jersey)
145. TELEGRAM & GAZETTE (Massachusetts)
146. Telegraph Herald (Dubuque, IA)
147. The Arizona Capitol Times
148. The Atlanta Journal-Constitution
149. The Augusta Chronicle (Georgia)
150. The Bakersfield Californian
151. The Baltimore Sun
152. The Berkshire Eagle (Pittsfield, Massachusetts)
153. The Bismarck Tribune
154. The Bond Buyer
155. The Capital (Annapolis, MD)
156. The Capital Times (Madison, Wisconsin)
157. The Cheraw Chronicle (South Carolina)

(continued)

158. The Christian Science Monitor
159. The Chronicle of Higher Education
160. The Chronicle of Philanthropy
161. The Columbian (Vancouver, Washington)
162. The Daily News of Los Angeles
163. The Daily Oklahoman (Oklahoma City, OK)
164. The Daily Record (Baltimore, MD)
165. The Daily Record of Rochester (Rochester, NY)
166. The Daily Reporter (Milwaukee, WI)
167. The Daily Star-Journal, Warrensburg, Mo
168. The Dallas Morning News
169. The Deal Pipeline
170. The Denver Post
171. The Detroit News (Michigan)
172. The Dispatch (Gilroy, California)
173. The Evening Sun (Hanover, Pennsylvania)
174. The Forward
175. The Gazette (Cedar Rapids, Iowa)
176. The Gazette (Fairlawn, North Jersey)
177. The Herald Bulletin (Anderson, Indiana)
178. The Herald Independent, Winnsboro, S.C.
179. The Herald-Palladium, St. Joseph, Mich.
180. The Herald-Sun (Durham, N.C.)
181. The Herald-Tribune (Batesville, Indiana)
182. The Hill
183. The Houston Chronicle
184. The Idaho Business Review (Boise, ID)
185. The Indianapolis Business Journal
186. The Issaquah Press (Washington)
187. The Item of Millburn and Short Hills (Essex, North Jersey)
188. The Journal Record (Oklahoma City, OK)
189. The Lebanon Daily News (Pennsylvania)
190. The Legal Ledger (St. Paul, MN)
191. The Maryland Gazette
192. The Mecklenburg Times (Charlotte, NC)
193. The Minnesota Lawyer (Minneapolis, MN)
194. The Neighbor News (Morris, North Jersey)
195. The New Hampshire Union Leader, Manchester
196. The News-Sentinel (Fort Wayne, Indiana)
197. The New York Post
198. The New York Times

(continued)

199. The Oklahoman (Oklahoma City, OK)
200. The Othello Outlook (Washington)
201. The Pantagraph (Bloomington, Illinois)
202. The Patriot-News (Harrisburg, Pennsylvania)
203. The Philadelphia Daily News
204. The Philadelphia Inquirer
205. The Pueblo Chieftain (Colorado)
206. The Record (Bergen County, NJ)
207. The Record-Eagle (Traverse City, Michigan)
208. The Register Guard (Eugene, Oregon)
209. The Ridgewood News (Bergen, North Jersey)
210. The Roanoke Times (Virginia)
211. The Salt Lake Tribune
212. The San Francisco Chronicle (California)
213. The Santa Fe New Mexican (New Mexico)
214. The State Journal- Register (Springfield, IL)
215. The Taos News (New Mexico)
216. The Times-Union (Albany, NY)
217. The Union Leader (Manchester, NH)
218. The Virginian-Pilot(Norfolk, VA.)
219. The Washington Post
220. The Washington Times
221. The York Dispatch (Pennsylvania)
222. Tire Business
223. Topeka Capital-Journal (Kansas)
224. Town Journal (Bergen, North Jersey)
225. Town News (Bergen, North Jersey)
226. Tribune-Review (Greensburg, PA)
227. Tulsa World (Oklahoma)
228. Twin-Boro News (Bergen, North Jersey)
229. USA TODAY
230. Vallejo Times Herald (California)
231. Variety
232. Verona-Cedar Grove Times (Essex, North Jersey)
233. Virginia Lawyers Weekly
234. Wayne Today (Passaic, North Jersey)
235. Westbrook Sentinel Tribune (Minnesota)
236. Whittier Daily News (California)
237. Wisconsin Law Journal (Milwaukee, WI)
238. Wisconsin State Journal (Madison, Wisconsin)
239. Wyoming Tribune-Eagle (Cheyenne)

APPENDIX D.
SAMPLE CODE

## Description of Provided Code

The following is a selection of files which contain code which was used for various tasks within this study – i.e., locating, collecting, processing, and analyzing data. These files, written in the Python programming language, are only part of the entire system, they cannot be used as a standalone application without the rest of the code, configuration settings and data. Besides, the system was implemented through an incremental process, which included writing code and storing it in files like the ones listed in this appendix, but also manipulating the data through the UNIX shell, executing countless one-time scripts, which were, naturally, not recorded as a file.

However, although the system is not packaged as a piece of software ready to be installed and run, the sample files provided here demonstrate the key implementation decisions, specific to this particular system. They will be helpful to anyone working on a similar project.

## List of Files

pr/crawler.py: collects press releases from corporate web sites.

pr/formatter.py: extracts relevant content from collected press releases and stores it in a structured way for further processing.

pr/sources/abbott.py: implements "abstract" methods, used in crawler.py and formatter.py, specific to a particular company (i.e., Abbott Laboratories). There are 40 such files in the system, one for each corporation in the data set.

news/formatter.py: extracts relevant content from collected news articles and stores it in a structured way for further processing.

shared/data_cleaner.py: provides basic data cleaning functionality (i.e., bringing the data to a common encoding, normalizing special characters and punctuation, etc.).

shared/tokenizer.py: provides access to word and sentence tokenizers (nltk library).

data/articles.py: provides access to news articles data associated with a company, loaded from text files.

data/releases.py: provides access to press release data associated with a company, loaded from text files.

data/scores.py: provides access to a collection of various scores for each company, including counts of sentences automatically labeled as positive, negative, both, or neutral, as well as sentiment scores.

data/subjlexicon.py: provides access to data from the subjectivity lexicon used in the study, loaded from a text file.

data/matches.py: provides access to a collection of matching blocks of text from press releases and news articles.

data/tokens.py: provides access to a tokens collection based on a supplied press release or news article identification number.

analyze/duplicate_finder.py: provides basic duplicate detection among press releases and news articles using the MD5 hashing algorithm.

analyze/block_finder.py: collection of scripts used to print out matching blocks of text based on different criteria.

analyze/match_finder.py: detects matching pairs of press releases and news articles.

analyze/match_writer.py: writes HTML files which display pairs of matching press releases and news articles, visualizing the matching text.

analyze/postag_writer.py: uses the nltk library for part-of-speech tagging.

analyze/sentence_writer.py: writes HTML files which display each press release and news article as lists of sentences. The words matching the subjectivity lexicon used in the study are marked as green for positive and red for negative.

analyze/matrix_maker.py: prints the main data set for further statistical data analysis.

## Source Code

### pr/crawler.py

```
import sys
import os.path
import re
import datetime
from urllib import urlopen
from bs4 import BeautifulSoup
from shared.config import ConfigReader
import source_factory
from shared import common

#retrieves html of all pr (press release) pages for a source and stores them
#for further processing. Needs to be run once per company.

class Crawler(object):

    def __init__(self, company_id, output_path):
        self._company_id = company_id
        self._src = source_factory.get_class(self._company_id)
        self._count = 0
        self._sb = []
        self._load_path(output_path)

    def run(self, istest=False):
        self._istest = istest

        linkpages = self._src.get_linkpages() #get html of all 'links' pages
        count = 0
        for page in linkpages:
            count += 1
            print 'processing page {0} of {1}'.format(count, len(linkpages))
            self._get_html(page)
```

```
        self._write_list()
        print 'EXTRACTED {0} PAGES'.format(self._count)


    def _load_path(self, output_path):
        self._path_dir = os.path.join(output_path, self._company_id)
        if not os.path.exists(self._path_dir):
            os.mkdir(self._path_dir)


    def _write_list(self):
        path = os.path.join(self._path_dir, common.get_list_file_name())
        with open(path, 'w') as f:
            f.write(''.join(self._sb))


    def _get_html(self, page): #get html of each 'pr' page and store it
        soup = BeautifulSoup(page)

        links = soup(self._src.is_link)
        titles = soup(self._src.is_title)
        dates = soup(self._src.is_date)

        if self._istest: #this is for developign a new scraper
            print 'links: ' + str(len(links))
            print 'titles: ' + str(len(titles))
            print 'dates: ' + str(len(dates))

        if not(len(links) == len(titles) == len(dates)):
            raise Exception("number of links/titles/dates not equal")

        count = 0
        for i in range(len(links)):
            count += 1
            if count % 5 == 0:
                print 'processing link {0} of {1}'.format(count, len(links))

            if self._company_id == '23': #special handling: liberty mutual
                if links[i].get('onclick') is None: #ignore pdf links
                    continue
                else:
                    link = self._src.get_link(links[i].get('onclick'))
            else:
                link = self._src.get_link(links[i].get('href'))

            if link is None: #special handling (target)
                continue

            if link.find('External.File') > -1: #ignore links to pdfs
                continue

            title = self._src.get_title(titles[i]).strip().encode('utf-8')
            title = title.replace('\r\n', ' ')
            title = title.replace('\n', ' ')
            title = title.replace('  ', ' ')

            date = self._src.get_date(dates[i])

            if date is None:
                date = datetime.date(2012, 1, 1) #don't forget to check for these when
loading!

            if date > datetime.date(2011,12,31) and date < datetime.date(2013,1,1):
                self._count += 1
                self._sb.append('LOTW-ID: {0}\n'.format(self._count))
                self._sb.append('LOTW-DATE: {0}\n'.format(date.isoformat()))
                self._sb.append('LOTW-TITLE: {0}\n'.format(title))
                self._sb.append('LOTW-URL: {0}\n'.format(link))
                self._sb.append('\n')
```

```
                if not self._istest:
                    self._save_html(link)


    def _save_html(self, link):
        html = urlopen(link).read()
        path = os.path.join(self._path_dir, str(self._count))
        with open(path, 'w') as f:
            f.write(html)
```

## pr/formatter.py

```
import os
import os.path
import sys
import re
from bs4 import BeautifulSoup
from shared import common
from shared.config import ConfigReader
from shared.data_cleaner import DataCleaner
import source_factory

class Formatter(object):

    def __init__(self, company_id, output_path):
        self._company_id = company_id
        self._output_path = output_path
        self._src = source_factory.get_class(self._company_id)
        self._marker = 'LOTW-BR-MARKER'
        self._pattern1 = re.compile('\n\s*')
        self._pattern2 = re.compile('(LOTW-BR-MARKER\s*)+')
        self._pattern3 = re.compile(' [ ]+')
        self._dc = DataCleaner(self._src.get_encoding())

        self._load_replacements()
        self._load_linebreaks()
        self._load_path()
        self._load_list()


    def run(self):
        sb_list = []
        sb_text = []
        counter = 0
        for item in self._list:
            #we don't care about item[0] - it's the id assigned during crawl -
            #   which we will override here even if it's identical
            date        = item[1]
            title       = item[2]
            url         = item[3]
            counter += 1

            path = os.path.join(self._path_dir, str(counter))
            with open(path) as f:
                html = f.readlines()
            body = self._format_body(''.join(html), title)

            #special handling for delta's dates
            if self._company_id == '24':
                date = self._src.get_date_from_body(body)

            sb_list.append('LOTW-ID: {0}\n'.format(counter))
            sb_list.append('LOTW-DATE: {0}\n'.format(date))
            sb_list.append('LOTW-TITLE: {0}\n'.format(title))
            sb_list.append('LOTW-URL: {0}\n\n'.format(url))

            sb_text.append('LOTW-ID: {0}\n'.format(counter))
            sb_text.append('{0}\n\n\n\n'.format(body))

        self._write_files(sb_list, sb_text)


    def _write_files(self, sb_list, sb_text):
        list_file = common.get_list_file_name(self._company_id)
```

```
    path_list = os.path.join(self._output_path, list_file) #refactor
    path_text = os.path.join(self._output_path, self._company_id)

    clean_list = self._dc.clean(''.join(sb_list))
    clean_text = ''.join(sb_text) #this has been cleaned in _format_body()

    with open(path_list, 'w') as f:
        f.write(clean_list)

    with open(path_text, 'w') as f:
        f.write(clean_text)


def _format_body(self, html, title):
    html = html.replace('\r\n', '\n') #kill windows/dos carriage returns
    for s in self._br:
        html = html.replace(s, self._marker + s) #mark all future line breaks
        html = html.replace(s.upper(), self._marker + s) #check for caps in tags

    body = self._src.get_text(html) #kill tags

    body = self._dc.clean_unicode(body) #clean data

    body = self._pattern1.sub(' ', body) #kill line breaks (MUST add a space instead!)
    body = self._pattern2.sub('\n\n', body) #insert my line breaks

    body = self._pattern3.sub(' ', body) #kill extra spaces

    for s in self._replacements:
        body = body.replace(s, self._replacements[s])

    body = body.strip()
    if body.startswith(title): #get rid of title
        body = body[len(title):]

    return body.strip()


def _load_replacements(self):
    self._replacements = {}
    self._replacements['&#8216;'] = "'"
    self._replacements['&#8217;'] = "'"
    self._replacements['&lsquo;'] = "'"
    self._replacements['&rsquo;'] = "'"
    self._replacements['&#8220;'] = '"'
    self._replacements['&#8221;'] = '"'
    self._replacements['&ldquo;'] = '"'
    self._replacements['&rdquo;'] = '"'
    self._replacements['&#8249;'] = "'"
    self._replacements['&#8250;'] = "'"
    self._replacements['&lsaquo;'] = "'"
    self._replacements['&rsaquo;'] = "'"


def _load_linebreaks(self):
    self._br = set()
    self._br.add('<p>')
    self._br.add('<p ')
    self._br.add('<li>')
    self._br.add('<li ')
    self._br.add('<div')
    self._br.add('<h1')
    self._br.add('<h2')
    self._br.add('<h3')
    self._br.add('<h4')
    self._br.add('<h5')
    self._br.add('<h6')
    self._br.add('<table')
    self._br.add('<tr')
    self._br.add('<td')
    self._br.add('<br>')
    self._br.add('<br ')
    self._br.add('<blockquote')
    self._br.add('<caption')
    self._br.add('<dd')
    self._br.add('<dl')
    self._br.add('<dt')
    self._br.add('<form')
    self._br.add('<iframe')
    self._br.add('<input')
    self._br.add('<label')
```

```
        self._br.add('<pre')


    def _load_path(self):
        cfr = ConfigReader()
        root = cfr.get('ROOT_ORIGINAL')
        path1 = cfr.get('DOWNLOADED_PR')
        path2 = os.path.join(root, path1)
        self._path_dir = os.path.join(path2, self._company_id)


    def _load_list(self):
        self._list = []

        filename = common.get_list_file_name()
        path = os.path.join(self._path_dir, filename)

        with open(path) as f:
            write = False #no doc has been read yet: there's nothing to write

            while True:
                line = f.readline()
                if line == '':
                    self._add_to_list(release_id, date, title, url) #flush the last doc
                    break

                if line.startswith('LOTW-ID: '): #found new doc
                    if write: #write if there is a previous doc to write
                        self._add_to_list(release_id, date, title, url)
                    release_id = line[9:].strip()
                    write = True

                if line.startswith('LOTW-DATE: '):
                    date = line[11:].strip()

                if line.startswith('LOTW-TITLE: '):
                    title = line[12:].strip()

                if line.startswith('LOTW-URL: '):
                    url = line[10:].strip()


    def _add_to_list(self, release_id, date, title, url):
        item = [release_id, date, title, url]
        self._list.append(item)
```

## pr/sources/abbot.py

```
import sys
import datetime
import re
from urlparse import urljoin
from urllib import urlopen
import nltk
from base_source import BaseSource

from shared import common

class Abbott(BaseSource):
    def __init__(self):
        self._ptn_date = re.compile('^(\w+) (\d{1,2})')


    def get_linkpages(self):
        all_html = []
        base =
'http://www.abbott.com/global/url/pressReleases/en_US/60.5:5/general_content/Press_Release_Selector_01
.htm?page={0}&year=2012'
        for i in range(0,3):
            url = base.format(i)
            print 'collecting links from {0}'.format(url)
            html = urlopen(url).read()
            start = html.index('<tr class="press-release">')
            end = html.index('bottom pagination', start)
            html = html[start:end]
            all_html.append(html)
```

```
        return all_html


    def is_link(self, tag):
        return tag.has_key('href') and tag.parent.name == 'td' and \
            tag.parent.has_key('class') and tag.parent.get('class')[0] == 'description'


    def is_title(self, tag):
        return self.is_link(tag)


    def is_date(self, tag):
        return tag.name == 'p' and tag.parent.name == 'td' and \
                tag.parent.has_key('class') and \
                tag.parent.get('class')[0] == 'date' and \
                self._ptn_date.match(tag.string.strip())


    def get_link(self, link):
        return urljoin('http://www.abbott.com/', link)


    def get_title(self, tag):
        sb = []
        for s in tag.stripped_strings:
            sb.append(s)
        return ''.join(sb)


    def get_date(self, tag):
        raw = tag.string.strip()
        match = self._ptn_date.match(raw)
        if not match:
            raise Exception('date format did not match pattern')
        year = 2012
        month = common.get_month_by_name(match.group(1))
        day = int(match.group(2))
        return datetime.date(year, month, day)


    def get_encoding(self):
        return "utf-8"

    def get_text(self, html):
        start = html.index('<div id="press-release-lower-content"')
        end = html.index('<div class="prcontact hr-stamp">', start)
        html = html[start:end]
        return self._filter_html(html)
```

## news/formatter.py

```
import datetime
import os
import sys
import re
from shared import common
from shared.config import ConfigReader
from shared.data_cleaner import DataCleaner


class Formatter(object):

    def __init__(self, company_id, output_path):
        self._company_id = company_id
        self._output_path = output_path

        self._sb_list = []
        self._sb_text = []
        self._counter = 0

        self._init_regex()
        self._set_flags(False, False, False, False)
        self._reset_content()
        self._load_nonpubs()
```

```
def run(self):
    path = self._get_input_path()
    with open(path) as f:
        self._scan(f)

    self._write_files()


def _write_files(self):
    list_file = common.get_list_file_name(self._company_id)
    path_list = os.path.join(self._output_path, list_file)
    path_text = os.path.join(self._output_path, self._company_id)

    dc = DataCleaner('utf-8')
    clean_list = dc.clean(''.join(self._sb_list))
    clean_text = dc.clean(''.join(self._sb_text))

    with open(path_list, 'w') as f:
        f.write(clean_list)

    with open(path_text, 'w') as f:
        f.write(clean_text)


def _scan(self, f):
    write = False #no doc has been read yet: there's nothing to write
    while True:
        line = f.readline()
        if line == '':
            self._write_doc() #flush the last doc
            break

        if self._re_start.search(line): #found new doc
            if write: #if there is a previous doc to write
                self._write_doc()
            self._set_flags(True, False, False, False)

        if line.startswith('HEADLINE'):
            self._set_flags(False, True, False, False)
            write = True
            line = line[10:] #drop 'HEADLINE: '

        if line.startswith('BYLINE'):
            self._set_flags(False, False, True, False)
            write = True
            line = line[8:] #drop 'BYLINE: '
            if line.lower().startswith('by '):
                line = line[3:] #drop 'By |by '

        if line.strip() == 'BODY:': #use this to protect against 'BODY:' in the text
            self._set_flags(False, False, False, True)
            write = True
            line = line[6:] #drop 'BODY: '

        if self._ishead:
            line = line.strip()
            if len(line) > 0:
                self._head.append(line)

        if self._isheadline:
            line = line.strip()
            self._headline = self._headline + ' ' + line

        if self._isbyline:
            line = line.strip()
            self._byline = self._byline + ' ' + line

        if self._isbody:
            if not (line.startswith('DOCUMENT-TYPE: ') or
                line.startswith('PUB-TYPE: ') or
                line.startswith('ORGANIZATION: ') or
                line.startswith('LOAD-DATE: ')):
                self._body.append(line)


def _write_doc(self):
    self._parse_head()

    pub = self._pub.strip()
    byline = self._byline.strip()
    headline = self._headline.strip()
```

```
    body = self._format_body(self._body)

    if len(byline) > 100:
        byline = byline[:100]

    if len(headline) > 500:
        headline = headline[:500]

    if self._date is None:
        raise Exception('Found no date for <<{0}>> in <<{1}>>'.format(headline, pub))
    else:
        date = self._date.strftime('%Y-%m-%d')

    self._counter += 1

    self._sb_list.append('LOTW-ID: {0}\n'.format(self._counter))
    self._sb_list.append('LOTW-PUB: {0}\n'.format(pub))
    self._sb_list.append('LOTW-DATE: {0}\n'.format(date))
    self._sb_list.append('LOTW-HEADLINE: {0}\n'.format(headline))
    self._sb_list.append('LOTW-BYLINE: {0}\n\n'.format(byline))

    self._sb_text.append('LOTW-ID: {0}\n'.format(self._counter))
    self._sb_text.append('{0}\n\n\n\n'.format(body))

    self._set_flags(False, False, False, False)
    self._reset_content()


def _format_body(self, body):
    newbody = []
    for line in body:
        line = line.replace('\r\n', '\n') #kill windows/dos carriage returns
        line = line.rstrip() #kill line breaks at the end of line
        if len(line) > 0: #if there is text, add whitspace
            line = line + ' '
        else:
            line = '\n\n' #otherwise, add a line break
        newbody.append(line)

    text = ''.join(newbody).strip()
    pattern = re.compile(r'\n{3,}')
    return pattern.sub('\n\n', text)


def _parse_head(self):
    publines = []
    header = self._head
    for line in header:
        linelower = line.lower()
        if '2012' in line and not 'copyright' in linelower:
            self._parse_date(line)
        else:
            ispub = True
            for nonpub in self._nonpubs: #check line against non-pub stopwords
                if nonpub in linelower:
                    ispub = False
            if ispub:
                publines.append(line)

    #special handling for this dataset only!
    if len(publines) == 2 and not (publines[0] == 'The New York Times' \
        and publines[1] == 'The International Herald Tribune'):
        publines[0] = publines[0] + ' ----- ' + publines[1]

    if publines[0].strip() == 'Richmond Times Dispatch (Virginia)':
        publines[0] = 'Richmond Times-Dispatch (Virginia)'
    elif publines[0].strip() == 'The New York Times ----- National':
        publines[0] = 'The New York Times'

    self._pub = publines[0]

def _parse_date(self, text):
    text = text.replace(',', ' ')
    words = text.split()
    if len(words) == 2:
        self._date = datetime.datetime.strptime(text, '%B %Y')
    elif len(words) == 3:
        self._date = datetime.datetime.strptime(text, '%B %d %Y')
    elif len(words) == 4:
        self._date = datetime.datetime.strptime(text, '%B %d %Y %A')
    elif len(words) == 7:
```

```
            words = words[:3]
            text = ' '.join(words)
            self._date = datetime.datetime.strptime(text, '%B %d %Y')
        else:
            print len(words)
            raise Exception('ERROR: cannot parse date: {0}'.format(text))


    def _init_regex(self):
        months = '(January)|(February)|(March)|(April)|(May)|(June)| \
            (July)|(August)|(September)|(October)|(November)|(December)'
        days = '(Monday)|(Tuesday)|(Wednesday)|(Thursday)|(Friday)|(Saturday)|(Sunday)'
        pattern_date = r'^\s*' + months + '\s\d+,\s\d+\s' + days + '\s*$'
        pattern_start = r'^\s*\d+\s+of\s+\d+\s+DOCUMENTS\s*$'

        self._re_start = re.compile(pattern_start)
        self._re_date = re.compile(pattern_date)


    def _set_flags(self, ishead, isheadline, isbyline, isbody):
        self._ishead = ishead
        self._isheadline = isheadline
        self._isbyline = isbyline
        self._isbody = isbody


    def _reset_content(self):
        self._head = []
        self._pub = ''
        self._date = None
        self._headline = ''
        self._byline = ''
        self._body = []


    def _load_nonpubs(self):
        self._nonpubs = set()
        cfr = ConfigReader()
        path = os.path.abspath(cfr.get('NONPUBS'))
        with open(path) as f:
            for line in f.readlines():
                line = line.strip().lower()
                if len(line) > 0:
                    self._nonpubs.add(line)


    def _get_input_path(self):
        cfr = ConfigReader()
        root = cfr.get('ROOT_ORIGINAL')
        path1 = cfr.get('DOWNLOADED_NEWS')
        path2 = os.path.join(root, path1)
        return os.path.join(path2, self._company_id)
```

## shared/data_cleaner.py

```
import sys
import unicodedata
from HTMLParser import HTMLParser

class DataCleaner(object):

    def __init__(self, encoding='utf-8'):
        self._encoding = encoding
        self._load_dic()


    def clean(self, text):
        #make unicode
        u = text.decode(self._encoding)
        return self.clean_unicode(u)


    #used for releases: the output from soup is unicode
    def clean_unicode(self, u):
        hp = HTMLParser()

        #translate important characters
```

```
        u = u.translate(self._dic)

        #unescape HTML entities and characters
        u = hp.unescape(u)

        #make ascii: transform or ignore the rest of non-ascii chars
        a = unicodedata.normalize('NFKD', u).encode('ascii', 'ignore')

        #clean upremaining entities
        a = hp.unescape(a)

        a = a.replace('`', "'") #replace apostrophies with single quotes
        a = a.replace("''", '"') #replace double single quotes with regular doubloe quotes

        return a


    def _load_dic(self):
        self._dic = {}
        self._dic[ord(u'\u2018')] = ord("'")  #Single curved quote, left
        self._dic[ord(u'\u2019')] = ord("'")  #Single curved quote, right
        self._dic[ord(u'\u2019')] = ord("'")  #Single curved quote, right
        self._dic[ord(u'\u201B')] = ord("'")  #single reversed comma, quotation mark
        self._dic[ord(u'\u201F')] = ord("'")  #double reversed comma, quotation mark
        self._dic[ord(u'\u301D')] = ord("'")  #reversed double prime quotation mark
        self._dic[ord(u'\u301E')] = ord("'")  #double prime quotation mark
        self._dic[ord(u'\uFF02')] = ord("'")  #Halfwidth and Fullwidth Forms
        self._dic[ord(u'\u0060')] = ord("'")  #GRAVE ACCENT
        self._dic[ord(u'\u00B4')] = ord("'")  #ACUTE ACCENT

        self._dic[ord(u'\u201C')] = ord('"')  #Double curved quote, or "curly quote," left
        self._dic[ord(u'\u201D')] = ord('"')  #Double curved quote, right
        self._dic[ord(u'\uFF07')] = ord('"')  #Halfwidth and Fullwidth Forms
        self._dic[ord(u'\xab')] = ord('"')    #Halfwidth and Fullwidth Forms
        self._dic[ord(u'\xbb')] = ord('"')    #Halfwidth and Fullwidth Forms

        self._dic[ord(u'\u00AE')] = None  #registered trademark
        self._dic[ord(u'\u2122')] = None  #trademark sign

        self._dic[ord(u'\u2010')] = ord('-')  #dash
        self._dic[ord(u'\u2011')] = ord('-')  #dash
        self._dic[ord(u'\u2012')] = ord('-')  #dash
        self._dic[ord(u'\u2013')] = ord('-')  #dash
        self._dic[ord(u'\u2014')] = ord('-')  #dash
        self._dic[ord(u'\u2015')] = ord('-')  #dash
        self._dic[ord(u'\u2500')] = ord('-')  #dash
        self._dic[ord(u'\u2212')] = ord('-')  #dash
```

## shared/tokenizer.py

```
import nltk
from nltk.tokenize import word_tokenize


class Tokenizer(object):

    def __init__(self):
        #create pre-trained sentence tokenizer: tokenizing words alone is incorrect
        self._s_tokenizer = nltk.data.load('/tokenizers/punkt/english.pickle')


    def get_tokens(self, text):
        #tokenize text into sentences, arg must be False to prevent modification
        sents = self.get_sentences(text)
        #tokenize sentences into word-tokens
        t_sents = (word_tokenize(s) for s in sents)
        #return flattened token list
        return sum(t_sents, [])


    def get_sentences(self, text):
        return self._s_tokenizer.tokenize(text, realign_boundaries=False)
```

## data/articles.py

```
import os.path
import datetime
from shared import common
from shared.config import ConfigReader
import pubs

class ArticleLoader(object):

    def __init__(self, company_id):
        self._news = {}
        self._load(company_id)


    def get_articles(self):
        return self._news


    def _load(self, company_id):
        cfr = ConfigReader()
        p1 = cfr.get('ROOT_ORIGINAL')
        p2 = cfr.get('FORMATTED_NEWS')
        p3 = os.path.join(p1, p2)
        metafile = common.get_list_file_name(company_id)

        path_text = os.path.join(p3, str(company_id))
        path_meta = os.path.join(p3, metafile)

        news_text = self._load_text(path_text)
        self._load_meta(path_meta, news_text)


    def _load_text(self, path):
        text = {}
        article_id = -1
        sb = []
        with open(path) as f:
            while True:
                line = f.readline()
                if line == '':
                    text[article_id] = ''.join(sb) #flush last record
                    break

                if line.startswith('LOTW-ID: '):
                    if article_id != -1:
                        text[article_id] = ''.join(sb)
                    article_id = int(line[9:].strip())
                    sb = []
                else:
                    sb.append(line)
        return text


    def _load_meta(self, path, news_text):
        article_id = -1
        with open(path) as f:
            while True:
                line = f.readline()
                if line == '':
                    self._news[article_id] = Article(article_id, pub, date, headline, byline,
news_text[article_id])
                    break

                if line.startswith('LOTW-ID: '):
                    if article_id != -1:
                        self._news[article_id] = Article(article_id, pub, date, headline, byline,
news_text[article_id])
                    article_id = int(line[9:].strip())


                if line.startswith('LOTW-PUB: '):
                    pub = line[10:]

                if line.startswith('LOTW-DATE: '):
                    date = line[11:]

                if line.startswith('LOTW-HEADLINE: '):
                    headline = line[15:]
```

```
                if line.startswith('LOTW-BYLINE: '):
                    byline = line[13:]


class Article(object):

    def __init__(self, article_id, pub, date, headline, byline, body):
        self._news = article_id, pub, self._load_date(date), headline, byline, body,
self._load_pub_id(pub)

    def id(self):
        return self._news[0]

    def pub(self):
        return self._news[1]

    def date(self):
        return self._news[2]

    def headline(self):
        return self._news[3]

    def byline(self):
        return self._news[4]

    def body(self):
        return self._news[5]

    def pub_id(self):
        return self._news[6]

    def _load_pub_id(self, pub):
        return pubs.get_pub_id(pub)

    def _load_date(self, date):
        date = date.strip()
        return datetime.datetime.strptime(date, '%Y-%m-%d')
```

## data/releases.py

```
import os.path
import datetime
from shared import common
from shared.config import ConfigReader


class ReleaseLoader(object):

    def __init__(self, company_id):
        self._releases = {}
        self._load(company_id)


    def get_releases(self):
        return self._releases


    def _load(self, company_id):
        cfr = ConfigReader()
        p1 = cfr.get('ROOT_ORIGINAL')
        p2 = cfr.get('FORMATTED_PR')
        p3 = os.path.join(p1, p2)
        metafile = common.get_list_file_name(company_id)

        path_text = os.path.join(p3, str(company_id))
        path_meta = os.path.join(p3, metafile)

        pr_text = self._load_text(path_text)
        self._load_meta(path_meta, pr_text)


    def _load_text(self, path):
        text = {}
        release_id = -1
        sb = []
        with open(path) as f:
```

```
            while True:
                line = f.readline()
                if line == '':
                    text[release_id] = ''.join(sb) #flush last record
                    break

                if line.startswith('LOTW-ID: '):
                    if release_id != -1:
                        text[release_id] = ''.join(sb)
                    release_id = int(line[9:].strip())
                    sb = []
                else:
                    sb.append(line)
        return text


    def _load_meta(self, path, pr_text):
        release_id = -1
        with open(path) as f:
            while True:
                line = f.readline()
                if line == '':
                    self._releases[release_id] = Release(release_id, date, title, url,
pr_text[release_id])
                    break

                if line.startswith('LOTW-ID: '):
                    if release_id != -1:
                        self._releases[release_id] = Release(release_id, date, title, url,
pr_text[release_id])
                    release_id = int(line[9:].strip())

                if line.startswith('LOTW-DATE: '):
                    date = line[11:]

                if line.startswith('LOTW-TITLE: '):
                    title = line[12:]

                if line.startswith('LOTW-URL: '):
                    url = line[10:]


class Release(object):

    def __init__(self, release_id, date, title, url, body):
        self._releases = release_id, self._load_date(date), title, url, body

    def id(self):
        return self._releases[0]

    def date(self):
        return self._releases[1]

    def title(self):
        return self._releases[2]

    def url(self):
        return self._releases[3]

    def body(self):
        return self._releases[4]

    def _load_date(self, date):
        date = date.strip()
        return datetime.datetime.strptime(date, '%Y-%m-%d')
```

## data/scores.py

```
import cPickle
import os.path
import string
from shared import common
from shared.config import ConfigReader

class ScoreLoader(object):

    def __init__(self, company_id):
        self._load_dictionaries(company_id)

    def count_pos_rel_sentences(self, release_id):
        return self._rel_dict[release_id].number_of_pos_sents()

    def count_neg_rel_sentences(self, release_id):
        return self._rel_dict[release_id].number_of_neg_sents()

    def count_posneg_rel_sentences(self, release_id):
        return self._rel_dict[release_id].number_of_posneg_sents()

    def count_subj_rel_sentences(self, release_id):
        return self._rel_dict[release_id].number_of_subj_sents()

    def count_all_rel_sentences(self, release_id):
        return self._rel_dict[release_id].number_of_all_sents()

    def count_pos_rel_words(self, release_id):
        return self._rel_dict[release_id].number_of_pos_words()

    def count_neg_rel_words(self, release_id):
        return self._rel_dict[release_id].number_of_neg_words()

    def count_pos_art_sentences(self, article_id):
        return self._art_dict[article_id].number_of_pos_sents()

    def count_neg_art_sentences(self, article_id):
        return self._art_dict[article_id].number_of_neg_sents()

    def count_posneg_art_sentences(self, article_id):
        return self._art_dict[article_id].number_of_posneg_sents()

    def count_subj_art_sentences(self, article_id):
        return self._art_dict[article_id].number_of_subj_sents()

    def count_all_art_sentences(self, article_id):
        return self._art_dict[article_id].number_of_all_sents()

    def count_pos_art_words(self, article_id):
        return self._art_dict[article_id].number_of_pos_words()

    def count_neg_art_words(self, article_id):
        return self._art_dict[article_id].number_of_neg_words()


    def _load_dictionaries(self, company_id):

        self._rel_dict = {}
        path_rel = common.get_sentiment_scores_path(company_id) + '-' + common.DOCTYPE_PR
        self._load_dict(path_rel, self._rel_dict)

        self._art_dict = {}
        path_art = common.get_sentiment_scores_path(company_id) + '-' + common.DOCTYPE_NEWS
        self._load_dict(path_art, self._art_dict)


    def _load_dict(self, path, dic):
        with open(path) as f:
            lines = f.readlines()

            is_first_line = True
            current_id = -1
            pos_sents = 0
            neg_sents = 0
            posneg_sents = 0
            subj_sents = 0
            all_sents = 0
            pos_words = 0
            neg_words = 0
```

```
            for line in lines:
                pairs = line.split()
                doc_id = int(pairs[0].split('=')[1])
                # sent_id = int(pairs[1].split('=')[1])
                pos = int(pairs[2].split('=')[1])
                neg = int(pairs[3].split('=')[1])

                if doc_id == current_id or is_first_line:
                    is_first_line = False
                    current_id = doc_id

                    all_sents += 1

                    if pos > 0 and neg == 0:
                        pos_sents += 1
                        subj_sents += 1
                    elif pos == 0 and neg > 0:
                        neg_sents += 1
                        subj_sents += 1
                    elif pos > 0 and neg > 0:
                        posneg_sents += 1
                        subj_sents += 1

                    pos_words += pos
                    neg_words += neg

                else:
                    newdoc = Document(current_id, pos_sents, neg_sents, posneg_sents, \
                            subj_sents, all_sents, pos_words, neg_words)

                    dic[current_id] = newdoc

                    #resert everything
                    current_id = doc_id
                    pos_sents = 0
                    neg_sents = 0
                    posneg_sents = 0
                    subj_sents = 0
                    all_sents = 0
                    pos_words = 0
                    neg_words = 0

                #write last record
                newdoc = Document(current_id, pos_sents, neg_sents, posneg_sents, \
                        subj_sents, all_sents, pos_words, neg_words)

                dic[current_id] = newdoc


class Document(object):

    def __init__(self, doc_id, pos_sents, neg_sents, posneg_sents, subj_sents,\
            all_sents, pos_words, neg_words):
        self._data = int(doc_id), int(pos_sents), int(neg_sents), \
                int(posneg_sents), int(subj_sents), int(all_sents), \
                int(pos_words), int(neg_words)

    def id(self):
        return self._data[0]

    def number_of_pos_sents(self):
        return self._data[1]

    def number_of_neg_sents(self):
        return self._data[2]

    def number_of_posneg_sents(self):
        return self._data[3]

    def number_of_subj_sents(self):
        return self._data[4]

    def number_of_all_sents(self):
        return self._data[5]

    def number_of_pos_words(self):
        return self._data[6]

    def number_of_neg_words(self):
        return self._data[7]
```

## data/subjlexicon.py

```
import cPickle
import string
import nltk
from shared import common
from shared.config import ConfigReader

class SubjLexiconLoader(object):

    def __init__(self, strongonly=False):
        self._words = {}
        self._stems = {}
        self._stemmer = nltk.PorterStemmer()
        self._load_codes()
        self._load_lexicon(strongonly)


    def get_polarity(self, word, pos):
        if not pos in self._codes:
            pos_code = 'anypos'
        else:
            pos_code = self._codes[pos]

        key1 = word, pos_code
        key2 = word

        stemmed = self._stemmer.stem(word)
        key3 = stemmed, pos_code
        key4 = stemmed

        if key1 in self._words:
            return self._words[key1]
        elif key2 in self._words:
            return self._words[key2]
        elif key3 in self._stems:
            return self._stems[key3]
        elif key4 in self._stems:
            return self._stems[key4]
        else:
            return None


    def _load_lexicon(self, strongonly):
        path = common.get_subjlexicon_path()
        with open(path) as f:
            lines = f.readlines()
            for line in lines:
                pairs = line.split()
                stype = pairs[0].split('=')[1]
                word = pairs[2].split('=')[1]
                pos_code = pairs[3].split('=')[1]
                stemmed = pairs[4].split('=')[1]
                polarity = pairs[5].split('=')[1]

                if strongonly and stype == 'weaksubj':
                    continue

                if pos_code == 'anypos':
                    key = word
                else:
                    key = word, pos_code

                if stemmed == 'n':
                    self._words[key] = polarity
                else:
                    self._stems[key] = polarity



    def _load_lexicon_stemmed(self):
        path = common.get_subjlexicon_path(True)
        with open(path) as f:
            lines = f.readlines()
            for line in lines:
                line = line.split()
                key = line[0]
                if not key in self._dict:
                    self._dict[key] = line[2]
```

```
                key = line[0], line[1]
                if not key in self._dict:
                    self._dict[key] = line[2]


    def _load_codes(self):
        self._codes = {}
        self._codes['JJ']   = 'adj'
        self._codes['JJR']  = 'adj'
        self._codes['JJS']  = 'adj'
        self._codes['NN']   = 'noun'
        self._codes['NNS']  = 'noun'
        self._codes['NNP']  = 'noun'
        self._codes['NNPS'] = 'noun'
        self._codes['RB']   = 'adverb'
        self._codes['RBR']  = 'adverb'
        self._codes['RBS']  = 'adverb'
        self._codes['WRB']  = 'adverb'
        self._codes['VB']   = 'verb'
        self._codes['VBD']  = 'verb'
        self._codes['VBG']  = 'verb'
        self._codes['VBN']  = 'verb'
        self._codes['VBP']  = 'verb'
        self._codes['VBZ']  = 'verb'
```

## data/matches.py

```
import cPickle
import string
from shared import common


class MatchMaker(object):

    def __init__(self, company_id, matches_name):
       self._company_id = company_id
       self._matches_name = matches_name
       self._r_a_blocks = {}


    def add_blocks(self, release_id, article_id, blocks):
        if not release_id in self._r_a_blocks:
            a_blocks = {}
            self._r_a_blocks[release_id] = a_blocks
        else:
            a_blocks = self._r_a_blocks[release_id]

        a_blocks[article_id] = blocks


    def save(self):
        path = common.get_pickled_matches_path(self._company_id, self._matches_name)
        with open(path, 'wb') as f:
            cPickle.dump(self._r_a_blocks, f, -1)


class MatchLoader(object):

    def __init__(self, company_id, matches_name):
        self._load(company_id, matches_name)


    def count_rel_art_pairs(self):
        count = 0
        for release_id in self._rel_art_blocks:
            art_blocks = self._rel_art_blocks[release_id]
            count += len(art_blocks)
        return count


    def count_matching_blocks(self):
        count = 0
        for release_id in self._rel_art_blocks:
            art_blocks = self._rel_art_blocks[release_id]
            for article_id in art_blocks:
                block = art_blocks[article_id]
```

```
                count += len(block)
        return count


    def get_release_ids(self):
        ids = set()
        for release_id in self._rel_art_blocks:
            ids.add(release_id)
        return ids


    def get_article_ids(self, release_id=None):
        ids = set()
        if release_id is None: #loop through all releases
            for release_id in self._rel_art_blocks:
                self._get_article_ids_helper(release_id, ids)
        else: #use only one release
            self._get_article_ids_helper(release_id, ids)
        return ids

    def _get_article_ids_helper(self, release_id, ids):
        art_blocks = self._rel_art_blocks[release_id]
        for article_id in art_blocks:
            ids.add(article_id)


    def get_matches(self, release_id, article_id):
        art_blocks = self._rel_art_blocks[release_id]
        return art_blocks[article_id]


    def _load(self, company_id, matches_name):
        pickle_path = common.get_pickled_matches_path(company_id, matches_name)
        with open(pickle_path, 'rb') as f:
            self._rel_art_blocks = cPickle.load(f)
```

## data/tokens.py

```
import cPickle
import string
from shared import common
from shared.config import ConfigReader

class TokenLoader(object):

    def __init__(self, company_id):
        self._load_dictionaries(company_id)
        self._exclude_tokens = set(string.punctuation)
        self._exclude_tokens.add(ConfigReader().get('MARKER_BR'))


    def get_release_tokens(self, release_id, lowercase):
        tokens = self._pr_tokens[release_id]
        if lowercase:
            tokens = [t.lower() for t in tokens]
        return tokens


    def get_article_tokens(self, article_id, lowercase):
        tokens = self._news_tokens[article_id]
        if lowercase:
            tokens = [t.lower() for t in tokens]
        return tokens


    def get_stripped_release_token_block(self, release_id, start, length):
        return self._strip_tokens(self._pr_tokens[release_id], start, length)


    def get_stripped_article_token_block(self, article_id, start, length):
        return self._strip_tokens(self._news_tokens[article_id], start, length)


    def _strip_tokens(self, tokens, start, end):
        token_list = tokens[start:end]
        return [t for t in token_list if t not in self._exclude_tokens]
```

```
    def _load_dictionaries(self, company_id):
        path = common.get_pickled_news_tokens_path(company_id)
        with open(path, 'rb') as f:
            self._news_tokens = cPickle.load(f)

        path = common.get_pickled_pr_tokens_path(company_id)
        with open(path, 'rb') as f:
            self._pr_tokens = cPickle.load(f)
```

## analyze/duplicate_finder.py

```
import sys
import hashlib
from shared import common
from data.articles import ArticleLoader
from data.releases import ReleaseLoader


class DuplicateFinder(object):

    def __init__(self, company_id):
        self._company_id = company_id


    def find_release_duplicates(self):
        hashes = self._load_rel_hashes()
        dups = self._get_duplicates(hashes)
        if len(dups) > 0:
            for d in dups:
                print d

    #Quick 'n dirty solution: print to file since I know there are duplicates.
    #   Run once and change permissions on output directory.
    def find_article_duplicates(self):
        hashes = self._load_art_hashes()
        dups = self._get_duplicates(hashes)
        if len(dups) > 0:
            path = common.get_art_duplicates_path(self._company_id)
            with open(path, 'w') as f:
                for d in dups:
                    f.write('{0}\n'.format(d))


    def _get_duplicates(self, hashes):
        dups = set()
        hashset = set()
        for id in hashes:
            if hashes[id] in hashset:
                dups.add(id)
            else:
                hashset.add(hashes[id])
        return dups


    def _load_rel_hashes(self):
        hashes = {}
        releases = ReleaseLoader(self._company_id).get_releases()
        for release_id in releases:
            release = releases[release_id]
            text = str(release.date()) + release.title() + release.body()
            m = hashlib.md5()
            m.update(text)
            hashes[release_id] = m.hexdigest()
        return hashes


    def _load_art_hashes(self):
        hashes = {}
        articles = ArticleLoader(self._company_id).get_articles()
        for article_id in articles:
            article = articles[article_id]
            text = str(article.date()) + article.pub() + article.headline() + article.body()
            m = hashlib.md5()
            m.update(text)
```

```
            hashes[article_id] = m.hexdigest()
        return hashes
```

## analyze/block_finder.py

```
import sys
from data.tokens import TokenLoader
from data.matches import MatchLoader
from shared.config import ConfigReader
from shared import common


POS_IN_BLOCK_ART = 0
POS_IN_BLOCK_REL = 1

#TIME_DELTA = 30

class BlockFinder(object):

    def __init__(self, company_id, matches_name):

        self._company_id = company_id
        self._matchloader = MatchLoader(company_id, matches_name)
        self._tokens = TokenLoader(company_id)
        self._br = ConfigReader().get('MARKER_BR')


    def print_all_matching_blocks(self, min_len, max_len):
        for release_id in self._matchloader.get_release_ids():
            for article_id in self._matchloader.get_article_ids(release_id):
                blocks = self._matchloader.get_matches(release_id, article_id)
                for block in blocks:
                    i = block[0]
                    j = block[1]
                    k = block[2]

                    rel_match = self._tokens.get_stripped_release_token_block(release_id, j,j+k)

                    if len(rel_match) >= min_len and len(rel_match) < max_len:
                        mb = ' '.join(rel_match)
                        mb = mb.replace(self._br, ' ')
                        print mb


    #prints blocks of min_length or larger occuring in more than one release -
    #   i.e., bad discriminators between releases
    def print_all_nondiscrim_release_blocks(self, min_len, max_len):

        blockset_dict = {}

        for release_id in self._matchloader.get_release_ids():

            blockset = set() #set of blocks for current release
            blockset_dict[release_id] = blockset

            for article_id in self._matchloader.get_article_ids(release_id):
                blocks = self._matchloader.get_matches(release_id, article_id)
                for block in blocks:
                    i = block[0]
                    j = block[1]
                    k = block[2]

                    rel_match = self._tokens.get_stripped_release_token_block(release_id, j,j+k)

                    if len(rel_match) >= min_len and len(rel_match) < max_len:
                        mb = ' '.join(rel_match)
                        mb = mb.replace(self._br, ' ')
                        mb = mb.lower().strip()
                        blockset.add(mb)

        #count occurances of each block per release
        bcounts = {}
        for release_id in blockset_dict:
            blockset = blockset_dict[release_id]
            for b in blockset:
                if b in bcounts:
                    bcounts[b] += 1
```

```
                else:
                    bcounts[b] = 1

        #print blocks which occur more than once per release
        result = [key for key in bcounts if bcounts[key] > 1]
        for r in result:
            print r
```

## analyze/match_finder.py

```
import sys
from operator import itemgetter
from difflib import SequenceMatcher
from data.releases import ReleaseLoader
from data.articles import ArticleLoader
from data.tokens import TokenLoader
from data.matches import MatchMaker
from data.blocks import BlockLoader
from data.duplicates import DuplicateLoader


class MatchFinder(object):

    def __init__(self, company_id, release_ids, article_ids, required_length, min_length,
blocks_name_toignore):

        self._company_id = company_id
        self._release_ids = release_ids
        self._article_ids = article_ids
        self._required_length = required_length
        self._min_length = min_length

        self._tokens = TokenLoader(company_id)

        self._releases = ReleaseLoader(company_id).get_releases()
        self._articles = ArticleLoader(company_id).get_articles()

        self._ignoreblocks = BlockLoader(company_id, blocks_name_toignore).get_blocks()
        self._count_ignore = 0

        dloader = DuplicateLoader(company_id)
        self._rel_duplicates = dloader.get_release_duplicates()
        self._art_duplicates = dloader.get_article_duplicates()


    def find_matches(self, output_name):

        matchmaker = MatchMaker(self._company_id, output_name)

        matcher = SequenceMatcher(autojunk=False)
        message = 'Processing company {0}: release {1} of {2}; article {3} of {4}'
        pairs_counter = 0

        for i, release_id in enumerate(self._release_ids): #loop through releases

            if release_id in self._rel_duplicates:
                continue

            matcher.set_seq2(self._tokens.get_release_tokens(release_id, True))
            release_date = self._releases[release_id].date()

            for j, article_id in enumerate(self._article_ids): #loop through articles

                if article_id in self._art_duplicates:
                    continue

                if j % 100 == 0:
                    print message.format(self._company_id, i+1, len(self._release_ids), j+1,
len(self._article_ids))

                matcher.set_seq1(self._tokens.get_article_tokens(article_id, True))
                article_date = self._articles[article_id].date()

                if article_date >= release_date: #search for matches if article appeared after the
release

                    blocks = matcher.get_matching_blocks() #block form: (i,j,k) where i = article
(seq1), j = release (seq2)
```

```
                    if len(blocks) > 0: #if there are blocks

                        valid_blocks = self._get_blocks(blocks, release_id, article_id)
                        if len(valid_blocks) > 0: #if there are valid blocks

                            matchmaker.add_blocks(release_id, article_id, valid_blocks)
                            print '\tfound match for release={0} and article={1}'.format(release_id,
article_id)
                            pairs_counter += 1

        print 'total matching pairs: {0}'.format(pairs_counter)
        print 'ignored bad discriminators: {0}'.format(self._count_ignore)
        matchmaker.save()


    def _get_blocks(self, blocks, release_id, article_id):

        blocklist = []
        required_length_check = False

        for b in blocks:
            i = b[0]
            j = b[1]
            k = b[2]

            rel_match = self._tokens.get_stripped_release_token_block(release_id, j,j+k)
            art_match = self._tokens.get_stripped_article_token_block(article_id, i,i+k)

            rel_temp = ' '.join(rel_match)
            art_temp = ' '.join(art_match)

            if rel_temp.lower() != art_temp.lower():
                print rel_temp.lower()
                print art_temp.lower()
                raise Exception("blocks don't match")

            #check against bad discriminators BEFORE updating required_length_check
            if rel_temp.lower() in self._ignoreblocks:
                self._count_ignore += 1
                continue

            #check for min_length BEFORE updating required_length_check
            if len(rel_match) < self._min_length:
                continue

            if len(rel_match) >= self._required_length:
                required_length_check = True

            blocklist.append(b)

        #sort by length, decending
        if len(blocklist) == 0:
            return []

        if not required_length_check:
            return []
        else:
            blocklist = sorted(blocklist, key=itemgetter(2), reverse=True)
            return blocklist
```

## analyze/postag_writer.py

```
import sys
import cPickle
import os.path
import nltk
from data.matches import MatchLoader
from data.tokens import TokenLoader
from shared.config import ConfigReader
from shared import common

#POS-tags releases and articles for a given macthes set and stores as pickles
class POSTagWriter(object):

    def __init__(self):
        self._br = ConfigReader().get('MARKER_BR')
```

```
    def write_tags(self, matches_name, company_id):

        dic_rel = {}
        dic_art = {}

        matches = MatchLoader(company_id, matches_name)
        tokens = TokenLoader(company_id)

        rel_ids = matches.get_release_ids()
        for count, release_id in enumerate(rel_ids):
            print 'processing release #{0} of {1}'.format(count+1, len(rel_ids))
            tmp = tokens.get_release_tokens(release_id, False)
            self._process_tokens(tmp, dic_rel, release_id)

        art_ids = matches.get_article_ids()
        for count, article_id in enumerate(art_ids):
            print 'processing article #{0} of {1}'.format(count+1, len(art_ids))
            tmp = tokens.get_article_tokens(article_id, False)
            self._process_tokens(tmp, dic_art, article_id)

        path1 = common.get_postags_path()
        path2 = os.path.join(path1, matches_name)

        path = os.path.join(path2, common.DOCTYPE_PR)
        self._pickle(company_id, dic_rel, path)

        path = os.path.join(path2, common.DOCTYPE_NEWS)
        self._pickle(company_id, dic_art, path)


    def _process_tokens(self, tmp, dic, doc_id):
            tokens = ['\n' if t == self._br else t for t in tmp]
            tagged = nltk.pos_tag(tokens)
            dic[doc_id] = tagged


    def _pickle(self, company_id, dic, path):
        filename = '{0}.pickle'.format(company_id)
        filepath = os.path.join(path, filename)
        with open(filepath, 'wb') as f:
            cPickle.dump(dic, f, -1)
```

## analyze/match_writer.py

```
import sys
import os.path
from operator import itemgetter
from data.releases import ReleaseLoader
from data.articles import ArticleLoader
from data.tokens import TokenLoader
from data.matches import MatchLoader
from shared.config import ConfigReader
from shared import common


POS_IN_BLOCK_ART = 0
POS_IN_BLOCK_REL = 1

TIME_DELTA = 14

class MatchWriter(object):

    def __init__(self, company_id, matches_name):

        self._company_id = company_id
        self._matchloader = MatchLoader(company_id, matches_name)
        self._tokens = TokenLoader(company_id)
        self._releases = ReleaseLoader(company_id).get_releases()
        self._articles = ArticleLoader(company_id).get_articles()

        self._br = ConfigReader().get('MARKER_BR')


    def write_matches(self, output_path):
```

```
    html = self._build_html()

    filename = '{0}.html'.format(self._company_id)
    filepath = os.path.join(output_path, filename)
    self._write_html_to_file(filepath, html)


def _build_html(self):
    sb = []
    counter = 0

    releases = self._get_sorted_releases()

    for release in releases:
        self._write_release_header(sb, release)
        articles = self._get_sorted_articles(release.id())

        for article in articles:

            #condition for id=35/32 only
            if self._company_id == '35':
                delta = article.date() - release.date()
                if delta.days >= TIME_DELTA and \
                        not (release.id() == 246 and article.id() == 944) and \
                        not (release.id() == 189 and article.id() == 1213) and \
                        not (release.id() == 71 and article.id() == 2557):

                    continue

            if self._company_id == '32':
                delta = article.date() - release.date()
                if delta.days >= TIME_DELTA:
                    continue


            #print '{0}-{1}'.format(release.id(), article.id())

            blocks = self._matchloader.get_matches(release.id(), article.id())

            self._write_article_summary(sb, blocks, release, article)
            self._write_texts(sb, blocks, release.id(), article.id())
            counter += 1

    print '{0}'.format(counter)
    return ''.join(sb)


def _get_sorted_releases(self):
    ids = self._matchloader.get_release_ids()
    rels = [self._releases[id] for id in ids]
    rels.sort(key = lambda x: x.date())
    return rels


def _get_sorted_articles(self, release_id):
    ids = self._matchloader.get_article_ids(release_id)
    arts = [self._articles[id] for id in ids]
    arts.sort(key = lambda x: x.date())
    return arts


def _write_release_header(self, sb, release):
    sb.append('\n\t<tr>\n\t\t<td colspan="2" class="release-title">')
    sb.append('{0} --- {1} --- {2}\n\t\t</td>\n\t</tr>'.format( \
            release.id(), release.date().strftime('%B %d'), release.title()))


def _write_article_summary(self, sb, blocks, release, article):
    sb.append('\n\t<tr><td colspan=2>')
    sb.append('\n\t\t<table class="tbl-inner1" cellpadding="5" border="1"i>')

    sb.append('\n\t\t\t<tr class="tbl-inner1-title"><td colspan="3" class="article-title">')
    sb.append('R: {0} --- {1} --- {2}\n\t\t</td>\n\t</tr>'.format( \
            release.id(), release.date().strftime('%B %d'), release.title()))

    sb.append('\n\t\t\t<tr class="tbl-inner1-title"><td colspan="3" class="article-title">')
    sb.append('A: {0} --- {1} --- {2} --- {3}\n\t\t</td>\n\t</tr>'.format( \
            article.id(), article.date().strftime('%B %d'), article.headline(), article.pub()))

    sb.append('\n\t\t\t<tr class="tbl-inner1-title"><td>#</td><td>length</td><td>match</td></tr>')
```

```
        for count, block in enumerate(blocks):
            i = block[0] #start in article
            j = block[1] #start in release
            k = block[2] #length

            rel_match = self._tokens.get_stripped_release_token_block(release.id(), j,j+k)
            art_match = self._tokens.get_stripped_article_token_block(article.id(), i,i+k)

            rel_temp = ' '.join(rel_match)
            art_temp = ' '.join(art_match)

            rel_temp = rel_temp.replace(self._br, ' ')
            art_temp = art_temp.replace(self._br, ' ')

            if rel_temp.lower() != art_temp.lower():
                print rel_temp.lower()
                print art_temp.lower()
                raise Exception("blocks don't match")

            sb.append('\n\t\t\t<tr valign="top">')
            sb.append('\n\t\t\t\t<td>{0}</td>'.format(count+1))
            sb.append('\n\t\t\t\t<td>{0}</td>'.format(k))
            sb.append('\n\t\t\t\t<td><span class="match match{0}">{1}</span>\n\t\t</td>'.format(count,
rel_temp))
            sb.append('\n\t\t\t</tr>')

        sb.append('\n\t\t</table>')
        sb.append('\n\t</td></tr>')


    def _write_texts(self, sb, blocks, release_id, article_id):

        rel_tokens = self._tokens.get_release_tokens(release_id, False)
        art_tokens = self._tokens.get_article_tokens(article_id, False)
        rel_html = self._get_text(blocks, rel_tokens, POS_IN_BLOCK_REL)
        art_html = self._get_text(blocks, art_tokens, POS_IN_BLOCK_ART)

        sb.append('\n\t<tr valign="top">')
        sb.append('\n\t\t<td width="50%">{0}\n\t\t</td>'.format(rel_html))
        sb.append('\n\t\t<td width="50%">{0}\n\t\t</td>'.format(art_html))
        sb.append('\n\t</tr>')


    def _get_text(self, blocks, orig_tokens, pos_in_block):
        span_start = '<span class="match match{0}">'
        span_end = '</span>'

        #clone list
        tokens = orig_tokens[:]

        #sort by position in article
        blocks = sorted(blocks, key=itemgetter(pos_in_block))
        a = 0
        for count, block in enumerate(blocks):
            pos = block[pos_in_block] #position in text
            k = block[2] #length

            tokens.insert(pos+a, span_start.format(count))
            tokens.insert(pos+k+a+1, span_end)
            a += 2

        html = ' '.join(tokens)
        html = html.replace(self._br, '<br/>')

        return html


    def _write_html_to_file(self, output_path, html):
        with open(output_path, 'w') as f:
            f.write('<html>\n<head>')
            f.write('\n\t<link rel="stylesheet" type="text/css" href="styles.css">')
            f.write('\n</head>\n<body>\n')
            f.write('\n<table class="tbl-main" cellpadding="5" border="1">')
            f.write(html)
            f.write('\n</table>')
            f.write('\n\n</body>\n</html>')
```

## analyze/sentence_writer.py

```
import sys
import os.path
from nltk.tag import pos_tag
from nltk.tokenize import word_tokenize
from data.releases import ReleaseLoader
from data.articles import ArticleLoader
from data.tokens import TokenLoader
from shared.config import ConfigReader
from shared.tokenizer import Tokenizer
from shared import common
from data.subjlexicon import SubjLexiconLoader


PR_CODE = 'R'
NEWS_CODE = 'A'


class SentenceWriter(object):

    def __init__(self, company_id, release_ids, article_ids, output_name):

        self._company_id = company_id
        self._release_ids = release_ids
        self._article_ids = article_ids
        self._output_name = output_name

        self._releases = ReleaseLoader(company_id).get_releases()
        self._articles = ArticleLoader(company_id).get_articles()

        self._tokenizer = Tokenizer()
        self._lexicon = SubjLexiconLoader()


        self._make_dirs()



    def write_and_calculate(self):

        words_rel_pos = []
        words_art_pos = []
        words_rel_neg = []
        words_art_neg = []
        scores_rel = []
        scores_art = []

        for i, release_id in enumerate(self._release_ids):
            print 'Processing release {0} of {1}'.format(i+1, len(self._release_ids))
            release = self._releases[release_id]
            text = release.title() + '.\n' + release.body()
            self._write_text(release_id, text, PR_CODE, common.DOCTYPE_PR, words_rel_pos,
words_rel_neg, scores_rel)


        for i, article_id in enumerate(self._article_ids):
            print 'Processing article {0} of {1}'.format(i+1, len(self._article_ids))
            article = self._articles[article_id]
            text = article.headline() + '.\n' + article.body()
            self._write_text(article_id, text, NEWS_CODE, common.DOCTYPE_NEWS, words_art_pos,
words_art_neg, scores_art)

        #save word lists and scores
        path = common.get_sentiment_scores_path(self._company_id) + '-' + common.DOCTYPE_PR
        with open(path, 'w') as f:
            f.write('\n'.join(scores_rel))

        path = common.get_sentiment_scores_path(self._company_id) + '-' + common.DOCTYPE_NEWS
        with open(path, 'w') as f:
            f.write('\n'.join(scores_art))

        path = common.get_sentiment_words_pos_path(self._company_id) + '-' + common.DOCTYPE_PR
        with open(path, 'w') as f:
            f.write('\n'.join(words_rel_pos))

        path = common.get_sentiment_words_neg_path(self._company_id) + '-' + common.DOCTYPE_PR
        with open(path, 'w') as f:
            f.write('\n'.join(words_rel_neg))
```

```
        path = common.get_sentiment_words_pos_path(self._company_id) + '-' + common.DOCTYPE_NEWS
        with open(path, 'w') as f:
            f.write('\n'.join(words_art_pos))

        path = common.get_sentiment_words_neg_path(self._company_id) + '-' + common.DOCTYPE_NEWS
        with open(path, 'w') as f:
            f.write('\n'.join(words_art_neg))


    def _write_text(self, text_id, text, code, doctype, words_pos, words_neg, scores):
        html = []
        name = common.get_company_name(self._company_id)

        html.append('<h4>{0}</h4>'.format(name))
        html.append('\n<table class="tbl-main" cellpadding="5" border="1">')

html.append('<tr><td>ID</td><td>text</td><td>POS</td><td>NEG</td><td>BOTH</td><td>N/A</td></tr>')

        sents = self._tokenizer.get_sentences(text)
        for i, s in enumerate(sents):

            pos = 0
            neg = 0

            tokens = word_tokenize(s)
            tagged = pos_tag(tokens)
            sb = []
            for pair in tagged:
                word = pair[0]
                polarity = self._lexicon.get_polarity(word, pair[1])

                if polarity == 'positive':
                    sb.append('<span class="pol-positive">{0}</span>'.format(word))
                    words_pos.append(word)
                    pos += 1

                elif polarity == 'negative':
                    sb.append('<span class="pol-negative">{0}</span>'.format(word))
                    words_neg.append(word)
                    neg += 1

                else:
                    sb.append(pair[0])

            scores.append('doc-id={0} sent-id={1} pos={2} neg={3}'.format(text_id, i+1, pos, neg))

            sent_id = '{0}-{1}-{2}-{3}'.format(self._company_id, code, text_id, i+1)
            html.append('<tr valign="top">')
            html.append('<td>{0}</td>'.format(sent_id))
            html.append('<td>{0}</td>'.format(' '.join(sb)))
            html.append('<td> </td><td> </td><td> </td><td> </td></tr>')

        html.append('\n</table>')

        path = self._get_filepath(doctype, text_id)
        self._write_html_to_file(path, '\n'.join(html))


    def _get_filepath(self, doctype, text_id):
        path_dir = common.get_sents_path(self._output_name, self._company_id)
        path_subdir = os.path.join(path_dir, doctype)
        return os.path.join(path_subdir, str(text_id))


    def _make_dirs(self):
        path = common.get_sents_path(self._output_name, self._company_id)
        if not os.path.exists(path):
            os.mkdir(path)

        rel_path = os.path.join(path, common.DOCTYPE_PR)
        if not os.path.exists(rel_path):
            os.mkdir(rel_path)

        art_path = os.path.join(path, common.DOCTYPE_NEWS)
        if not os.path.exists(art_path):
            os.mkdir(art_path)


    def _write_html_to_file(self, output_path, html):
        with open(output_path, 'w') as f:
            f.write('<html>\n<head>')
```

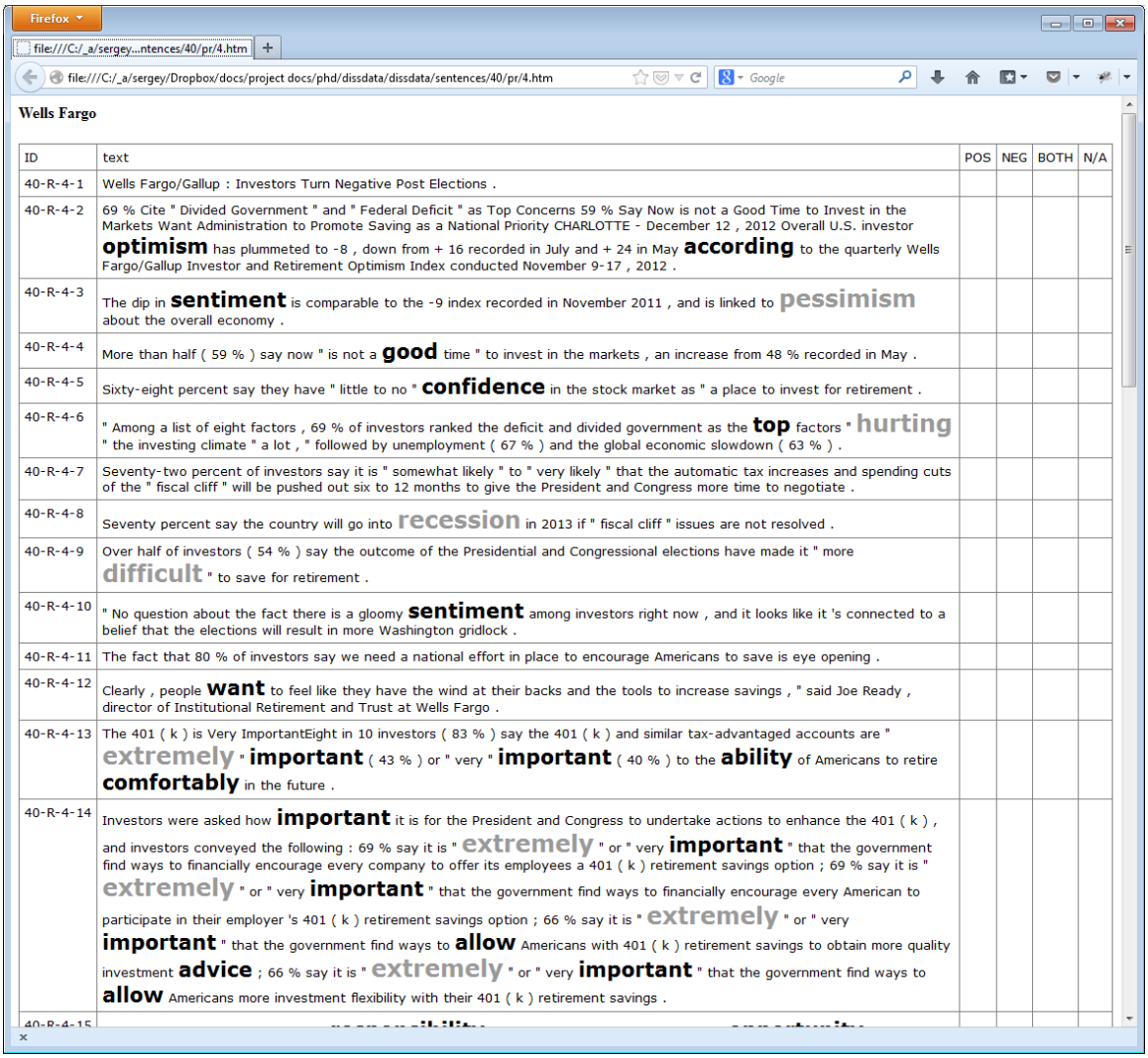
```
            f.write('\n\t<link rel="stylesheet" type="text/css" href="../../../styles.css">')
            f.write('\n</head>\n<body>\n')
            f.write(html)
            f.write('\n\n</body>\n</html>')
```

## analyze/matrix_maker.py

```
from __future__ import division
import sys
from data.matches import MatchLoader
from data.matches import MatchMaker
from data.tokens import TokenLoader
from data.scores import ScoreLoader
from data.articles import ArticleLoader

class MatrixMaker(object):

    def __init__(self, match_name):
        self._match_name = match_name


    def print_pairs(self):
        pass
#        sb = []
#        for company_id in range(1, 41):
#            matches = MatchLoader(company_id, self._match_name)
#            for release_id in matches.get_release_ids():



    def print_matrix(self):

        sb = []
        sb.append('co-id, rel-id, art-id, rel-len, art-len, rel-used, art-added, rel-subj-score, art-
subj-score, rel-sent-score, art-sent-score, pub, atrib\n')
#        sb.append('co-id rel-id art-id rel-len art-len rel-used art-added rel-subj-score art-subj-
score rel-sent-score art-sent-score\n')

        for company_id in range(1, 41):
            matches = MatchLoader(company_id, self._match_name)
            tokens = TokenLoader(company_id)
            scores = ScoreLoader(company_id)
            articles = ArticleLoader(company_id).get_articles()

            for release_id in matches.get_release_ids():
                rel_tokens = tokens.get_stripped_release_token_block(release_id, 0, sys.maxint)

                #release subjectivity score
                rel_subj = scores.count_subj_rel_sentences(release_id) /
scores.count_all_rel_sentences(release_id)
                #release sentiment score
                if scores.count_subj_rel_sentences(release_id) == 0:
                    rel_sents = 0
                else:
                    pos_minus_neg = scores.count_pos_rel_sentences(release_id) -
scores.count_neg_rel_sentences(release_id)
                    rel_sent = pos_minus_neg / scores.count_subj_rel_sentences(release_id)

                for article_id in matches.get_article_ids(release_id):
                    art_tokens = tokens.get_stripped_article_token_block(article_id, 0, sys.maxint)

                    blocks = matches.get_matches(release_id, article_id)
                    blocklen = 0
                    for b in blocks:
                        start = b[1]
                        length = b[2]
                        end = start + length
                        block_tokens = tokens.get_stripped_release_token_block(release_id, start, end)
                        blocklen += len(block_tokens)

                    rel_used = blocklen/len(rel_tokens)

                    art_added = 1 - blocklen/len(art_tokens)

                    #article subjectivity score
                    art_subj = scores.count_subj_art_sentences(article_id) /
scores.count_all_art_sentences(article_id)
                    #article sentiment score
```

```
                    if scores.count_subj_art_sentences(article_id) == 0:
                        art_sents = 0
                    else:
                        pos_minus_neg = scores.count_pos_art_sentences(article_id) -
scores.count_neg_art_sentences(article_id)
                        art_sent = pos_minus_neg / scores.count_subj_art_sentences(article_id)


                    pub = articles[article_id].pub_id()
                    atrib = self._has_attrib(articles[article_id].body())

                    sb.append('{0}, {1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}, {11},
{12}\n'.format( \
#                    sb.append('{0} {1} {2} {3} {4} {5} {6} {7} {8} {9} {10}\n'.format( \
                       company_id, release_id, article_id, len(rel_tokens), len(art_tokens),
rel_used, art_added, rel_subj, art_subj, rel_sent, art_sent, pub, atrib))

        text = ''.join(sb)
        print text


    def _has_attrib(self, body):
        body = body.lower()
        if 'press release' in body or 'news release' in body:
            return 1
        else:
            return 0
```

# APPENDIX E.
# AUTOMATED CODING OF EVALUATIVE LANGUAGE



**Wells Fargo**

| ID | text | POS | NEG | BOTH | N/A |
|---|---|---|---|---|---|
| 40-R-4-1 | Wells Fargo/Gallup : Investors Turn Negative Post Elections . | | | | |
| 40-R-4-2 | 69 % Cite " Divided Government " and " Federal Deficit " as Top Concerns 59 % Say Now is not a Good Time to Invest in the Markets Want Administration to Promote Saving as a National Priority CHARLOTTE - December 12 , 2012 Overall U.S. investor **optimism** has plummeted to -8 , down from + 16 recorded in July and + 24 in May **according** to the quarterly Wells Fargo/Gallup Investor and Retirement Optimism Index conducted November 9-17 , 2012 . | | | | |
| 40-R-4-3 | The dip in **sentiment** is comparable to the -9 index recorded in November 2011 , and is linked to **pessimism** about the overall economy . | | | | |
| 40-R-4-4 | More than half ( 59 % ) say now " is not a **good** time " to invest in the markets , an increase from 48 % recorded in May . | | | | |
| 40-R-4-5 | Sixty-eight percent say they have " little to no " **confidence** in the stock market as " a place to invest for retirement . | | | | |
| 40-R-4-6 | " Among a list of eight factors , 69 % of investors ranked the deficit and divided government as the **top** factors " **hurting** " the investing climate " a lot , " followed by unemployment ( 67 % ) and the global economic slowdown ( 63 % ) . | | | | |
| 40-R-4-7 | Seventy-two percent of investors say it is " somewhat likely " to " very likely " that the automatic tax increases and spending cuts of the " fiscal cliff " will be pushed out six to 12 months to give the President and Congress more time to negotiate . | | | | |
| 40-R-4-8 | Seventy percent say the country will go into **recession** in 2013 if " fiscal cliff " issues are not resolved . | | | | |
| 40-R-4-9 | Over half of investors ( 54 % ) say the outcome of the Presidential and Congressional elections have made it " more **difficult** " to save for retirement . | | | | |
| 40-R-4-10 | " No question about the fact there is a gloomy **sentiment** among investors right now , and it looks like it 's connected to a belief that the elections will result in more Washington gridlock . | | | | |
| 40-R-4-11 | The fact that 80 % of investors say we need a national effort in place to encourage Americans to save is eye opening . | | | | |
| 40-R-4-12 | Clearly , people **want** to feel like they have the wind at their backs and the tools to increase savings , " said Joe Ready , director of Institutional Retirement and Trust at Wells Fargo . | | | | |
| 40-R-4-13 | The 401 ( k ) is Very ImportantEight in 10 investors ( 83 % ) say the 401 ( k ) and similar tax-advantaged accounts are " **extremely** " **important** ( 43 % ) or " very " **important** ( 40 % ) to the **ability** of Americans to retire **comfortably** in the future . | | | | |
| 40-R-4-14 | Investors were asked how **important** it is for the President and Congress to undertake actions to enhance the 401 ( k ) , and investors conveyed the following : 69 % say it is " **extremely** " or " very **important** " that the government find ways to financially encourage every company to offer its employees a 401 ( k ) retirement savings option ; 69 % say it is " **extremely** " or " very **important** " that the government find ways to financially encourage every American to participate in their employer 's 401 ( k ) retirement savings option ; 66 % say it is " **extremely** " or " very **important** " that the government find ways to **allow** Americans with 401 ( k ) retirement savings to obtain more quality investment **advice** ; 66 % say it is " **extremely** " or " very **important** " that the government find ways to **allow** Americans more investment flexibility with their 401 ( k ) retirement savings . | | | | |

*Figure E1*. Screenshot of a coded press release.

*Note*. Terms coded as positive are marked in bold and use a larger font size. Terms coded as negative are highlighted in grey and use a larger font size. The application uses the same font size and colors (green for positive, red for negative); this adjustment is made for printing. Manual coding was done using the same pages, except the evaluative terms coded by a computer were not highlighted in any way. Scoring based on the results of automated coding was executed based on counts of positive and negative terms by sentence. The counts for the visible sentences displayed in Figure E1 are displayed in Figure E2.

```
doc-id=4 sent-id=1 pos=0 neg=0
doc-id=4 sent-id=2 pos=2 neg=0
doc-id=4 sent-id=3 pos=1 neg=1
doc-id=4 sent-id=4 pos=1 neg=0
doc-id=4 sent-id=5 pos=1 neg=0
doc-id=4 sent-id=6 pos=1 neg=1
doc-id=4 sent-id=7 pos=0 neg=0
doc-id=4 sent-id=8 pos=0 neg=1
doc-id=4 sent-id=9 pos=0 neg=1
doc-id=4 sent-id=10 pos=1 neg=0
doc-id=4 sent-id=11 pos=0 neg=0
doc-id=4 sent-id=12 pos=1 neg=0
doc-id=4 sent-id=13 pos=4 neg=1
doc-id=4 sent-id=14 pos=8 neg=4
```

*Figure E2*. Partial result of automatic coding the text displayed in Figure E1.

Table E1. *Top 20 terms coded as subjective.*

| | Positive Terms | | Negative Terms | |
|---|---|---|---|---|
| | Press Releases | News Articles | Press Releases | News Articles |
| Common Terms | help | help | vice | vice |
| | well | well | risk | risk |
| | support | support | loss | loss |
| | just | just | least | least |
| | good | good | too | too |
| | better | better | crisis | crisis |
| | best | best | low | low |
| | agreement | agreement | division | division |
| | strong | strong | against | against |
| | interest | interest | | |
| | want | want | | |
| | able | able | | |
| | potential | potential | | |
| | top | top | | |
| Unique Terms | ability | deal | differ | cut |
| | great | according | decrease | problems |
| | important | even | limited | stake |
| | benefits | profit | decline | problem |
| | commitment | free | disease | rival |
| | innovative | large | need | close |
| | | | hard | despite |
| | | | difficult | concerns |
| | | | unfavorable | little |
| | | | failure | failed |
| | | | disruption | slightly |

# APPENDIX F.
# SCREENSHOT DEMONSTRATING MATCHING TEXT

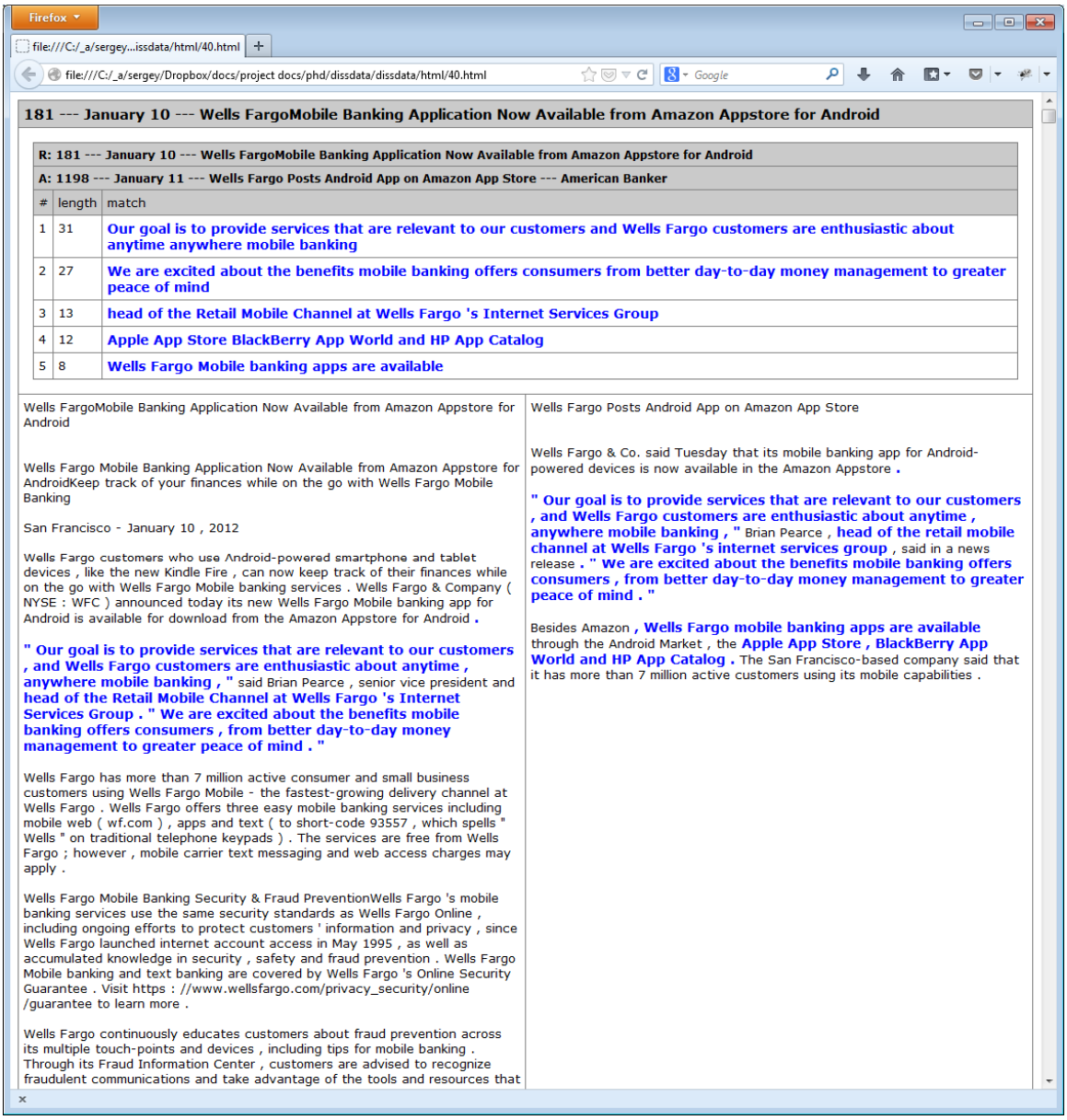



*Figure F1*. Matching text: press release (left) and news article (right).

APPENDIX G.
PROPORTIONS OF PRESS RELEASE AND NEWS ARTICLE CONTENT

**Ninety Percent Used and Ten Percent Added**

**Company:** Lockheed Martin
**Publication:** Alamogordo Daily News (New Mexico)
**Date of press release:** November 29, 2012
**Date of article:** December 6, 2012

Table G1. *Matching text spans (90%/10%).*

| # | Length | Matching Text |
|---|---|---|
| 1 | 196 | and modern hardware and software ensure high reliability rates and dramatically reduced operational and support costs The MFCR is an X-band solid-state active electronically scanned array radar which provides precision tracking and wideband discrimination and classification capabilities For extremely rapid deployments the MEADS MFCR can provide both surveillance and fire control capabilities until a surveillance radar joins the network An advanced identify friend-or-foe subsystem supports improved passive threat identification and typing Using its 360-degree defensive capability the advanced MEADS radars and PAC-3 MSE Missile MEADS defends up to eight times the coverage area with far fewer system assets and significantly reduces demand for deployed personnel and equipment which reduces demand for airlift MEADS successfully completed its first flight test on November 17 2011 against a simulated target attacking from behind A PAC-3 MSE Certified Missile Round was employed during the test along with the MEADS lightweight launcher and battle manager MEADS International a multinational joint venture headquartered in Orlando Fla. is the prime contractor for the MEADS system |
| 2 | 127 | System MEADS detected tracked intercepted and destroyed an air-breathing target in its first-ever intercept flight test today at White Sands Missile Range N.M The test achieved all criteria for success MEADS is a next-generation ground-mobile air and missile defense system that incorporates 360-degree radars netted and distributed battle management easily transportable launchers and the hit-to-kill PAC-3 Missile Segment Enhancement MSE Missile The system combines superior battlefield protection with new flexibility to protect forces and critical assets against tactical ballistic missiles cruise missiles unmanned aerial vehicles and aircraft The MEADS test configuration included a networked MEADS battle manager lightweight launcher firing a PAC-3 MSE Certified Missile Round |

(continued)

| | | |
|---|---|---|
| 3 | 106 | Manager Gregory Kee MEADS is proving its capability to defend our warfighters and key assets against a growing 21st century threat The test exploited the MEADS capability for full-perimeter 360-degree defense with the PAC-3 MSE Missile performing a unique over-the-shoulder maneuver to defeat the target attacking from behind the MEADS emplacement MEADS provides advanced capabilities that detect track and intercept evolving threats from farther away and without blind spots said MEADS International President Dave Berganini Today 's successful intercept proves MEADS advertised capabilities are real Its digital designs |
| 4 | 50 | MQM-107 target and guided the missile to a successful intercept Today 's successful flight test further demonstrates MEADS ability to identify track engage and defeat targets attacking from any direction using a single mobile launcher said NATO MEADS Management Agency General |
| 5 | 15 | and a 360-degree MEADS Multifunction Fire Control Radar MFCR which tracked the |

Table G2. *Press release and news article texts (90%/10%).*

| Press Release | News Article |
|---|---|
| MEADS Successfully Intercepts Air-Breathing Target at White Sands Missile Range | Medium Extended Air Defense System 's successful mission |
| ORLANDO/MUNICH/ROME , November 29 , 2012 - The Medium Extended Air Defense **System ( MEADS ) detected , tracked , intercepted and destroyed an air-breathing target in its first-ever intercept flight test today at White Sands Missile Range , N.M . The test achieved all criteria for success .** | ORLANDO/MUNICH/ROME , -- The Medium Extended Air Defense **System ( MEADS ) detected , tracked , intercepted and destroyed an air-breathing target in its first-ever intercept flight test today at White Sands Missile Range , N.M . The test achieved all criteria for success .** |
| **MEADS is a next-generation , ground-mobile air and missile defense system that incorporates 360-degree radars , netted and distributed battle management , easily transportable launchers and the hit-to-kill PAC-3 Missile Segment Enhancement ( MSE ) Missile . The system combines superior battlefield protection with new flexibility to protect forces and critical assets against tactical ballistic missiles , cruise missiles , unmanned aerial vehicles and aircraft .** | **MEADS is a next-generation , ground-mobile air and missile defense system that incorporates 360-degree radars , netted and distributed battle management , easily transportable launchers and the hit-to-kill PAC-3 Missile Segment Enhancement ( MSE ) Missile . The system combines superior battlefield protection with new flexibility to protect forces and critical assets against tactical ballistic missiles , cruise missiles , unmanned aerial vehicles and aircraft .** |

(continued)

**The MEADS test configuration included a networked MEADS battle manager , lightweight launcher firing a PAC-3 MSE Certified Missile Round and a 360-degree MEADS Multifunction Fire Control Radar ( MFCR ) , which tracked the MQM-107 target and guided the missile to a successful intercept .**

**" Today 's successful flight test further demonstrates MEADS ' ability to identify , track , engage and defeat targets attacking from any direction using a single mobile launcher , " said NATO MEADS Management Agency General Manager Gregory Kee . " MEADS is proving its capability to defend our warfighters and key assets against a growing 21st century threat . "**

**The test exploited the MEADS capability for full-perimeter , 360-degree defense with the PAC-3 MSE Missile performing a unique over-the-shoulder maneuver to defeat the target attacking from behind the MEADS emplacement .**

**" MEADS provides advanced capabilities that detect , track and intercept evolving threats from farther away and without blind spots , " said MEADS International President Dave Berganini . " Today 's successful intercept proves MEADS ' advertised capabilities are real . Its digital designs and modern hardware and software ensure high reliability rates and dramatically reduced operational and support costs . "**

**The MFCR is an X-band , solid-state , active electronically scanned array radar which provides precision tracking and wideband discrimination and classification capabilities . For extremely rapid deployments , the MEADS MFCR can provide both surveillance and fire control capabilities until a surveillance radar joins the network . An advanced identify friend-or-foe subsystem supports improved passive threat identification and typing .**

**The MEADS test configuration included a networked MEADS battle manager , lightweight launcher firing a PAC-3 MSE Certified Missile Round , and a 360-degree MEADS Multifunction Fire Control Radar ( MFCR ) , which tracked the**

**MQM-107 target and guided the missile to a successful intercept .**

**" Today 's successful flight test further demonstrates MEADS ' ability to identify , track , engage and defeat targets attacking from any direction using a single mobile launcher , " said NATO MEADS Management Agency General**

**Manager Gregory Kee . " MEADS is proving its capability to defend our warfighters and key assets against a growing 21st century threat . "**

**The test exploited the MEADS capability for full-perimeter , 360-degree defense with the PAC-3 MSE Missile performing a unique over-the-shoulder maneuver to defeat the target attacking from behind the MEADS emplacement .**

**" MEADS provides advanced capabilities that detect , track and intercept evolving threats from farther away and without blind spots , " said MEADS International President Dave Berganini . " Today 's successful intercept proves MEADS ' advertised capabilities are real . Its digital designs , and modern hardware and software ensure high reliability rates and dramatically reduced operational and support costs . "**

**The MFCR is an X-band , solid-state , active electronically scanned array radar which provides precision tracking and wideband discrimination and classification capabilities . For extremely rapid deployments , the MEADS MFCR can provide both surveillance and fire control capabilities until a surveillance radar joins the network . An advanced identify friend-or-foe subsystem supports improved passive threat identification and typing .**

(continued)

**Using its 360-degree defensive capability , the advanced MEADS radars and PAC-3 MSE Missile , MEADS defends up to eight times the coverage area with far fewer system assets and significantly reduces demand for deployed personnel and equipment , which reduces demand for airlift .**

**MEADS successfully completed its first flight test on November 17 , 2011 , against a simulated target attacking from behind . A PAC-3 MSE Certified Missile Round was employed during the test along with the MEADS lightweight launcher and battle manager .**

**MEADS International , a multinational joint venture headquartered in Orlando , Fla. , is the prime contractor for the MEADS system .** Major subcontractors and joint venture partners are MBDA in Italy and Germany , and Lockheed Martin in the United States .

The MEADS program management agency NAMEADSMA is located in Huntsville , Ala.

**Using its 360-degree defensive capability , the advanced MEADS radars and PAC-3 MSE Missile , MEADS defends up to eight times the coverage area with far fewer system assets and significantly reduces demand for deployed personnel and equipment , which reduces demand for airlift .**

**MEADS successfully completed its first flight test on November 17 , 2011 , against a simulated target attacking from behind . A PAC-3 MSE Certified Missile Round was employed during the test along with the MEADS lightweight launcher and battle manager .**

**MEADS International , a multinational joint venture headquartered in Orlando , Fla. , is the prime contractor for the MEADS system .**

Major subcontractors and joint venture partners are MBDA in Italy and Germany , and Lockheed Martin in the United States .

The MEADS program management agency NAMEADSMA is located in Huntsville , Ala.

**Seventy Percent Used and Ten Percent Added**

**Company:** Lockheed Martin
**Publication:** Mississippi Business Journal (Jackson, MS)
**Date of press release:** May 24, 2012
**Date of article:** May 28, 2012

Table G3. *Matching text spans (70%/10%).*

| # | Length | Matching Text |
|---|---|---|
| 1 | 97 | Based on lessons learned from the first two SBIRS geosynchronous satellites production of GEO-3 and GEO-4 is proceeding very well In addition we have a number of affordability initiatives in place jointly with the Air Force to continually reduce the cost of each follow-on SBIRS satellite Lockheed Martin engineers and technicians will now integrate the propulsion subsystem with the core structure which is essential for maneuvering the satellite during transfer orbit to its final location as well as conducting on-orbit repositioning maneuvers throughout its mission life |
| 2 | 80 | The structure was delivered to Lockheed Martin 's Mississippi Space Technology Center where engineers and technicians will integrate the spacecraft 's propulsion subsystem Featuring a mix of GEO satellites hosted payloads in highly elliptical earth orbit and associated ground hardware and software SBIRS delivers resilient and improved missile warning capabilities for the nation while simultaneously providing significant contributions to the military 's missile defense technical intelligence and battlespace awareness mission areas |
| 3 | 63 | The team expects to receive funding to begin long lead parts procurement for the fifth and sixth GEO satellite by the end of the year Additionally under the Air Force 's Overhead Persistent Infrared OPIR Space Modernization Initiative SMI Lockheed Martin will evolve technologies to improve capability and affordability for future SBIRS spacecraft |
| 4 | 55 | The integrated core propulsion module will then be shipped to Sunnyvale Calif. for final assembly integration and test SBIRS GEO-4 is on schedule to be available for launch in 2015 Lockheed Martin 's SBIRS contracts include four highly elliptical orbiting HEO payloads four GEO satellites |
| 5 | 44 | The GEO-4 structure identical to the previous three SBIRS GEO spacecraft is made from lightweight high-strength composite materials designed to withstand the accelerations and vibrations generated during launch and support the spacecraft throughout on-orbit operations |
| 6 | 16 | has received the core structure for the U.S. Air Force 's fourth Space Based Infrared System |
| 7 | 15 | Louie Lombardo Director of Lockheed Martin 's SBIRS Follow-on Production SFP program |

Table G4. *Press release and news article texts (70%/10%).*

| Press Release | News Article |
|---|---|
| Lockheed Martin Delivers Core Structure for Fourth SBIRS Satellite<br><br>STENNIS , Miss. , May 24 , 2012 - Lockheed Martin [ NYSE : LMT ] **has received the core structure for the U.S. Air Force 's fourth Space Based Infrared System** ( SBIRS ) geosynchronous satellite ( GEO-4 ) . **The structure was delivered to Lockheed Martin 's Mississippi Space & Technology Center , where engineers and technicians will integrate the spacecraft 's propulsion subsystem .**<br><br>**Featuring a mix of GEO satellites , hosted payloads in highly elliptical earth orbit , and associated ground hardware and software , SBIRS delivers resilient and improved missile warning capabilities for the nation while simultaneously providing significant contributions to the military 's missile defense , technical intelligence and battlespace awareness mission areas .**<br><br>**The GEO-4 structure , identical to the previous three SBIRS GEO spacecraft , is made from lightweight , high-strength composite materials designed to withstand the accelerations and vibrations generated during launch and support the spacecraft throughout on-orbit operations .**<br><br>" Delivery of the SBIRS GEO-4 core structure is a major milestone indicating the program is continuing to meet its commitments , " said **Louie Lombardo , Director of Lockheed Martin 's SBIRS Follow-on Production ( SFP ) program** . " **Based on lessons learned from the first two SBIRS geosynchronous satellites , production of GEO-3 and GEO-4 is proceeding very well . In addition , we have a number of affordability initiatives in place jointly with the Air Force to continually reduce the cost of each follow-on SBIRS satellite .**<br>" | Lockheed Martin 's Mississippi Space & Technology Center receives core structure for GEO-4 satellite<br><br>Lockheed Martin **has received the core structure for the U.S. Air Force 's fourth Space Based Infrared System** geosynchronous satellite ( GEO-4 ) .<br><br>**The structure was delivered to Lockheed Martin 's Mississippi Space & Technology Center , where engineers and technicians will integrate the spacecraft 's propulsion subsystem .**<br><br>**Featuring a mix of GEO satellites , hosted payloads in highly elliptical earth orbit , and associated ground hardware and software , SBIRS delivers resilient and improved missile warning capabilities for the nation while simultaneously providing significant contributions to the military 's missile defense , technical intelligence and battlespace awareness mission areas** , according to Lockheed Martin **.**<br><br>**The GEO-4 structure , identical to the previous three SBIRS GEO spacecraft , is made from lightweight , high-strength composite materials designed to withstand the accelerations and vibrations generated during launch and support the spacecraft throughout on-orbit operations .**<br><br>**Louie Lombardo , director of Lockheed Martin 's SBIRS Follow-on Production ( SFP ) program** , said , **" Based on lessons learned from the first two SBIRS geosynchronous satellites , production of GEO-3 and GEO-4 is proceeding very well . In addition , we have a number of affordability initiatives in place jointly with the Air Force to continually reduce the cost of each follow-on SBIRS satellite . "** |

(continued)

**Lockheed Martin engineers and technicians will now integrate the propulsion subsystem with the core structure , which is essential for maneuvering the satellite during transfer orbit to its final location , as well as conducting on-orbit repositioning maneuvers throughout its mission life . The integrated core propulsion module will then be shipped to Sunnyvale , Calif. , for final assembly , integration and test . SBIRS GEO-4 is on schedule to be available for launch in 2015 .**

**Lockheed Martin 's SBIRS contracts include four highly elliptical orbiting ( HEO ) payloads , four GEO satellites** , and ground assets to receive , process , and disseminate the infrared mission data . **The team expects to receive funding to begin long lead parts procurement for the fifth and sixth GEO satellite by the end of the year . Additionally , under the Air Force 's Overhead Persistent Infrared ( OPIR ) Space Modernization Initiative ( SMI ) , Lockheed Martin will evolve technologies to improve capability and affordability for future SBIRS spacecraft .**

The SBIRS team is led by the Infrared Space Systems Directorate at the U.S. Air Force Space and Missile Systems Center . Lockheed Martin is the SBIRS prime contractor , Northrop Grumman is the payload integrator . Air Force Space Command operates the SBIRS system .

Headquartered in Bethesda , Md. , Lockheed Martin is a global security and aerospace company that employs about 123,000 people worldwide and is principally engaged in the research , design , development , manufacture , integration and sustainment of advanced technology systems , products and services . The Corporation 's net sales for 2011 were $ 46.5 billion .

**Lockheed Martin engineers and technicians will now integrate the propulsion subsystem with the core structure , which is essential for maneuvering the satellite during transfer orbit to its final location , as well as conducting on-orbit repositioning maneuvers throughout its mission life .**

**The integrated core propulsion module will then be shipped to Sunnyvale , Calif. , for final assembly , integration and test . SBIRS GEO-4 is on schedule to be available for launch in 2015 .**

**Lockheed Martin 's SBIRS contracts include four highly elliptical orbiting ( HEO ) payloads , four GEO satellites** and ground assets to receive , process and disseminate the infrared mission data .

**The team expects to receive funding to begin long lead parts procurement for the fifth and sixth GEO satellite by the end of the year . Additionally , under the Air Force 's Overhead Persistent Infrared ( OPIR ) Space Modernization Initiative ( SMI ) , Lockheed Martin will evolve technologies to improve capability and affordability for future SBIRS spacecraft .**

**Fourty Eight Percent Used and Ten Percent Added**

**Company:** Abbott Laboratories
**Publication:** Chicago Daily Herald
**Date of press release:** September 21, 2012
**Date of article:** October 1, 2012

Table G5. *Matching text spans (48%/10%).*

| # | Length | Matching Text |
|---|---|---|
| 1 | 80 | Volwiler Society Established in 1985 and named for the late Ernest H. Volwiler Ph.D. an internationally recognized scientist and former Abbott president and chairman of the board the society encourages professional growth and is meant to spur recognize and showcase scientific innovation at Abbott Each year the Volwiler Society issues Outstanding Research awards to the Abbott scientists and teams who have made significant contributions to their fields |
| 2 | 78 | 15 overall receiving high marks for important quality research social responsibility and employee loyalty The complete rankings are available online and in the Oct. 19 2012 issue of Science This is the ninth time Abbott has been ranked among the top companies in the survey Scientific innovation is the driving force behind Abbott 's mission to discover new ways to help patients manage their health said |
| 3 | 77 | Scientists working at Abbott are given opportunities to succeed lead and grow in their careers through training mentoring networking groups and development programs In addition a number of internal recognition efforts including chairman 's awards president 's awards and patent/inventor awards work to highlight scientific excellence and contribution throughout the company The company recognizes its most distinguished scientists and engineers with induction into the |
| 4 | 48 | We strive to create an environment that values invention creativity and collaboration which helps top scientists to conduct important work while building interesting and meaningful careers at Abbott We are honored by this recognition from the scientific community as a benchmark of our success |
| 5 | 18 | annual survey identifies the most respected employers in the biotechnology and pharmaceutical industry Abbott ranked |
| 6 | 12 | named one of the Top 20 Employers by the journal Science |
| 7 | 8 | Based on the views of working scientists |

Table G6. *Press release and news article texts (48%/10%).*

| Press Release | News Article |
|---|---|
| The JournalScienceAgain Recognizes Abbott as One of the Top Employers in the Biotech and Pharmaceutical Industry<br><br>Date : September 21 , 2012<br><br>Abbott Park , Illinois ( NYSE : ABT ) a Abbott today was again **named one of the Top 20 Employers by the journal Science . Based on the views of working scientists ,** this prestigious **annual survey identifies the most respected employers in the biotechnology and pharmaceutical industry .**<br><br>**Abbott ranked** No . **15 overall , receiving high marks for important , quality research , social responsibility and employee loyalty . The complete rankings are available online and in the Oct. 19 , 2012 issue of Science . This is the ninth time Abbott has been ranked among the top companies in the survey .**<br><br>**" Scientific innovation is the driving force behind Abbott 's mission to discover new ways to help patients manage their health , " said** John Leonard , M.D. , senior vice president , Pharmaceuticals , Research and Development , Abbott . **" We strive to create an environment that values invention , creativity and collaboration , which helps top scientists to conduct important work while building interesting and meaningful careers at Abbott . We are honored by this recognition from the scientific community as a benchmark of our success** . "<br><br>**Scientists working at Abbott are given opportunities to succeed , lead and grow in their careers through training , mentoring , networking groups and development programs . In addition , a number of internal recognition efforts , including chairman 's awards , president 's awards and patent/inventor awards , work to highlight scientific excellence and contribution throughout the company .**<br>**The company recognizes its most distinguished scientists and engineers with induction into the** prestigious **Volwiler Society . Established in 1985** | Abbott recognized as top employer in industry<br><br>LIBERTYVILLE TOWNSHIP -- Abbott Laboratories was **named one of the Top 20 Employers by the journal Science .**<br><br>**Based on the views of working scientists ,** the **annual survey identifies the most respected employers in the biotechnology and pharmaceutical industry .**<br><br>**Abbott ranked 15 overall , receiving high marks for important , quality research , social responsibility and employee loyalty . The complete rankings are available online and in the Oct. 19 , 2012 issue of Science . This is the ninth time Abbott has been ranked among the top companies in the survey .**<br><br>**" Scientific innovation is the driving force behind Abbott 's mission to discover new ways to help patients manage their health , " said** Dr. John Leonard , senior vice president , pharmaceuticals , research and development at Abbott .<br><br>**" We strive to create an environment that values invention , creativity and collaboration , which helps top scientists to conduct important work while building interesting and meaningful careers at Abbott . We are honored by this recognition from the scientific community as a benchmark of our success** , " Leonard said .<br><br>**Scientists working at Abbott are given opportunities to succeed , lead and grow in their careers through training , mentoring , networking groups and development programs . In addition , a number of internal recognition efforts , including chairman 's awards , president 's awards and patent/inventor awards , work to highlight scientific excellence and contribution throughout the company .**<br><br>**The company recognizes its most distinguished scientists and engineers with induction into the Volwiler Society . Established in 1985 and named** |

(continued)

**and named for the late Ernest H. Volwiler , Ph.D. , an internationally recognized scientist and former Abbott president and chairman of the board , the society encourages professional growth , and is meant to spur , recognize and showcase scientific innovation at Abbott . Each year , the Volwiler Society issues " Outstanding Research " awards to the Abbott scientists and teams who have made significant contributions to their fields .**

About the 2012 Science Survey

The annual survey polls biotechnology , biopharmaceutical , pharmaceutical , and related industries to determine the 20 best employers in these industries as well as their driving characteristics . Respondents to the web-based survey were asked to rate companies based on 23 characteristics , including leadership and direction , work culture/environment , and academic and intellectual challenge . The 2012 rankings were determined based on 4,276 survey responses from readers of Science and other survey invitees . The study was conducted and rankings were determined by an independent research organization .

Abbott Widely Recognized as a Great Place to Work

In addition to being recognized as a top employer by Science , Abbott has been honored by The Scientist magazine as a Top BioPharma Employer nine times since 2003 . Abbott also has been honored for workplace leadership in more than 25 countries around the world . The company has been included on the Working Mother Best Companies list for 12 consecutive years , on DiversityInc magazine 's list of the top companies for diversity for nine years , and Hispanic Business magazine 's list of the Top Companies for Diversity Practices for six years . Additionally , FORTUNE has named Abbott as one of " America 's Most Admired Companies " every year since the list 's inception in 1983 .

About Abbott
Abbott ( NYSE : ABT ) is a global , broad-based health care company devoted to the discovery , development , manufacturing and marketing of pharmaceuticals and medical products , including nutritionals , devices and diagnostics . The company employs approximately 91,000 people and markets its products in more than 130 countries .

**for the late Ernest H. Volwiler , Ph.D. , an internationally recognized scientist and former Abbott president and chairman of the board , the society encourages professional growth , and is meant to spur , recognize and showcase scientific innovation at Abbott . Each year , the Volwiler Society issues " Outstanding Research " awards to the Abbott scientists and teams who have made significant contributions to their fields .**

PUBLICATION-TYPE : Newspaper

**Thirty Percent Used and Sixty Seven Percent Added**

**Company:** Ford Motor
**Publication:** Detroit Free Press (Michigan)
**Date of press release:** October 29, 2012
**Date of article:** October 29, 2012

Table G7. *Matching text spans (30%/67%).*

| # | Length | Matching Text |
|---|---|---|
| 1 | 33 | From the start our eye was on what was required to transform these operations into businesses that would attract the world 's best suppliers needed to move Ford 's business forward |
| 2 | 33 | This new joint venture will enable Valeo to broaden its offering of innovative thermal system products and support its customers in meeting their CO2 emissions reduction challenges |
| 3 | 32 | Detroit Thermal Systems will provide Ford with an experienced high-quality climate control supplier and contribute significantly to the renaissance of manufacturing in the Detroit region said |
| 4 | 28 | The combined capabilities of two world-class suppliers will provide customers with a capable long-term partner committed to support their needs today and for the future |
| 5 | 21 | Few companies take the longer-term comprehensive approach we took with the restructuring of ACH |
| 6 | 20 | and at the same time preserve as many jobs as possible We are proud of what we accomplished |
| 7 | 18 | intends to apply for certification as a minority business enterprise through the Michigan Minority Business Development Council |
| 8 | 16 | The sale also expands Ford 's business with minority-owned suppliers V. Johnson Enterprises |
| 9 | 9 | from the Sheldon Road Plant to a new DTS |
| 10 | 7 | climate control business to Detroit Thermal Systems |

Table G8. *Press release and news article texts (30%/67%).*

| Press Release | News Article |
|---|---|
| Sale Announced for Last Remaining ACH Operation - Climate Control Business<br><br>Ford Motor Company and Automotive Components Holdings , LLC ( ACH ) today announced the sale of ACH 's **climate control business to Detroit Thermal Systems** , LLC ( DTS )<br><br>This sale involves the last remaining automotive components operation and is the 10th sale of an ACH operation or plant<br><br>This sale marks the culmination of the ACH strategy and the successful transition of ACH-managed operations to key strategic automotive suppliers<br><br>DEARBORN , Mich. , Oct. 29 , 2012 - Ford Motor Company and Automotive Components Holdings , LLC ( ACH ) today announced the sale of the last remaining ACH operation - the climate control business currently located at the ACH Sheldon Road Plant in Plymouth Township , Mich. Ford and ACH have signed definitive agreements for the sale of the climate control business to Detroit Thermal Systems , LLC ( DTS ) , a joint venture between Valeo and V. Johnson Enterprises , LLC . V. Johnson Enterprises is owned by Vincent Johnson , a Detroit entrepreneur and former Detroit Pistons basketball star .<br><br>This announcement marks the fulfillment of the ACH strategy and the 10th ACH sale of an operation or plant . ACH was started in October 2005 with 17 automotive components plants .<br><br>The initial closing of the sales transaction has taken place , contingent upon the approval of state and local incentives . Asset and operation transfers **from the Sheldon Road Plant to a new DTS** manufacturing facility in Romulus , Mich. , will start in mid-2013 and conclude by the end of 2014 **.** | Ford to sell **climate control business to Detroit Thermal Systems**<br><br>Oct. 29--Ford announced today an agreement to sell the last of its Automotive Components Holdings ( ACH ) businesses , a divestiture that has been years in the making .<br><br>ACH 's climate control business in Plymouth Township is being purchased by Detroit Thermal Systems ( DTS ) , which is a new joint venture between French supplier Valeo and V. Johnson Enterprises , a Detroit supplier owned by former Piston basketball player Vincent Johnson .<br><br>The parties hope to transfer all assets and operations **from the Sheldon Road plant to a new DTS** facility in Romulus starting in mid-2013 , to be completed by the end of 2014 .<br><br>Production of climate control systems for Ford vehicles in North America is expected to start in the third quarter of 2013 and current ACH employees at the Sheldon Road plant can apply for jobs at the new DTS plant which wants a diverse workforce offering opportunities for minorities and military veterans .<br><br>" We are very proud to announce the creation of Detroit Thermal Systems . Our team is committed to manufacturing world-class climate control systems and to creating employment opportunities for the talented workforce in the Detroit area , " said Johnson , chief executive officer of Detroit Thermal Systems .<br><br>" The growth of our new organization will help to revitalize local communities by creating new job opportunities for minorities , military veterans and others hard hit during the economic crisis , " Johnson said . |

(continued)

**" Few companies take the longer-term , comprehensive approach we took with the restructuring of ACH , "** said Mark Fields , Ford president of The Americas . **" From the start , our eye was on what was required to transform these operations into businesses that would attract the world 's best suppliers needed to move Ford 's business forward - and at the same time , preserve as many jobs as possible . We are proud of what we accomplished** . "

**The sale also expands Ford 's business with minority-owned suppliers . V. Johnson Enterprises** holds a controlling interest in DTS , while Valeo has the balance of the ownership in the new company . DTS **intends to apply for certification as a minority business enterprise through the Michigan Minority Business Development Council .**

" With each of these ACH sales , we worked with buyers to ensure alignment on the evolution of our business relationship over time , " said Tony Brown , group vice president , Ford Global Purchasing . " We see these business relationships as integral to our Aligned Business Framework . "

**" Detroit Thermal Systems will provide Ford with an experienced , high-quality climate control supplier and contribute significantly to the renaissance of manufacturing in the Detroit region , " said** Vincent Johnson , chairman and chief executive officer of both V. Johnson Enterprises and DTS . **" The combined capabilities of two world-class suppliers will provide customers with a capable , long-term partner , committed to support their needs today and for the future** . "

ACH started in October 2005 with 17 components plants and now all will have been sold or closed . This is the 10th sale of ACH-managed operations .

Ford wanted to get out of the parts business to focus on its core business of building and selling cars , now down to two brands : Ford and Lincoln **.**

**" Few companies take the longer-term , comprehensive approach we took with the restructuring of ACH , "** Mark Fields , Ford president of The Americas , said in a statement .

**" From the start , our eye was on what was required to transform these operations into businesses that would attract the world 's best suppliers needed to move Ford 's business forward -- and at the same time , preserve as many jobs as possible . We are proud of what we accomplished** , " Fields said .

Ford and ACH have signed agreements to sell the Sheldon Road Plant to DTS . Finalizing the deal is contingent on approval of state and local incentives .

**The sale also expands Ford 's business with minority-owned suppliers . V. Johnson Enterprises** has a controlling interest in DTS which **intends to apply for certification as a minority business enterprise through the Michigan Minority Business Development Council .**

**" Detroit Thermal Systems will provide Ford with an experienced , high-quality climate control supplier and contribute significantly to the renaissance of manufacturing in the Detroit region , " said** Johnson .

(continued)

" This acquisition is a strategic breakthrough for Valeo that will not only strengthen our ties with Ford Motor Company in North America and the rest of the world , but also enhance our presence across North America , " said Jacques Aschenbroich , Valeo chief executive officer . " Valeo , in addition , will be a member of Ford 's Aligned Business Framework supplier program .

**" This new joint venture will enable Valeo to broaden its offering of innovative thermal system products and support its customers in meeting their CO2 emissions reduction challenges , "** he added .

# # #

About Ford Motor Company

Ford Motor Company , a global automotive industry leader based in Dearborn , Mich. , manufactures or distributes automobiles across six continents . With about 168,000 employees and about 65 plants worldwide , the company 's automotive brands include Ford and Lincoln . The company provides financial services through Ford Motor Credit Company . For more information regarding Ford and its products worldwide , please visit http : //corporate.ford.com .

**" The combined capabilities of two world-class suppliers will provide customers with a capable , long-term partner , committed to support their needs today and for the future** , " Johnson said .

Valeo sees the deal as " a strategic breakthrough , " that will strengthen its ties with Ford , said CEO Jacques Aschenbroich . It " also enhances our presence across North America , " he said .

**" This new joint venture will enable Valeo to broaden its offering of innovative thermal system products and support its customers in meeting their CO2 emissions reduction challenges , "** Aschenbroich said .

Contact Alisa Priddle at 313-222-5394 or apriddle @ freepress.com

**Twenty Percent Used and Sixty Four Percent Added**

**Company:** Wells Fargo
**Publication:** American Banker
**Date of press release:** June 25, 2012
**Date of article:** July 26, 2012

Table G9. *Matching text spans (20%/64%).*

| # | Length | Matching Text |
|---|---|---|
| 1 | 35 | This acquisition enhances our position in the marketplace and provides our clients with dedicated customer service as well as Wells Fargo 's strength stability and broad product set |
| 2 | 18 | We have been growing our subscription finance business organically for many years |
| 3 | 13 | Dee Dee Sklar former head of WestLB 's subscription finance group |
| 4 | 12 | letters of credit mainly to private equity and real estate investment funds |
| 5 | 9 | Julie Caperton head of Asset-Backed Finance and Securitization |
| 6 | 8 | will lead a team of 14 including |

Table G10. *Press release and news article texts (20%/64%).*

| Press Release | News Article |
|---|---|
| Wells Fargoto Acquire Fund Financing Portfolio from WestLB<br><br>Wells Fargo to Acquire Fund Financing Portfolio from WestLB $ 3 billion portfolio enhances Wells Fargo 's presence in subscription finance<br><br>CHARLOTTE - June 25 , 2012<br><br>Wells Fargo & Company ( NYSE : WFC ) today announced that it has reached a definitive agreement to acquire WestLB 's subscription finance portfolio . The portfolio contains approximately $ 6 billion in commitments ( approximately $ 3 billion outstanding ) . Terms of the agreement were not disclosed .<br><br>Subscription finance provides committed revolving and term loans as well as **letters of credit mainly to private equity and real estate investment funds** to facilitate the funds ' investment activities . The financing is secured by the uncalled capital | Wells Fargo to Acquire Subscription Finance Business<br><br>Wells Fargo ( WFC ) has agreed to buy a subscription finance portfolio with $ 6 billion in commitments from WestLB , European commercial bank based in Germany .<br><br>The San Francisco bank , like some of its competitors , has been opting to buy niche businesses , such as specialty lenders , insurance brokerages and asset-management firms , rather than whole banks . It has not made a whole bank acquisition in more than three years .<br><br>Subscription financing provides revolving and term loans in addition to **letters of credit mainly to private equity and real estate investment funds** . The financial terms of the deal , which is expected to close in the second quarter , were not disclosed , Wells Fargo said Monday **.** |

(continued)

commitments from the funds ' institutional investors .

**" We have been growing our subscription finance business organically for many years , "** said **Julie Caperton , head of Asset-Backed Finance and Securitization . " This acquisition enhances our position in the marketplace and provides our clients with dedicated customer service as well as Wells Fargo 's strength , stability and broad product set . "**

" Subscription finance clients include some of the industry 's strongest fund managers who have commitments from high quality institutional investors , " said Mary Katherine DuBose , head of Corporate Debt Finance . " In addition to continuing to provide subscription finance services to these clients , we look forward to being able to offer them our full suite of banking products and services . "

**Dee Dee Sklar , former head of WestLB 's subscription finance group ,** has been hired to run Wells Fargo 's subscription finance business and will report to DuBose . Sklar **will lead a team of 14 , including** 8 former WestLB employees .

The transaction is expected to close by the end of the second quarter of 2012 .

About Wells FargoWells Fargo & Company ( NYSE : WFC ) is a nationwide , diversified , community-based financial services company with $ 1.3 trillion in assets . Founded in 1852 and headquartered in San Francisco , Wells Fargo provides banking , insurance , investments , mortgage , and consumer and commercial finance through more than 9,000 stores , 12,000 ATMs , the Internet ( wellsfargo.com ) , and other distribution channels across North America and internationally . With more than 270,000 team members , Wells Fargo serves one in three households in America . Wells Fargo & Company was ranked No . 26 on Fortune 's 2012 rankings of America 's largest corporations . Wells Fargo 's vision is to satisfy all our customers ' financial needs and help them succeed financially .

MediaElise Wilkinsonelise.wilkinson @ wellsfargo.com704-374-6512

InvestorsJim Rowejim.rowe @ wellsfargo.com415-396-8216

**" We have been growing our subscription finance business organically for many years , "** **Julie Caperton , head of asset-backed finance and securitization** , said in a press release **. " This acquisition enhances our position in the marketplace and provides our clients with dedicated customer service as well as Wells Fargo 's strength , stability and broad product set . "**

Wells Fargo has hired **Dee Dee Sklar , former head of WestLB 's subscription finance group ,** to run its subscription finance business . Sklar will report to Mary Katherine DuBose , head of corporate debt finance for Wells Fargo , and **will lead a team of 14 , including** eight former WestLB employees .

## Ten Percent Used and Eighty Eight Percent Added

**Company:** Target
**Publication:** San Jose Mercury News (California)
**Date of press release:** December 14, 2012
**Date of article:** December 14, 2012

Table G11. *Matching text spans (10%/88%).*

| # | Length | Matching Text |
|---|---|---|
| 1 | 27 | California continues to be a strong market for Target and we 're excited to expand our presence there in 2013 |
| 2 | 24 | Target is committed to being a good neighbor and developing long-lasting relationships with guests and the Alameda community |

Table G12. *Press release and news article texts (10%/88%).*

| Press Release | News Article |
|---|---|
| Target to Open New Store in Alameda , Calif.<br><br>December 14 , 2012 MINNEAPOLIS<br><br>Target is pleased to announce plans to open a new store in Alameda , Calif. , in October 2013 . The new store will be located at the Alameda Landing Shopping Center at the intersection of Mariner Square Loop and Stargell Avenue . To date , Target has announced plans to open fifteen Target stores in 2013 .<br><br>The Alameda store will be approximately 140,000 square feet and will offer guests the everyday essentials and exclusive brands they have come to expect from Target . In addition , the store will include a selection of fresh produce , fresh packaged meat and pre-packaged baked goods to further enhance guests ' shopping experience . | Alameda : Target announces Alameda store<br><br>ALAMEDA -- The announcement that Target would open at Alameda Landing was initially made last year by Catellus Development Corp. , the master developer behind the mixed-use project now under way in the city 's West End .<br><br>But on Friday Target representatives helped make it official , saying the Alameda store will open in October next year and be one of 15 to open nationwide .<br><br>The store is expected to anchor the shopping center at Alameda Landing , which will also feature 275 homes . The overall project covers about 77 acres at what was once a supply center for the former Alameda Naval Air Station near the Oakland-Alameda Estuary . |

(continued)

The Alameda location will employ approximately 200 team members . Target will host job fairs approximately two months prior to the new store opening , at which prospective candidates may apply and interview for open team member positions . Candidates may also apply online at Target.com/careers or at in-store kiosks located in all Target stores approximately three months prior to the new store opening .

**" California continues to be a strong market for Target , and we 're excited to expand our presence there in 2013 , "** said Cary Strouse , Target 's senior vice president of stores in the Western region **. " Target is committed to being a good neighbor and developing long-lasting relationships with guests and the Alameda community . "**

Target creates strong partnerships with local organizations in all of the communities where the company does business through Target 's community giving programs . This store will start a local grant program , contribute to the United Way and donate food to a Feeding America member , or approved agency . Target also encourages team members to volunteer their time to serve the needs of their community .

About Target

Minneapolis-based Target Corporation ( NYSE : TGT ) serves guests at 1,782 stores across the United States and at Target.com . The company plans to open its first stores in Canada in 2013 . Since 1946 , Target has given 5 percent of its income through community grants and programs ; today , that giving equals more than $ 4 million a week . For more information about Target 's commitment to corporate responsibility , visit Target.com/hereforgood .

For more information , visit Target.com/pressroom .

The Alameda store will measure about 140,000 square feet and have about 200 employees .

Along with clothes and household items , the store will offer produce , packaged meat and pre-packaged baked goods **.**

**" California continues to be a strong market for Target , and we 're excited to expand our presence there in 2013 , "** Cary Strouse , a senior vice president with the chain , said in a statement **. " Target is committed to being a good neighbor and developing long-lasting relationships with guests and the Alameda community . "**

The Alameda store is also expected to generate up to $ 300,000 in annual sales tax revenue , while an additional $ 200,000 is expected from other businesses when the Alameda Landing shopping center is finished , according to city officials .

In January , the Planning Board approved tweaking the plan for Alameda Landing to help accommodate the Target store . But the board also opted to cap the amount of space dedicated inside the business for groceries and other non-taxable items to 10 percent for five years , aiming to prevent the chain from undermining smaller , neighborhood stores .

Before agreeing to open at Alameda Landing , Target was in talks to possibly open at Alameda South Shore Center , but that plan fell through in June 2007 .

Reach Peter Hegarty at 510-748-1654 or follow him onTwitter.com/Peter_Hegarty/ .

## One Percent Used and Ninety Six Percent Added

**Company:** American Express
**Publication:** Deseret Morning News (Salt Lake City)
**Date of press release:** March 6, 2012
**Date of article:** March 7, 2012

Table G13. *Matching text spans (1%/96%).*

| # | Length | Matching Text |
|---|---|---|
| 1 | 8 | at participating merchants the savings are automatically |

Table G14. *Press release and news article texts (1%/96%).*

| Press Release | News Article |
|---|---|
| American Express Turns Twitter # Hashtags into Couponless National Merchant Offers and Cardmember Savings Using Its Smart Offer APIs<br><br>NEW YORK , March 6 , 2012 --<br><br>American Express today introduced a new experience for US Cardmembers using Twitter , that turns customized Twitter # hashtags into savings . Now , American Express Cardmembers can sync their eligible Card with Twitter at sync.americanexpress.com/twitter , and when they tweet using special offer hashtags , couponless savings are loaded directly to their synced Cards - no coupons , no print-outs . Then , when Cardmembers use their synced Card for qualifying purchases online or in-store **at participating merchants , the savings are automatically** delivered via a statement credit within days . Participating merchant partners include Best Buy , McDonald 's , Whole Foods Market , Zappos and more . This unique user experience is facilitated by American Express ' Smart Offer APIs , which also enables American Express to provide detailed reporting to merchants about the online and offline spend behavior of their customers . | Credit card company pays customers to Tweet<br><br>Making money is now as simple as a tweet . American Express has launched a program that gives customers discounts for tweets , according to Mashable . Customers go to a website where they sync their cards to their Twitter accounts . Whenever they tweet special offer tags in the messages , customers receive savings loaded onto their cards without the need of a coupon . When customers go to buy items **at participating merchants , the savings are automatically** deducted . Best Buy , McDonald ? s , Whole Foods Market and Zappos have all signed on to the new programs . The trade-off of discounts for the customers and free marketing for American Express is the idea behind the program . This isn ? t the first time AmEx has used social networks . The company launched a ? Link , Like , Love ? campaign on Facebook last July that gave users offers if customers liked participating merchants . AmEx also offered deals through Foursquare , a location-based social network . |

(continued)

" American Express is turning Twitter content into commerce by connecting Cardmembers to merchants and delivering real world value to both , " said Ed Gilligan , Vice Chairman , American Express . " With the continued convergence of online and offline commerce , our closed loop continues to enable us to bring seamless , relevant ways to connect our cardmembers and merchants on the most powerful social and digital platforms . "

" Every day , millions of people use Twitter to get special offers from the brands and retailers they care about , " said Adam Bain , Twitter 's president of global revenue . " Now , American Express is making it even easier for people to act on those offers simply by sending Tweets with special offer hashtags from retailers . It 's exciting to see American Express build on Twitter in a way that benefits both consumers and retailers . "

Sync . Tweet . Save .

Cardmembers can tweet their way to savings by syncing an eligible American Express Card with Twitter at sync.americanexpress.com/twitter . To take advantage of the exclusive offers available from a variety of the nation 's largest merchants , Cardmembers tweet the special offer hashtags to load offers directly to their synced Cards . When the Cardmember uses that same Card to make a qualified purchase in-store or online with a participating merchant , an automatic statement credit is issued within days .

Participating merchants at launch include 1-800-FLOWERS.COM , Best Buy , Century 21 Department Store , The Cheesecake Factory , Dell , FedEx Office , FTD , Gulf , H&M ; , McDonald 's , Seamless.com , Sports Authority , Ticketmaster , Virgin America , Whole Foods Market and Zappos.com . All of the current American Express special offers are highlighted as " favorites " on the @ americanexpress Twitter page .

Read the entire article at Mashable . EMAIL :

jferguson @ desnews.com TWITTER : @ joeyferguson

(continued)

American Express created @ amexsync , an automated notification handle that is activated when someone sends a Tweet that includes one of the special offer hashtags . This sophisticated handle detects if a user is already synced and either confirms their offer enrollment or provides a link to first sync their American Express Card to enroll in the selected offer . Additionally , if an offer is no longer available , @ amexsync will direct users to check out the Tweets @ americanexpress has " favorited " to find the latest offers . Questions or additional assistance will be provided by @ askamex , the American Express customer service team on Twitter .

Experience Sync at SXSW

American Express will showcase the new Twitter experience during a special , nationally live-streamed concert from SXSW on March 12 at 7 p.m. CT featuring an award-winning recording artist . For details , please visit youtube.com/AmericanExpress . The Amex Sync Show will be broadcast live at youtube.com/americanexpress and also be available on the VEVO mobile and tablet platform .

Local and visiting eligible Cardmembers in Austin can take advantage of a special city-wide offer and get $ 10 when sync their Card , Tweet the special offer # AmexAustin10 and use their synced Card to spend in Austin during SXSW from March 9 to March 13 . Limit one statement credit per Cardmember . Click here for program terms .

About American Express

American Express is a global services company , providing customers with access to products , insights and experiences that enrich lives and build business success . Learn more at americanexpress.com and connect with us on facebook.com/americanexpress , foursquare.com/americanexpress , twitter.com/americanexpress , and youtube.com/americanexpress .

APPENDIX H.
AGGREGATE DATA BY PUBLICATION (RQ1 AND RQ2)

Table H1. *Aggregate data by publications (RQ1 and RQ2).*

| Publication | Used | Added | Cases |
|---|---|---|---|
| AIM West Milford (Passaic, North Jersey) | 0.80% | 98.30% | 1 |
| Alamogordo Daily News (New Mexico) | 45.94% | 54.02% | 2 |
| American Banker | 9.11% | 76.53% | 86 |
| Austin American-Statesman (Texas) | 6.59% | 88.88% | 4 |
| Automotive News | 4.30% | 97.37% | 38 |
| Bangor Daily News (Maine) | 7.75% | 91.40% | 2 |
| Brattleboro Reformer (Vermont) | 3.47% | 92.32% | 6 |
| Buffalo News (New York) | 4.01% | 96.32% | 10 |
| Business Insurance | 6.81% | 91.84% | 7 |
| Chapel Hill Herald (Durham, N.C.) | 1.21% | 98.59% | 1 |
| Charleston Daily Mail (West Virginia) | 6.47% | 90.41% | 6 |
| Chicago Daily Herald | 12.89% | 82.36% | 29 |
| Colorado Springs Business Journal (Col. Springs, CO) | 17.34% | 58.63% | 4 |
| Contra Costa Times (California) | 7.42% | 92.88% | 61 |
| Crain's Detroit Business | 7.59% | 96.25% | 9 |
| Daily Camera (Boulder, Colorado) | 1.37% | 94.80% | 5 |
| Daily Journal of Commerce (Portland, OR) | 3.70% | 89.93% | 1 |
| Daily News (New York) | 2.69% | 89.38% | 15 |
| Daily Variety | 8.29% | 79.67% | 14 |
| Dayton Daily News (Ohio) | 4.17% | 92.98% | 18 |
| Deseret Morning News (Salt Lake City) | 7.42% | 84.10% | 11 |
| Detroit Free Press (Michigan) | 3.75% | 93.12% | 47 |
| Education Week | 0.88% | 98.20% | 1 |
| Finance & Commerce (Minneapolis, MN) | 11.23% | 65.44% | 5 |
| Florida Times-Union (Jacksonville) | 1.99% | 94.99% | 9 |
| Government Technology | 6.78% | 58.30% | 1 |
| Hartford Courant (Connecticut) | 5.57% | 82.67% | 12 |
| Herald News (Passaic County, NJ) | 3.21% | 91.06% | 10 |

(continued)

| | | | |
|---|---|---|---|
| Inland Valley Daily Bulletin (Ontario, CA) | 2.49% | 94.10% | 6 |
| Intelligencer Journal/New Era (Lancaster, Pennsylvania) | 4.01% | 83.56% | 4 |
| Investment News | 1.63% | 95.32% | 4 |
| Investor's Business Daily | 3.20% | 95.39% | 46 |
| Journal of Commerce Online | 5.84% | 86.27% | 15 |
| Las Cruces Sun-News (New Mexico) | 11.79% | 57.52% | 4 |
| Las Vegas Review-Journal (Nevada) | 5.12% | 94.94% | 14 |
| Lewiston Morning Tribune (Idaho) | 3.54% | 96.82% | 13 |
| Long Island Business (Long Island, NY) | 9.32% | 88.96% | 7 |
| Los Angeles Times | 4.27% | 93.06% | 46 |
| Lowell Sun (Massachusetts) | 3.84% | 95.25% | 11 |
| Marin Independent Journal (California) | 1.09% | 98.74% | 1 |
| Maryland Gazette | 8.42% | 58.33% | 2 |
| Mississippi Business Journal (Jackson, MS) | 68.53% | 9.92% | 1 |
| Missouri Lawyers Media | 1.58% | 98.35% | 1 |
| Monterey County Herald (California) | 4.29% | 95.54% | 4 |
| Morning Call (Allentown, Pennsylvania) | 4.74% | 91.53% | 4 |
| Newsday (New York) | 2.17% | 97.89% | 4 |
| New York Observer | 2.89% | 95.95% | 2 |
| Orange County Register (California) | 3.51% | 95.39% | 9 |
| Oroville Mercury Register (California) | 0.09% | 98.98% | 1 |
| Palm Beach Post (Florida) | 6.15% | 94.91% | 6 |
| Pensions & Investments | 2.28% | 97.44% | 6 |
| Pittsburgh Post-Gazette | 3.48% | 94.08% | 22 |
| Pittsburgh Tribune Review | 2.96% | 96.05% | 6 |
| Plastics News | 6.67% | 90.26% | 7 |
| Providence Journal | 10.69% | 82.47% | 3 |
| Public Opinion (Chambersburg, Pennsylvania) | 2.56% | 94.89% | 2 |
| Richmond Times-Dispatch (Virginia) | 7.34% | 85.59% | 3 |
| Rubber & Plastics News | 10.92% | 91.74% | 2 |
| San Antonio Express-News | 4.76% | 95.42% | 7 |
| San Gabriel Valley Tribune (California) | 3.51% | 91.82% | 6 |
| San Jose Mercury News (California) | 7.50% | 92.04% | 82 |
| Sarasota Herald Tribune (Florida) | 5.91% | 88.52% | 5 |
| Sentinel & Enterprise (Fitchburg, Massachusetts) | 1.46% | 96.33% | 2 |
| South Bend Tribune (Indiana) | 6.53% | 90.18% | 2 |

(continued)

| | | | |
|---|---|---|---|
| South Florida Sun-Sentinel (Fort Lauderdale) | 1.95% | 96.73% | 4 |
| Spokesman Review (Spokane, WA) | 3.91% | 95.59% | 13 |
| Star-News (Wilmington, NC) | 3.01% | 95.99% | 5 |
| Star Tribune (Minneapolis, MN) | 7.25% | 91.70% | 29 |
| St. Louis Post-Dispatch (Missouri) | 3.70% | 97.11% | 5 |
| St. Paul Pioneer Press (Minnesota) | 4.36% | 92.23% | 1 |
| Suburban Trends (Morris, North Jersey) | 17.45% | 77.87% | 4 |
| Sunday News (Lancaster, Pennsylvania) | 13.59% | 85.47% | 1 |
| Tampa Bay Times | 2.06% | 96.82% | 15 |
| Tampa Tribune (Florida) | 0.95% | 98.59% | 1 |
| TELEGRAM & GAZETTE (Massachusetts) | 6.66% | 92.29% | 3 |
| Telegraph Herald (Dubuque, IA) | 3.16% | 77.11% | 6 |
| The Atlanta Journal-Constitution | 2.36% | 94.83% | 29 |
| The Augusta Chronicle (Georgia) | 4.65% | 79.74% | 5 |
| The Baltimore Sun | 2.67% | 96.49% | 11 |
| The Berkshire Eagle (Pittsfield, Massachusetts) | 4.49% | 92.67% | 10 |
| The Bismarck Tribune | 3.15% | 94.53% | 15 |
| The Bond Buyer | 6.36% | 92.65% | 6 |
| The Capital (Annapolis, MD) | 5.75% | 90.44% | 9 |
| The Christian Science Monitor | 11.27% | 87.07% | 1 |
| The Chronicle of Philanthropy | 1.90% | 99.35% | 1 |
| The Columbian (Vancouver, Washington) | 13.33% | 86.76% | 1 |
| The Daily News of Los Angeles | 1.89% | 97.51% | 2 |
| The Daily Oklahoman (Oklahoma City, OK) | 4.40% | 92.38% | 9 |
| The Daily Record (Baltimore, MD) | 11.54% | 90.59% | 33 |
| The Daily Record of Rochester (Rochester, NY) | 16.40% | 94.85% | 2 |
| The Daily Reporter (Milwaukee, WI) | 17.27% | 53.54% | 2 |
| The Daily Star-Journal, Warrensburg, Mo | 0.50% | 99.57% | 1 |
| The Dallas Morning News | 2.96% | 92.65% | 14 |
| The Deal Pipeline | 4.14% | 89.54% | 52 |
| The Denver Post | 7.03% | 86.52% | 3 |
| The Detroit News (Michigan) | 5.06% | 94.17% | 93 |
| The Dispatch (Gilroy, California) | 31.94% | 47.31% | 1 |
| The Evening Sun (Hanover, Pennsylvania) | 2.56% | 96.87% | 1 |
| The Gazette (Cedar Rapids, Iowa) | 11.13% | 90.95% | 3 |

(continued)

| | | | |
|---|---|---|---|
| The Herald Bulletin (Anderson, Indiana) | 17.99% | 83.56% | 1 |
| The Herald-Sun (Durham, N.C.) | 7.91% | 94.63% | 9 |
| The Hill | 13.95% | 83.67% | 2 |
| The Houston Chronicle | 5.29% | 94.62% | 35 |
| The Indianapolis Business Journal | 1.96% | 97.80% | 5 |
| The Journal Record (Oklahoma City, OK) | 6.65% | 95.43% | 8 |
| The Lebanon Daily News (Pennsylvania) | 1.50% | 91.08% | 4 |
| The Minnesota Lawyer (Minneapolis, MN) | 1.56% | 98.81% | 1 |
| The New Hampshire Union Leader, Manchester | 22.37% | 89.47% | 1 |
| The New York Post | 6.49% | 92.98% | 16 |
| The New York Times | 3.86% | 95.91% | 104 |
| The Oklahoman (Oklahoma City, OK) | 15.49% | 89.92% | 1 |
| The Pantagraph (Bloomington, Illinois) | 4.49% | 92.38% | 9 |
| The Philadelphia Daily News | 3.46% | 94.73% | 1 |
| The Philadelphia Inquirer | 3.52% | 95.51% | 28 |
| The Record (Bergen County, NJ) | 3.13% | 92.21% | 12 |
| The Roanoke Times (Virginia) | 7.38% | 94.37% | 7 |
| The Salt Lake Tribune | 6.54% | 86.88% | 11 |
| The San Francisco Chronicle (California) | 5.44% | 95.97% | 13 |
| The State Journal- Register (Springfield, IL) | 5.18% | 91.61% | 8 |
| The Times-Union (Albany, NY) | 15.86% | 90.80% | 2 |
| The Union Leader (Manchester, NH) | 10.46% | 87.84% | 3 |
| The Virginian-Pilot (Norfolk, VA.) | 2.01% | 97.02% | 36 |
| The Washington Post | 4.49% | 92.89% | 56 |
| The Washington Times | 4.81% | 94.72% | 12 |
| The York Dispatch (Pennsylvania) | 6.26% | 87.41% | 5 |
| Topeka Capital-Journal (Kansas) | 5.03% | 92.05% | 5 |
| Tulsa World (Oklahoma) | 8.52% | 87.38% | 22 |
| USA TODAY | 3.23% | 95.51% | 50 |
| Wayne Today (Passaic, North Jersey) | 0.85% | 98.48% | 1 |
| Whittier Daily News (California) | 4.30% | 91.38% | 1 |
| Wisconsin State Journal (Madison, Wisconsin) | 2.60% | 96.48% | 6 |
| Wyoming Tribune-Eagle (Cheyenne) | 14.63% | 91.63% | 1 |

APPENDIX I.
INTERCODER RELIABILITY

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| rater 2 * rater 1 | 33 | 100.0% | 0 | .0% | 33 | 100.0% |

**rater 2 * rater 1 Crosstabulation**

| | | | rater 1 | | | | |
|---|---|---|---|---|---|---|---|
| | | | positive | negative | both | neutral | Total |
| rater 2 | positive | Count | 4 | 0 | 0 | 1 | 5 |
| | | Expected Count | .6 | .2 | .2 | 4.1 | 5.0 |
| | negative | Count | 0 | 0 | 0 | 1 | 1 |
| | | Expected Count | .1 | .0 | .0 | .8 | 1.0 |
| | both | Count | 0 | 0 | 1 | 0 | 1 |
| | | Expected Count | .1 | .0 | .0 | .8 | 1.0 |
| | neutral | Count | 0 | 1 | 0 | 25 | 26 |
| | | Expected Count | 3.2 | .8 | .8 | 21.3 | 26.0 |
| Total | | Count | 4 | 1 | 1 | 27 | 33 |
| | | Expected Count | 4.0 | 1.0 | 1.0 | 27.0 | 33.0 |

**Symmetric Measures**

| | | Value | Asymp. Std. Error[a] | Approx. T[b] | Approx. Sig. |
|---|---|---|---|---|---|
| Measure of Agreement | Kappa | .729 | .146 | 5.445 | .000 |
| N of Valid Cases | | 33 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

*Figure I1*. Intercoder reliability testing for human coders.

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| coded by human * coded by computer | 1220 | 100.0% | 0 | .0% | 1220 | 100.0% |

**coded by human * coded by computer Crosstabulation**

| | | | coded by computer | | | | |
|---|---|---|---|---|---|---|---|
| | | | positive | negative | both | neutral | Total |
| coded by human | positive | Count | 103 | 6 | 30 | 20 | 159 |
| | | Expected Count | 51.2 | 12.5 | 15.8 | 79.5 | 159.0 |
| | negative | Count | 11 | 9 | 10 | 8 | 38 |
| | | Expected Count | 12.2 | 3.0 | 3.8 | 19.0 | 38.0 |
| | both | Count | 2 | 5 | 7 | 3 | 17 |
| | | Expected Count | 5.5 | 1.3 | 1.7 | 8.5 | 17.0 |
| | neutral | Count | 277 | 76 | 74 | 579 | 1006 |
| | | Expected Count | 324.1 | 79.2 | 99.8 | 503.0 | 1006.0 |
| Total | | Count | 393 | 96 | 121 | 610 | 1220 |
| | | Expected Count | 393.0 | 96.0 | 121.0 | 610.0 | 1220.0 |

**Symmetric Measures**

| | | Value | Asymp. Std. Error[a] | Approx. T[b] | Approx. Sig. |
|---|---|---|---|---|---|
| Measure of Agreement | Kappa | .210 | .018 | 12.443 | .000 |
| N of Valid Cases | | 1220 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

*Figure I2*. Intercoder reliability testing for manual vs. automated coding.

Table I1. *Data used for testing RQ3 and RQ4 (results of manual coding).*

| Case | rel_id[a] | art_id[b] | subj_pr[c] | subj_news[d] | pol_pr[e] | pol_news[f] |
|---|---|---|---|---|---|---|
| 1 | 122 | 68 | 0.24000 | 0.15385 | 0.33333 | 1.00000 |
| 2 | 19 | 185 | 0.19048 | 0.33333 | 1.00000 | 1.00000 |
| 3 | 72 | 161 | 0.20000 | 0.06667 | -0.50000 | 1.00000 |
| 4 | 30 | 575 | 0.08333 | 0.12500 | 1.00000 | 1.00000 |
| 5 | 154 | 580 | 0.35000 | 0.12500 | 1.00000 | 1.00000 |
| 6 | 36 | 71 | 0.03333 | 0.23529 | 1.00000 | 0.75000 |
| 7 | 306 | 1452 | 0.22222 | 0.15152 | 0.87500 | -0.40000 |
| 8 | 30 | 1077 | 0.31250 | 0.06897 | 1.00000 | 1.00000 |
| 9 | 359 | 1905 | 0.36364 | 0.16129 | 0.37500 | -1.00000 |
| 10 | 76 | 2990 | 0.15152 | 0.28571 | 1.00000 | 0.08333 |
| 11 | 23 | 265 | 0.10000 | 0.00000 | 1.00000 | 0.00000 |
| 12 | 389 | 76 | 0.29630 | 0.09091 | 1.00000 | 1.00000 |
| 13 | 310 | 435 | 0.07143 | 0.12500 | 1.00000 | 1.00000 |
| 14 | 66 | 65 | 0.16667 | 0.09091 | 1.00000 | 1.00000 |
| 15 | 19 | 181 | 0.19048 | 0.07143 | 0.50000 | -1.00000 |
| 16 | 101 | 245 | 0.11765 | 0.16667 | 1.00000 | 1.00000 |
| 17 | 71 | 80 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 18 | 92 | 123 | 0.00000 | 0.04348 | 0.00000 | 1.00000 |
| 19 | 115 | 42 | 0.10714 | 0.22449 | 1.00000 | 0.36364 |
| 20 | 49 | 141 | 0.04167 | 0.12500 | 1.00000 | 1.00000 |
| 21 | 37 | 109 | 0.05263 | 0.10000 | 1.00000 | 1.00000 |
| 22 | 117 | 260 | 0.29032 | 0.12500 | 1.00000 | 1.00000 |
| 23 | 113 | 197 | 0.58333 | 0.24138 | 0.85714 | -0.14286 |
| 24 | 231 | 98 | 0.40000 | 0.27586 | 0.50000 | -0.75000 |
| 25 | 154 | 660 | 0.32258 | 0.21429 | 1.00000 | 1.00000 |
| 26 | 222 | 2480 | 0.13333 | 0.10000 | 1.00000 | 1.00000 |
| 27 | 251 | 161 | 0.23077 | 0.15789 | 1.00000 | 0.00000 |
| 28 | 180 | 1164 | 0.22727 | 0.18182 | 1.00000 | 1.00000 |

[a]Press release number. [b]Article number. [c]Press release subjectivity score. [d]Article subjectivity score. [e]Press release polarity score. [f]Article polarity score.

# APPENDIX J.
# SUBJECTIVITY AND POLARITY TESTING

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | subjectivity / press releases | .1956639 | 28 | .13646195 | .02578889 |
| | subjectivity / news articles | .1443129 | 28 | .08275441 | .01563911 |

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | subjectivity / press releases & subjectivity / news articles | 28 | .357 | .063 |

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference Lower | Upper | | | |
| Pair 1 | subjectivity / press releases - subjectivity / news articles | .05135107 | .13197691 | .02494129 | .00017577 | .10252637 | 2.059 | 27 | .049 |

*Figure J1*. Paired-samples *t* test for RQ3 (subjectivity scores).

*Note*. SPSS does not offer a way to conduct one-tailed tests, so these results demonstrate a two-tailed test. Since I am conducting a one-tailed test, I use $t_{crit}(27) = 1.7033$. Thus, one-tailed significance is $p = .025$ (.049/2).

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | polarity / press releases | .7835882 | 28 | .39952537 | .07550320 |
| | polarity / news articles | .5322896 | 28 | .68072085 | .12864415 |

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | polarity / press releases & polarity / news articles | 28 | .222 | .257 |

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference Lower | Upper | | | |
| Pair 1 | polarity / press releases - polarity / news articles | .25129857 | .70874719 | .13394063 | -.02352490 | .52612204 | 1.876 | 27 | .071 |

*Figure J2*. Paired-samples *t* test for RQ4 (polarity scores).

*Note*: SPSS does not offer a way to conduct one-tailed tests, so these results demonstrate a two-tailed test. Since I am conducting a one-tailed test, I use $t_{crit}(27) = 1.7033$. Thus, one-tailed significance is $p = .036$ (.071/2).

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | subjectivity: press release | .5422621 | 1642 | .12496385 | .00308388 |
| | subjectivity: news article | .5097029 | 1642 | .15318720 | .00378038 |

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | subjectivity: press release & subjectivity: news article | 1642 | .229 | .000 |

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | | | | |
| | | Mean | Std. Deviation | Std. Error Mean | Lower | Upper | | | |
| Pair 1 | subjectivity: press release - subjectivity: news article | .03255920 | .17409890 | .00429645 | .02413210 | .04098630 | 7.578 | 1641 | .000 |

*Figure J3*. Paired-samples *t* test on automated coding (subjectivity scores).

*Note.* SPSS does not offer a way to conduct one-tailed tests, so these results demonstrate a two-tailed test. Since I am conducting a one-tailed test, I use $t_{crit}(1,641) = 1.645$. One-tailed significance is $p < .001$.

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | sentiment: press release | .6334244 | 1642 | .27259030 | .00672704 |
| | sentiment: news article | .3669479 | 1642 | .37225311 | .00918654 |

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | sentiment: press release & sentiment: news article | 1642 | .254 | .000 |

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | | | | |
| | | Mean | Std. Deviation | Std. Error Mean | Lower | Upper | | | |
| Pair 1 | sentiment: press release - sentiment: news article | .26647652 | .40173574 | .00991411 | .24703087 | .28592217 | 26.879 | 1641 | .000 |

*Figure J4*. Paired-samples *t* test on automated coding (polarity scores).

*Note*. SPSS does not offer a way to conduct one-tailed tests, so these results demonstrate a two-tailed test. Since I am conducting a one-tailed test, I use $t_{crit}(1,641) = 1.645$. One-tailed significance is $p < .001$.

APPENDIX K.
EXPLICIT ATTRIBUTION ANALYSIS

**attribution**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | no | 1471 | 89.6 | 89.6 | 89.6 |
| | yes | 171 | 10.4 | 10.4 | 100.0 |
| | Total | 1642 | 100.0 | 100.0 | |

*Figure K1*. Descriptive statistics for measuring explicit attribution (RQ5).

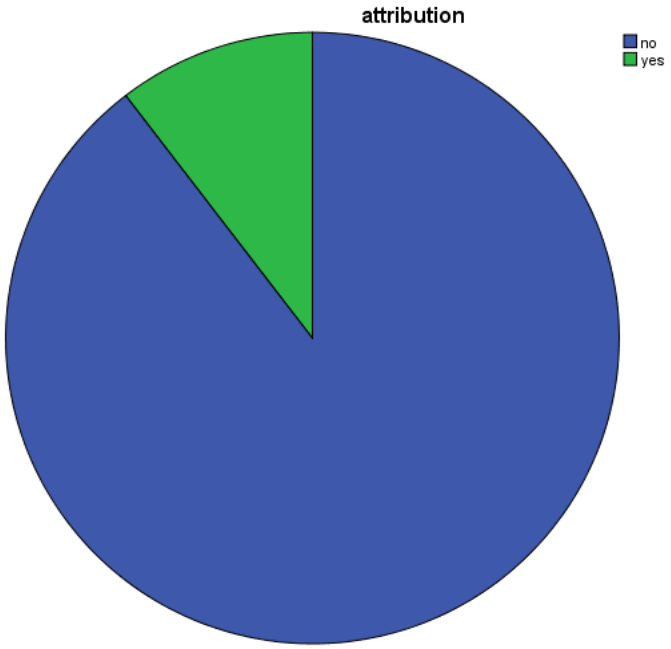



*Figure K2*. Pie chart displaying explicit attribution proportions (RQ5).

Table K1. *Attribution data by publication.*

| Publication | Attribution Percentage | Cases |
|---|---|---|
| AIM West Milford (Passaic, North Jersey) | 0% | 1 |
| Alamogordo Daily News (New Mexico) | 0% | 2 |
| American Banker | 55% | 86 |
| Austin American-Statesman (Texas) | 0% | 4 |
| Automotive News | 0% | 38 |
| Bangor Daily News (Maine) | 0% | 2 |
| Brattleboro Reformer (Vermont) | 0% | 6 |
| Buffalo News (New York) | 0% | 10 |
| Business Insurance | 0% | 7 |
| Chapel Hill Herald (Durham, N.C.) | 0% | 1 |
| Charleston Daily Mail (West Virginia) | 33% | 6 |
| Chicago Daily Herald | 0% | 29 |
| Colorado Springs Business Journal (Colorado Springs, CO) | 25% | 4 |
| Contra Costa Times (California) | 31% | 61 |
| Crain's Detroit Business | 0% | 9 |
| Daily Camera (Boulder, Colorado) | 20% | 5 |
| Daily Journal of Commerce (Portland, OR) | 0% | 1 |
| Daily News (New York) | 0% | 15 |
| Daily Variety | 7% | 14 |
| Dayton Daily News (Ohio) | 11% | 18 |
| Deseret Morning News (Salt Lake City) | 27% | 11 |
| Detroit Free Press (Michigan) | 0% | 47 |
| Education Week | 0% | 1 |
| Finance & Commerce (Minneapolis, MN) | 40% | 5 |
| Florida Times-Union (Jacksonville) | 11% | 9 |
| Government Technology | 100% | 1 |
| Hartford Courant (Connecticut) | 0% | 12 |
| Herald News (Passaic County, NJ) | 0% | 10 |
| Inland Valley Daily Bulletin (Ontario, CA) | 0% | 6 |
| Intelligencer Journal/New Era (Lancaster, Pennsylvania) | 0% | 4 |
| Investment News | 0% | 4 |
| Investor's Business Daily | 4% | 46 |

(continued)

| | | |
|---|---|---|
| Journal of Commerce Online | 0% | 15 |
| Las Cruces Sun-News (New Mexico) | 0% | 4 |
| Las Vegas Review-Journal (Nevada) | 0% | 14 |
| Lewiston Morning Tribune (Idaho) | 0% | 13 |
| Long Island Business (Long Island, NY) | 0% | 7 |
| Los Angeles Times | 0% | 46 |
| Lowell Sun (Massachusetts) | 0% | 11 |
| Marin Independent Journal (California) | 0% | 1 |
| Maryland Gazette | 0% | 2 |
| Mississippi Business Journal (Jackson, MS) | 0% | 1 |
| Missouri Lawyers Media | 100% | 1 |
| Monterey County Herald (California) | 0% | 4 |
| Morning Call (Allentown, Pennsylvania) | 0% | 4 |
| Newsday (New York) | 25% | 4 |
| New York Observer | 50% | 2 |
| Orange County Register (California) | 0% | 9 |
| Oroville Mercury Register (California) | 0% | 1 |
| Palm Beach Post (Florida) | 0% | 6 |
| Pensions & Investments | 50% | 6 |
| Pittsburgh Post-Gazette | 9% | 22 |
| Pittsburgh Tribune Review | 0% | 6 |
| Plastics News | 57% | 7 |
| Providence Journal | 0% | 3 |
| Public Opinion (Chambersburg, Pennsylvania) | 0% | 2 |
| Richmond Times-Dispatch (Virginia) | 0% | 3 |
| Rubber & Plastics News | 0% | 2 |
| San Antonio Express-News | 0% | 7 |
| San Gabriel Valley Tribune (California) | 0% | 6 |
| San Jose Mercury News (California) | 23% | 82 |
| Sarasota Herald Tribune (Florida) | 0% | 5 |
| Sentinel & Enterprise (Fitchburg, Massachusetts) | 0% | 2 |
| South Bend Tribune (Indiana) | 0% | 2 |
| South Florida Sun-Sentinel (Fort Lauderdale) | 0% | 4 |
| Spokesman Review (Spokane, WA) | 15% | 13 |
| Star-News (Wilmington, NC) | 20% | 5 |
| Star Tribune (Minneapolis, MN) | 14% | 29 |

(continued)

| | | |
|---|---|---|
| St. Louis Post-Dispatch (Missouri) | 0% | 5 |
| St. Paul Pioneer Press (Minnesota) | 0% | 1 |
| Suburban Trends (Morris, North Jersey) | 25% | 4 |
| Sunday News (Lancaster, Pennsylvania) | 100% | 1 |
| Tampa Bay Times | 7% | 15 |
| Tampa Tribune (Florida) | 0% | 1 |
| TELEGRAM & GAZETTE (Massachusetts) | 33% | 3 |
| Telegraph Herald (Dubuque, IA) | 0% | 6 |
| The Atlanta Journal-Constitution | 3% | 29 |
| The Augusta Chronicle (Georgia) | 0% | 5 |
| The Baltimore Sun | 0% | 11 |
| The Berkshire Eagle (Pittsfield, Massachusetts) | 0% | 10 |
| The Bismarck Tribune | 0% | 15 |
| The Bond Buyer | 0% | 6 |
| The Capital (Annapolis, MD) | 11% | 9 |
| The Christian Science Monitor | 0% | 1 |
| The Chronicle of Philanthropy | 0% | 1 |
| The Columbian (Vancouver, Washington) | 0% | 1 |
| The Daily News of Los Angeles | 0% | 2 |
| The Daily Oklahoman (Oklahoma City, OK) | 0% | 9 |
| The Daily Record (Baltimore, MD) | 0% | 33 |
| The Daily Record of Rochester (Rochester, NY) | 0% | 2 |
| The Daily Reporter (Milwaukee, WI) | 50% | 2 |
| The Daily Star-Journal, Warrensburg, Mo | 0% | 1 |
| The Dallas Morning News | 29% | 14 |
| The Deal Pipeline | 2% | 52 |
| The Denver Post | 0% | 3 |
| The Detroit News (Michigan) | 1% | 93 |
| The Dispatch (Gilroy, California) | 100% | 1 |
| The Evening Sun (Hanover, Pennsylvania) | 0% | 1 |
| The Gazette (Cedar Rapids, Iowa) | 0% | 3 |
| The Herald Bulletin (Anderson, Indiana) | 100% | 1 |
| The Herald-Sun (Durham, N.C.) | 22% | 9 |
| The Hill | 0% | 2 |
| The Houston Chronicle | 29% | 35 |

(continued)

| | | |
|---|---|---|
| The Indianapolis Business Journal | 0% | 5 |
| The Journal Record (Oklahoma City, OK) | 13% | 8 |
| The Lebanon Daily News (Pennsylvania) | 50% | 4 |
| The Minnesota Lawyer (Minneapolis, MN) | 0% | 1 |
| The New Hampshire Union Leader, Manchester | 0% | 1 |
| The New York Post | 6% | 16 |
| The New York Times | 5% | 104 |
| The Oklahoman (Oklahoma City, OK) | 0% | 1 |
| The Pantagraph (Bloomington, Illinois) | 0% | 9 |
| The Philadelphia Daily News | 0% | 1 |
| The Philadelphia Inquirer | 4% | 28 |
| The Record (Bergen County, NJ) | 0% | 12 |
| The Roanoke Times (Virginia) | 43% | 7 |
| The Salt Lake Tribune | 9% | 11 |
| The San Francisco Chronicle (California) | 8% | 13 |
| The State Journal- Register (Springfield, IL) | 0% | 8 |
| The Times-Union (Albany, NY) | 0% | 2 |
| The Union Leader (Manchester, NH) | 33% | 3 |
| The Virginian-Pilot (Norfolk, VA.) | 0% | 36 |
| The Washington Post | 2% | 56 |
| The Washington Times | 8% | 12 |
| The York Dispatch (Pennsylvania) | 20% | 5 |
| Topeka Capital-Journal (Kansas) | 20% | 5 |
| Tulsa World (Oklahoma) | 5% | 22 |
| USA TODAY | 0% | 50 |
| Wayne Today (Passaic, North Jersey) | 100% | 1 |
| Whittier Daily News (California) | 0% | 1 |
| Wisconsin State Journal (Madison, Wisconsin) | 17% | 6 |
| Wyoming Tribune-Eagle (Cheyenne) | 100% | 1 |

REFERENCES

## Literature


Alcoceba-Hernando, José Antonio. (2010). Analysis of Institutional Press Releases and its Visibility in the Press. *Revista Latina de Comunicación Social, 13*(65), 1-13. doi: 10.4185/rlcs-65-2010-904-354-367-en

Althaus, Scott L., & Tewksbury, David. (2002). Agenda Setting and the "New" News: Patterns of Issue Importance Among Readers of the Paper and Online Versions of the New York Times. *Communication Research, 29*(2), 180-207. doi: 10.1177/0093650202029002004

Amazon Mechanical Turk. (2013). Retrieved at https://www.mturk.com/mturk/.

Ambrosio, Joanne Angela. (1980). It's in the Journal. But this is reporting? *Columbia Journalism Review, 18*(6), 34-36.

American Journalism Review. (2012). *http://www.ajr.org/ajrabout.asp.* 11/27/2012

Anderson, William B. (2001). The media battle between Celebrex and Vioxx: influencing media coverage but not content. *Public Relations Review, 27*(4), 449-460. doi: 10.1016/s0363-8111(01)00100-x

Applegate, Edd. (2005). Mistakes Made in Companies' Press Releases (How to Improve Your Company's Press Releases). *Public Relations Quarterly, 50*(4), 25-30.

Bagdikian, Ben H. (1974). Congress and the media: partners in propaganda. *Columbia Journalism Review, 12*(5), 3-10.

Bajkiewicz, Timothy E., Kraus, Jeffrey J., & Hong, Soo Yeon. (2011). The impact of newsroom changes and the rise of social media on the practice of media relations. *Public Relations Review, 37*(3), 329-331. doi: 10.1016/j.pubrev.2011.05.001

Baxter, Bill L. (1981). The news release: An idea whose time has gone? *Public Relations Review, 7*(1), 27-31. doi: 10.1016/s0363-8111(81)80095-1

Berelson, Bernard. (1971). *Content analysis in communications research*.

Berelson, Bernard, & Lazarsfeld, Paul Felix. (1948). *The analysis of communication content*.

Berkowitz, Dan, & Lee, Jonghyuk. (2004). Media relations in Korea: Cheong between journalist and public relations practitioner. *Public Relations Review, 30*(4), 431-437. doi: 10.1016/j.pubrev.2004.08.011

Bernays, E.L. . (1955). The theory and practice of public relations: A resume. . In E. L. Bernays (Ed.), *The engineering of consent* (pp. 3-25). Norman, OK: University of Oklahoma Press.

Bollinger, Lee. (2001). A New Scoring Method for the Press Release. *Public Relations Quarterly, 46*(1), 31-35.

Brechman, Jean M., Lee, Chul-joo, & Cappella, Joseph N. (2011). Distorting Genetic Research About Cancer: From Bench Science to Press Release to Published News. *Journal of Communication, 61*(3), 496-513. doi: 10.1111/j.1460-2466.2011.01550.x

Bressers, Bonnie, & Gordon, Joye. (2010). Increasing Publicity and Thematic News Coverage: The Impact of Localizing News Releases in a State-Wide Experimental Field Study. *Public Relations Journal*.

Brooks, David S. (1999). The Media Supply Chain: How to Increase Media Coverage for your Product or Service by Understanding and Meeting Shared Responsibilities with the Media. *Public Relations Quarterly, 44*(4), 26.

Bryman, Alan. (1984). The Debate about Quantitative and Qualitative Research: A Question of Method or Epistemology? *The British Journal of Sociology, 35*(1), 75-92.

Budd, Richard W., Thorp, Robert K., & Donohew, Lewis. (1967). *Content analysis of communications*.

Carney, Thomas. (1972). *Content Analysis: A Technique for Systematic Inference from Communications*.

Catenaccio, Paola. (2008). Press Releases as a Hybrid Genre: Addressing the Informative/Promotional Conundrum. *Pragmatics, 18*(1), 9-31.

Chomsky, Noam. (1997). What Makes Mainstream Media Mainstream. from http://www.chomsky.info/articles/199710--.htm

CNN Money. (2013). Fortune 500. Retrieved at http://money.cnn.com/magazines/fortune/fortune500/2012/full_list/.

Cohen, Bernard Cecil. (1963). *The press and foreign policy*.

Cohen, Jacob. (1990). Things I Have Learned (So Far). *American Psychologist, 45*(12), 1304-1312.

Columbia Journalism Review. (2012). *http://www.cjr.org/about_us/mission_statement.php.* Retrieved 11/27/2012

Connolly-Ahern, Colleen, Ahern, Lee A., & Bortree, Denise Sevick. (2009). The effectiveness of stratified constructed week sampling for content analysis of electronic news source archives: AP Newswire, Business Wire, and PR Newswire. *Journalism & Mass Communication Quarterly, 86*(4), 862-883.

Cooley, S.C., & Besova, A. . (2009). *A not so distant past: an examination of distance and salience in media attribute assignment and agenda building.* Paper presented at the NCA 95th Annual Convention, Chicago.

Costco. (2013). Investor relations. Retrieved at http://phx.corporate-ir.net/phoenix.zhtml?c=83830&p=irol-news.

Cutlip, Scott M. (1994). *The Unseen Power: Public Relations: A History*: Routledge

DeLorme, Denise E., & Fedler, Fred. (2003). Journalists' hostility toward public relations: an historical analysis. *Public Relations Review, 29*(2), 99-124. doi: 10.1016/s0363-8111(03)00019-5

DiStaso, Marcia W. (2012). The Annual Earnings Press Release's Dual Role: An Examination of Relationships with Local and National Media Coverage and Reputation. *Journal of Public Relations Research, 24*(2), 123-143. doi: 10.1080/1062726x.2012.626131

Donsbach, Wolfgang, Jandura, Olaf, & Jandura, Grit. (2005). *Against Long Odds? Self-Portrayals of Political Parties in Their Press Releases and the Media*. Paper presented at the International Communication Association. Conference Paper retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=18654940&site=ehost-live

Entman, Robert M. (1991). Framing U.S. Coverage of International News: Contrasts in Narratives of the KAL and Iran Air Incidents. *The Journal of Communication, 41*(4), 6-27.

Entman, Robert M. . (1993). Framing: Toward Clarification of a Fractured Paradigm. *The Journal of Communication, 43*(4), 51-58.

Gandy, O.H. (1982). *Beypnd agenda setting: information subsidies and public policy*.

Gentzkow, M., & Shapiro, J.M. (2007). What Drives Media Slant? Evidence from U.S. Daily Newspapers. *National Bureau of Economic Research*.

George, Alexander L. (1973). *Propaganda analysis: a study of inferences made from Nazi propaganda in World War II*.

Gilpin, Dawn. (2007). *Attractor Basins in the Phase Space of Reputation: The Example of Wal-Mart and the Media*. Paper presented at the International Communication Association. Conference Paper retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=26949772&site=ehost-live

Goffman, Erving. (1974). *Frame analysis*.

Golitsinski, Sergei. (2007). *Significance of the General Public for Public Relations: A Study of the Blogosphere's Impact on the October 2006 Edelman/Wal-Mart Crisis.* (Master of Arts), University of Norrthern Iowa, Cedar Falls.

Google. (2013). News from Google. Retrieved at https://www.google.com/intl/en/press.

Google. (2013). 2012 Press releases. Retrieved at https://investor.google.com/releases/2012.

Gower, Karla. (2007). *Public relations and the press: The troubled embrace*: Northwestern University Press.

Goya-Martinez, M. (2009). *The visual context and representation of tragic events in news websites.* Paper presented at the NCA 95th Annual Convention.

Grimmer, Justin, & Stewart, Brandon M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*. doi: 10.1093/pan/mps028

Grunig, James E., & Grunig, Larissa A. (1992). Models of public relations and communication. In J. E. Grunig (Ed.), *Excellence in public relations and communication management* (pp. 285-322). Hillsdale, NJ: Lawrence Erlbaum.

Grunig, James E., & Hunt, T. (1984). *Managing public relations*. New York: Holt, Rinehart and Winston.

Hale, Dennis. (1978). Press Releases vs. Newspaper Coverage of California Supreme Court Decisions. *Journalism Quarterly, 55*(4), 696-710.

Hanson, Ralph E. (2011). *Mass communication: Living in a media world* (3 ed.). Washington, DC: CQ Press.

Heath, Robert L. (2010). *The SAGE Handbook of Public Relations*.

Henry, Elaine. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication, 45*(4), 363-407.

Hiebert, Ray E. (1966). *Courtier to the Crowd: The Story of Ivy Lee and the Development of Public Relations*. Ames, IA: Iowa State University Press.

Holliday, Adrian. (2007). *Doing and writing qiualitative research*.

Holody, Kyle. (2009). *Framing Public Discourse on Physician-Assisted Suicide: Analysis of Newspaper Coverage and Death With Dignity Press Releases*. Paper presented at the International Communication Association. Article retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=45286732&site=ehost-live

Holsti, Ole R. (1969). *Content analysis for the social sciences and humanities*.

Hong, Soo Yeon. (2008). The relationship between newsworthiness and publication of news releases in the media. *Public Relations Review, 34*(3), 297-299. doi: 10.1016/j.pubrev.2008.03.033

Hou, J., & Ma, Y. (2009). *China's national power and U.S. news coverage.* Paper presented at the NCA 95th Annual Convention.

Hyejoon, Rim, Byung-Gu, Lee, & Ji Won, Han. (2009). *The Influence of News Sources on Health News Content: Does Localization Really Matter?* Paper presented at the International Communication Association. Article retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=45286751&site=ehost-live

Johnson Avery, Elizabeth, & Kim, Sora. (2008). Comprising or Compromising Credibility? Use of Spokesperson Quotations in News Releases Issued by Major Health Agencies. *Public Relations Journal*.

Jones, Harold Y. (1975). Filling up the white space. *Columbia Journalism Review, 14*(1), 10-11.

Jurafsky, Daniel, & Martin, James H. (2008). *Speech and language processing*.

Kiousis, Spiro, Mitrook, Michael, Xu, Wu, & Seltzer, Trent. (2006). First- and Second-Level Agenda-Building and Agenda-Setting Effects: Exploring the Linkages Among Candidate News Releases, Media Coverage, and Public Opinion During the 2002 Florida Gubernatorial Election. *Journal of Public Relations Research, 18*(3), 265-285. doi: 10.1207/s1532754xjprr1803_4

Kiousis, Spiro, Popescu, Cristina, & Mitrook, Michael. (2007). Understanding Influence on Corporate Reputation: An Examination of Public Relations Efforts, Media Coverage, Public Opinion, and Financial Performance From an Agenda-Building and Agenda-Setting Perspective. *Journal of Public Relations Research, 19*(2), 147-165. doi: 10.1080/10627260701290661

Kiousis, Spiro, Soo-Yeon, Kim, McDewitt, Michael, & Ostrowksi, Ally. (2009). Competing for attention: information subsidy influence in agenda building during election campaigns. *Journalism & Mass Communication Quarterly, 86*(3), 545-562.

Kogan, Shimon, Levin, Dimitry, Routledge, Bryan R., Sagi, Jacob S., & Smith, Noah A. (2009). *Predicting risk from financial reports with regression*. Paper presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on ZZZ, Boulder, Colorado.

Kopenhaver, Lillian Lodge. (1985). Aligning values of practitioners and journalists. *Public Relations Review, 11*(2), 34-42. doi: 10.1016/s0363-8111(82)80117-3

Krippendorff, Klaus. (2012). *Content Analysis: An Introduction to its Methodology* (3d ed.).

Kruckeberg, Dean, & Starck, Kenneth. (1988). *Public relations and community : a reconstructed theory*. New York: Praeger.

Lamme, M.O., & Russell, K.M. (2010). Removing the Spin: Toward a New Theory of Public Relations History. *Journalism & Communication Monographs 11*(Winter (4)), 281-362.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lang, Kurt, & Lang, Gladys Engel. (1952). The Unique Perspective of Television and its Effects: A Pilot Study. *American Sociological Review, 18*.

Ledingham, John A., & Bruning, S.D. (2000). Introduction: background and current trends in the study of relationship management. In J. A. Ledingham & S. D. Bruning (Eds.), *Public relations as a relationship management: A relational approach to the study and practice of public relations* (pp. xi-xvii). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Lehman-Wilzig, Sam, & Seletzky, Michal. (2012). Elite and Popular Newspaper Publication of Press Releases: Differential Success Factors? *Public Relations Journal*.

Len-Ríos, María E., Hinnant, Amanda, Sun, A. Park, Cameron, Glen T., Frisby, Cynthia M., & Youngah, Lee. (2009). Health news agenda building: journalists' perceptions of the role of public relations. *Journalism & Mass Communication Quarterly, 86*(2), 315-331.

Leskovec, Jure, Backstrom, Lars, & Kleinberg, Jon. (2009). *Meme-tracking and the dynamics of the news cycle*. Paper presented at the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France.

Levin, David. (2002). Making a Good Impression: Peace Movement Press Release Styles and Newspaper Coverage. *Harvard International Journal of Press/Politics, 7*(1), 79.

Lewis, Justin, Williams, Andrew, & Franklin, Bob. (2008a). A compromised fourth estate? *Journalism Studies, 9*(1), 1-20. doi: 10.1080/14616700701767974

Lewis, Justin, Williams, Andrew, & Franklin, Bob. (2008b). Four rumours and an explanation. *Journalism Practice, 2*(1), 27-45. doi: 10.1080/17512780701768493

LexisNexis. (2013). SmartIndexing. Retrieved at http://wiki.lexisnexis.com.proxy-um.researchport.umd.edu/academic/index.php?title=SmartIndexing#Searching:__Power_Search

Lindlof, Thomas R., & Taylor, Bryan C. (2002). *Qualitative communication research methods*.

Lippmann, Walter. (1922). *Public opinion*.

Liu, Bing. (2010). Sentiment Analysis and Subjectivity Handbook of Natural Language Processing (2 ed.).

Lowery, S., & De Fleur, M. L. (1983). *Milestones in mass communication research.*

Maat, Henk Pander. (2007). How promotional language in press releases is dealt with by journalists. *Journal of Business Communication, 44*(1), 59-95.

Maat, Henk Pander. (2008). Editing and Genre Conflict: How Newspaper Journalists Clarify and Neutralize Press Release Copy. *Pragmatics, 18*(1), 87-113.

Maat, Henk Pander, & de Jong, Caro. (2012). How newspaper journalists reframe product press release information. *Journalism*.

Manning, C.D., Raghavan, Prabhakar , & Schütze, Hinrich (2008). *Introduction to Information Retrieval*

Manning, C.D., & Schutze, H. (1999). *Foundations of statistical natural language processing*.

Marken, G. A. (1994). Let's Do Away with Press Releases. *Public Relations Quarterly, 39*(1), 46-48.

Martin, William P., & Singletary, Michael W. (1981). Newspaper Treatment of State Government Releases. *Journalism Quarterly, 58*(1), 93-96.

McCombs, Maxwell. (1983). Beyond Agenda Setting: Information Subsidies and Public Policy (Book). *Journalism Quarterly, 60*(2), 376-377.

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

McCombs, Maxwell, & Estrada, G. (1997). The news media and the pictures in our heads. In S. Iyengar & R. Reeves (Eds.), *Do the media govern?* (pp. 237-247). Thousand Oaks, CA: Sage Publications Ltd.

McCombs, Maxwell, & Reynolds, A. (2002). News influence on our pictures of the world. In J. Bryant & D. Zillmann (Eds.), *Media effects: Advances in theory and research.* (pp. 1-18). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McCombs, Maxwell, & Reynolds, Amy. (2009). How the news shapes our civic agenda. In J. Bryant & M. B. Oliver (Eds.), *Media Effects: Advances in Theory and Research* (pp. 1-16). New York: Routledge.

McCombs, Maxwell, & Shaw, Donald L. (1993). The evolution of agenda-setting research: Twenty-five years in the marketplace of ideas. *Journal of Communication, 43*(2), 58-67.

McQuail, Denis. (2005). *McQuail's mass communication theory* (5th ed.). London ; Thousand Oaks, Calif.: Sage Publications.

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

Miller, David, & Dinan, William. (2007). Public relations and the subversion of democracy. In D. Miller & W. Dinan (Eds.), *Thinker, faker, spinner, spy* (pp. 11-21). London: Pluto Press.

Monroe, B. L., Colaresi, M.P., & Quinn, K.M. (2008). Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis, 16*(4). doi: 10.1093/pan/mpn018

Moody, Reginald F. (2009). Writing the Narrative Press Release: Is it the Magic Potion for More Usable Press Communications? *Public Relations Journal*.

Morton, Linda P., & Warren, John. (1992). Proximity: localization vs. Distance in PR news releases. *Journalism Quarterly, 69*(4), 1023-1028.

Murphy, Priscilla. (2001). Affiliation Bias and Expert Disagreement in Framing the Nicotine Addiction Debate. *Science, Technology, & Human Values, 26*(3), 278-299.

Murphy, Priscilla. (2010). The Intractability of Reputation: Media Coverage as a Complex System in the Case of Martha Stewart. *Journal of Public Relations Research, 22*(2), 209-237. doi: 10.1080/10627261003601648

Natural Language Toolkit. (2013). Retrieved at http://nltk.org/.

Neuendorf, K. A. (2002). *The content analysis guidebook*.

Newsom, Douglas Ann, Turk, Judy VanSlyke, & Dean, Kruckeberg. (2004). *This is PR: The realities of public relations* (8 ed.). Belmont, CA: Wadsworth.

NIST. (2013). string matching. Retrieved at http://xlinux.nist.gov/dads/HTML/stringMatching.html.

Oard, Douglas W. (2009). A Whirlwind Tour of Automated Language Processing for the Humanities and Social Sciences *Working Together or Apart: Promoting the Next Generation of Digital Scholarship: Report of a Workshop Cosponsored by the Council on Library and Information Resources and The National Endowment for the Humanities* (pp. 34-43 ): Council on Library and Information Resources.

Obston, Andrea. (2004). The Eight Words You Can't Say in a Press Release. *Public Relations Quarterly, 49*(3), 9-29.

Ohl, Coral M., Pincus, J. David, Rimmer, Tony, & Harrison, Denise. (1995). Agenda building role of news releases in corporate takeovers. *Public Relations Review, 21*(2), 89-101. doi: 10.1016/0363-8111(95)90001-2

Palser, Barb. (2006). Artful Disguises. *American Journalism Review, 28*(5), 90-90.

Paluszek, John L. (2002). Propaganda, Public Relations, and Journalism: when bad things happen to good words. *Journalism Studies, 3*(3), 441-446.

Pang, Bo, & Lee, Lillian. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135. doi: http://dx.doi.org/10.1561/1500000011

Pang, Bo, Lee, Lillian, & Vaithyanathan, Shivakumar (2002). *Thumbs up? Sentiment classification using machine learning techniques.* Paper presented at the EMNLP.

Pedersen, Wesley. (1976). The Government News Release: Tedium Need Not Be The Message. *Public Relations Quarterly, 21*(4), 17.

Pelham, Fran. (2000). The Triple Crown of Public Relations: Pitch Letter, News Release, Feature Article. *Public Relations Quarterly, 45*(1), 38-43.

Pinkleton, Bruce. (1994). The Campaign of the Committee on Public Information: Its Contributions to the History and Evolution of Public Relations. *Journal of Public Relations Research, 6*(4), 229-240.

Popping, R. (2000). *Computer-assisted text analysis*. London ; Thousand Oaks, Calif.: Sage Publications.

Potter, Deborah. (2004). Virtual News Reports. *American Journalism Review, 26*(3), 68-68.

A press release and a news story. (1975). *Columbia Journalism Review, 14*(4), 4-6.

Python Standard Library. (2013). Retrieved at http://docs.python.org/2/library/difflib.html.

Ragas, Matthew W. (2012). Issue and Stakeholder Intercandidate Agenda Setting among Corporate Information Subsidies. *Journalism & Mass Communication Quarterly, 89*(1), 91-111. doi: 10.1177/1077699011430063

Reese, Stephen D. (2007). The Framing Project: A Bridging Model for Media Research Revisited. *Journal of Communication, 57*, 148-154.

Rodgers, Ronald R. (2010). The press and public relations through the lens of the periodicals, 1890–1930. *Public Relations Review, 36*(1), 50-55. doi: 10.1016/j.pubrev.2009.10.012

Rogers, E.M., Hart, W.B., & Dearing, J.W. (1997). A paradigmatic history of agenda-setting research. In S. Iyengar & R. Reeves (Eds.), *Do the media govern?* (pp. 225-236). Thousand Oaks, CA: Sage Publications Ltd.

Russell, Karen Miller, & Bishop, Carl O. (2009). Understanding Ivy Lee's declaration of principles: U.S. newspaper and magazine coverage of publicity and press agentry, 1865–1904. *Public Relations Review, 35*(2), 91-101. doi: 10.1016/j.pubrev.2009.01.004

Ryan, Michael. (1995). Models Help Writers Produce Publishable Releases. *Public Relations Quarterly, 40*(2), 25-27.

Ryan, Michael, & Martinson, David L. (1988). Journalists and Public Relations Practitioners: Why the Antagonism? *Journalism Quarterly, 65*(1), 131-140.

Sallot, Lynne M., & Johnson, Elizabeth A. (2006). Investigating relationships between journalists and public relations practitioners: Working together to set, frame and build the public agenda, 1991–2004. *Public Relations Review, 32*(2), 151-159. doi: 10.1016/j.pubrev.2006.02.008

Scheufele, Dietram A. (1999). *Participation as individual choice : comparing motivational and informational variables and their relevance for participatory behavior.* (Thesis (Ph D )), University of Wisconsin--Madison, 1999. Retrieved from http://www.library.wisc.edu/databases/connect/dissertations.html

Scheufele, Dietram A., & Tewksbury, David. (2007). Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models. *Journal of Communication, 57*, 9-20.

Schudson, Michael. (2001). The objectivity norm in American journalism*. *Journalism, 2*(2), 149-170.

Schultz, Friederike, Kleinnijenhuis, Jan, Oegema, Dirk, Utz, Sonja, & van Atteveldt, Wouter. (2012). Strategic framing in the BP crisis: A semantic network analysis of associative frames. *Public Relations Review, 38*(1), 97-107. doi: 10.1016/j.pubrev.2011.08.003

Seletzky, Michal, & Lehman-Wilzig, Sam. (2010). Factors Underlying Organizations' Successful Press Release Publication in Newspapers: Additional PR Elements for the Evolving 'Press Agentry' and 'Public Information' Models. *International Journal of Strategic Communication, 4*(4), 244-266. doi: 10.1080/1553118x.2010.515538

Shannon, Claude Elwood. (1948). The mathematical theory of communication. *Bell System technical journal, July*.

Shannon, Claude Elwood, & Weaver, Warren. (1963). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Shehata, Adam, & Strömbäck, Jesper. (2013). Not (Yet) a New Era of Minimal Effects: A Study of Agenda Setting at the Aggregate and Individual Levels. *International Journal of Press/Politics, 18*(2), 234-255. doi: 10.1177/1940161212473831

Sibbison, Jim. (1988). Dead fish and red herrings: how the EPA pollutes the news. (cover story). *Columbia Journalism Review, 27*(4), 25-28.

Sleurs, Kim, Jacobs, Geert, & Van Waes, Luuk. (2003). Constructing press releases, constructing quotations: A case study. *Journal of Sociolinguistics, 7*(2), 192-212. doi: 10.1111/1467-9481.00219

Smith, R. Jeffrey. (1977). The media's sweet tooth. *Columbia Journalism Review, 16*(1), 28-29.

Snover, Matthew, Madnani, Nitin, Dorr, Bonnie, & Schwartz, Richard. (2009). Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. Paper presented at the The Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009), Athens, Greece.

Spicer, Christopher H. (1993). Images of Public Relations in the Print Media. *Journal of Public Relations Research, 5*(1), 47-61.

Spinn3r. (2013). We index the blogosphere so you don't have to! Retrieved at http://www.spinn3r.com/features.

Stauber, J.C., & Rampton, S. (1995). *Toxic sludge is good for you: Lies, damn lies and the public relations industry*. Monroe, Maine: Common Courage Press.

Steel, R. (1980). *Walter Lippmann and the American century*. Boston, MA: Little, Brown & Co.

Stempel, G.H., Weaver, D.H., & Wilhoit, G.C. (2003). *Mass communication research and theory*: Pearson Education, Inc.

Strobbe, Ilse, & Jacobs, Geert. (2005). E-releases: A view from linguistic pragmatics. *Public Relations Review, 31*(2), 289-291. doi: 10.1016/j.pubrev.2005.02.009

Student. (1908). The probable error of a mean. *Biometrica, 6*(1).

Sullivan, John. (2011). True Enough. *Columbia Journalism Review, 50*(1), 34-39.

Supa, Dustin W., & Zoch, Lynn M. (2009). Maximizing Media Relations Through a Better Understanding of the Public Relations-Journalist Relationship: A Quantitative Analysis of Changes Over the Past 23 Years. *Public Relations Journal*.

Superceanu, Rodica. (2011). Intertextuality and informativity of press releases: Factors determining the communication between PR practitioner and journalist. *PCTS Proceedings (Professional Communication & Translation Studies), 4*(1/2), 21-30.

Sweetser, Kaye D., & Brown, Charles W. (2008). Information subsidies and agenda-building during the Israel–Lebanon crisis. *Public Relations Review, 34*(4), 359-366. doi: 10.1016/j.pubrev.2008.06.008

Takeshita, Toshio. (2006). Current Critical Problems in Agenda-Setting Research. *International Journal of Public Opinion Research, 18*(3), 275-296. doi: 10.1093/ijpor/edh104

Tausczik, Yla R., & Pennebaker, James W. (2009). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*.

Taylor, Maureen. (2009). Protocol journalism as a framework for understanding public relations–media relationships in Kosovo. *Public Relations Review, 35*(1), 23-30. doi: 10.1016/j.pubrev.2008.12.002

Tremayne, Mark. (2004). The web of context: applying network theory to the use of hyperlinks in journalism on the web. *Journalism & Mass Communication Quarterly, 81*(2), 237-253.

Tremayne, Mark, Nan, Zheng, Jae Kook, Lee, & Jaekwan, Jeong. (2006). Issue Publics on the Web: Applying Network Theory to the War Blogosphere. *Journal of Computer-Mediated Communication, 12*(1), 290-310. doi: 10.1111/j.1083-6101.2006.00326.x

Tremayne, Mark, Weiss, Amy Schmitz, & Alves, Rosental Calmon. (2007). From product to service: the diffusion of dynamic content in online newspapers. *Journalism & Mass Communication Quarterly, 84*(4), 825-839.

Tuchman, Gaye. (1972). Objectivity as Strategic Ritual: An Examination of Newsmen's Notions of Objectivity. *American Journal of Sociology, 77*(4), 660-679. doi: 10.2307/2776752

Turk, Judy VanSlyke, & Franklin, Bob. (1987). Information subsidies: Agenda-setting traditions. *Public Relations Review, 13*(4), 29-41. doi: 10.1016/s0363-8111(87)80015-2

US Pages. (2013). Fortune 500. Retrieved at http://www.uspages.com/fortune500.htm.

USA Today. (2008). On 225th birthday, newspapers dying? Retrieved at http://usatoday30.usatoday.com/printedition/news/20080523/al23.art.htm.

van Hoof, Anita M. J., Hermans, Liesbeth, & van Gorp, Baldwin. (2008). *The Influence of Press Releases on the Use of Strategic and Issue Frames*. Paper presented at the International Communication Association. Article retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=36956776&site=ehost-live

Van Hout, Tom, Pander Maat, Henk, & De Preter, Wim. (2011). Writing from news sources: The case of Apple TV. *Journal of Pragmatics, 43*(7), 1876-1889. doi: 10.1016/j.pragma.2010.09.024

Verhoeven, Piet, Tench, Ralph, Zerfass, Ansgar, Moreno, Angeles, & Verčič, Dejan. (2012). How European PR practitioners handle digital and social media. *Public Relations Review, 38*(1), 162-164. doi: 10.1016/j.pubrev.2011.08.015

Vos, Tim P. (2011). Explaining the Origins of Public Relations: Logics of Historical Explanation. *Journal of Public Relations Research, 23*(2), 119-140. doi: 10.1080/1062726x.2010.504793

Warren, John, & Morton, Linda P. (1991). Readability and Acceptance of Public Relations Releases from Institutions of Higher Education. *Communication Research Reports, 8*(1/2), 113-119.

Waters, Richard D., Tindall, Natalie T. J., & Morton, Timothy S. (2010). Media Catching and the Journalist-Public Relations Practitioner Relationship: How Social Media are Changing the Practice of Media Relations. *Journal of Public Relations Research, 22*(3), 241-264. doi: 10.1080/10627261003799202

Wikipedia. (2013). Fortune 500. Retrieved at http://en.wikipedia.org/wiki/Fortune_500.

Wikipedia. (2013). List of newspapers in the United States. Retrieved at https://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_States

Wikipedia. (2013). String searching algorithm. Retrieved at http://en.wikipedia.org/wiki/String_searching_algorithm.

Wikipedia. (2013). Algorithm. Retrieved at http://en.wikipedia.org/wiki/Algorithm.

Wikipedia. (2013). tf-idf. Retrieved at http://en.wikipedia.org/wiki/Tf%E2%80%93idf.
Wikipedia. (2013). Text corpus. Retrieved at http://en.wikipedia.org/wiki/Text_corpus.

Williams, Doug. (1994). In Defense of the (Properly Executed) Press Release. *Public Relations Quarterly, 39*(3), 5-7.

Wilson, Theresa, Wiebe, Janyce, & Hoffmann, Paul. (2009). Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics, 35*(3), 399-433.

Wimmer, Roger D., & Dominick, Joseph R. (1997). *Mass media research: an introduction*.

Wing, J. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33-35.

Wright, Donald K., & Hinson, Michelle D. (2009). An Updated Look at the Impact of Social Media on Public Relations Practice. *Public Relations Journal*.

Zhang, Juyan. (2004). *International agenda building and media response: How U.S. major newspapers used Saudi Arabia's press releases in its public relations campaign*. Paper presented at the International Communication Association. Article retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=45283719&site=ehost-live

Zhou, Liang, Lin, Chin-Yew, & Hovy, Eduard. (2006). *Re-evaluating machine translation results with paraphrase support*. Paper presented at the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia.

Zoch, Lynn M., & Molleda, Juan-Carlos. (2009). Building a theoretical model of media relations using framing, information subsidies, and agenda-building. In C. H. Botan & V. Hazleton (Eds.), *Public relations theory II* (pp. 279-309). New York, NY: Routledge.

**Cited Data: Press Releases**

Abbott. (2012, September 21). The Journal Science Again Recognizes Abbott as One of the Top Employers in the Biotech and Pharmaceutical Industry. Retrieved at http://www.abbott.com/news-media/press-releases/the-journal-science-again-recognizes-abbott-as-one-of-the-top-employers-in-the-biotech-and-pharmace.htm

Aetna. (2012, June 28). Aetna Statement on Supreme Court Ruling on the Affordable Care Act. Retrieved at http://newshub.aetna.com/press-release/corporate-and-financial/aetna-statement-supreme-court-ruling-affordable-care-act

Allstate. (2012, January 6). Superior Court of New Jersey Rules In Favor of Allstate New Jersey in Agency Franchise Lawsuits. Retrieved at http://www.allstatenewsroom.com/channels/News-Releases/releases/superior-court-of-new-jersey-rules-in-favor-of-allstate-new-jersey-in-agency-franchise-lawsuits?mode=print

Allstate. (2012, January 23). Protect Holiday Gifts with a Home Inventory and Insurance Smarts. Retrieved at http://www.allstatenewsroom.com/channels/NewsReleases/releases/protect-holiday-gifts-with-a-home-inventory-and-insurancesmarts?mode=print

Allstate. (2012, February 21). Allstate Increases Quarterly Dividend 4.8 Percent. Retrieved at http://www.allstatenewsroom.com/channels/News-Releases/releases/allstate-increases-quarterly-dividend-4-8-percent?mode=print

Allstate. (2012, February 27) Allstate Announces Senior Leadership Changes. Retrieved at http://www.allstatenewsroom.com/channels/News-Releases/releases/allstate-announces-senior-leadership-changes?mode=print

Amazon.com (2012, August 17). The Hunger Games Trilogy is now the Best-Selling Book Series of All Time on Amazon.com. Retrieved at http://phx.corporate-ir.net/phoenix.zhtml?c=176060&p=irol-newsArticle_Print&ID=1726645

Boeing. (2012, January 4). Boeing to Close Wichita Facility by the End of 2013. Retrieved at http://boeing.mediaroom.com/index.php?s=43&item=2090

Comcast. (2012, January 4). The Walt Disney Company and Comcast Corporation Announce a Long-Term, Comprehensive Distribution Agreement That Advances the Successful Multichannel Business Model. Retrieved at http://corporate.comcast.com/news-information/news-feed/the-walt-disney-company-and-comcast-corporation-announce-a-long-term-comprehensive-distribution-agreement-that-advances-the-successful-multichannel-business-model?print=1

CVS. (2012, May 23). MinuteClinic Signs Clinical Collaboration with Atlantic Health System for Clinic Locations in Five-County Region of Northern and Central New Jersey. http://info.cvscaremark.com/newsroom/press-releases/minuteclinic-signs-clinical-collaboration-atlantic-health-system-clinic-loca

Deere. (2012, January 26). John Deere Marks the Company's 175th Anniversary. Retrieved at http://www.deere.com/wps/dcom/en_US/corporate/our_company/news_and_media/press_releases/2012/corporate/2012jan26_corporaterelease.page

ExxonMobil. (2012, April 16). Rosneft and ExxonMobil today signed agreements. Retrieved at http://news.exxonmobil.com/press-release/rosneft-and-exxonmobil-announce-progress-strategic-cooperation-agreement

ExxonMobil. (2012, April 23). ExxonMobil and Employees Donate $8.8 Million to Texas Colleges and Universities. Retrieved at http://news.exxonmobil.com/press-release/exxonmobil-and-employees-donate-88-million-texas-colleges-and-universities

Fannie Mae. (2012, January 9). Consumer Attitudes Improve in December. Retrieved at http://www.fanniemae.com/portal/about-us/media/corporate-news/2012/5603.html

Fannie Mae. (2012, November 9). Speeches - Remarks by Tim Mayopoulos, National... Retrieved at http://www.fanniemae.com/portal/about-us/media/speeches/2012/speech-mayopoulos-2012national-association-of-realtors-conference.html

FedEx. (2012, August 14). FedEx Express Expands 'International First' Service for Early Deliveries. Retrieved at http://news.van.fedex.com/fedex-express-expands-%E2%80%98international-first%E2%80%99-service-early-deliveries

Home Depot. (2012, September 13). The Home Depot Closes Seven Big Box Stores In China. Retrieved from http://phx.corporate-ir.net/phoenix.zhtml?c=63646&p=irol-newsArticle_Print&ID=1735130&highlight=

Kroger. (2012, March 22). Kroger Will No Longer Purchase Ground Beef Made with Lean Finely Textured Beef. Retrieved from

Liberty Mutual. (2012, Feburary 22). Hazy Logic: Liberty Mutual Insurance/SADD Study Finds driving Under the Influence of Marijuana a Greater Threat to Teen Drivers than Alcohol. Retrieved from http://www.libertymutualgroup.com/omapps/ContentServer?kw=true&c=cms_asset&pagename=LMGroup%2FViews%2FlmgView98&cid=1240007590982

MetLife. (2012, January 10). MetLife Exits Forward Mortgage Business. Retrieved from https://www.metlife.com/about/press-room/us-press-releases/2012/index.html?compID=73539

Microsoft. (2012, January 12). Microsoft and LG Sign Patent Agreement Covering Android and Chrome OS Based Devices. Retrieved from http://www.microsoft.com/en-us/news/press/NewsArchive.aspx?feedid=PressReleases

PepsiCo. (2012, March 29). Doritos Debuts Distinctly Rolled Doritos Dinamita Flavored Tortilla Chips. Retrieved at http://www.pepsico.com/PressRelease/Doritos-Debuts-Distinctly-Rolled-Doritos-Dinamita-Flavored-Tortilla-Chips03292012.html

Wells Fargo. (2012, September 26). Wells Fargo brings CityLIFT program to Washington, D.C. and Prince George's County to help local housing market. Retrieved at https://www.wellsfargo.com/press/2012/20120926_DCCityLIFT

Wells Fargo. (2012, November 27). Wells Fargo Brings CityLIFT Program to Alameda and Contra Costa Counties to Help Local Housing Market. Retrieved at https://www.wellsfargo.com/press/2012/20121127_WellsFargobringsCityLifttoAlameda

**Cited Data: News Articles**

Automotive News. (2012, May 14). Racer, entrepreneur Shelby dies at 89. Retrieved from LexisNexis Academic on February 10, 2012.

Chicago Daily Herald. (2012, February 28). Shakeup at Allstate claims chief marketing officer. Retrieved from LexisNexis Academic on February 10, 2012.

Chicago Daily Herald. (2012, October 01). Abbott recognized as top employer in industry. Retrieved from LexisNexis Academic on February 10, 2012.

Contra Costa Times. (2012, November 27). Wells Fargo to launch $5 million effort to provide down payments for East Bay potential home buyers. Retrieved from LexisNexis Academic on February 10, 2012.

Deseret Morning News (Salt Lake City). (2012, January 16). Survey finds consumers more confident.

The Deal Pipeline. (2012, April 16). Exxon Mobil, Rosneft finalize partnership. Retrieved from LexisNexis Academic on February 10, 2012.

Long Island Business (Long Island, NY). (2012, February 6). MetLife laying off 55 in Hauppauge. Retrieved from LexisNexis Academic on February 10, 2012.

San Jose Mercury News. (2012, November 27). Wells Fargo to launch $5 million effort to provide down payments for East Bay potential home buyers. Retrieved from LexisNexis Academic on February 10, 2012.

Star Tribune. (2012, July 24). Target assails swipe-fee settlement. Retrieved from LexisNexis Academic on February 10, 2012.

Spokesman Review (Spokane, WA). (2012, January 5). Boeing to Leave Wichita. Retrieved from LexisNexis Academic on February 10, 2012.

The New York Times. (2012, October 25). European Regulator Says Microsoft Violated a Deal. Retrieved from LexisNexis Academic on February 10, 2012.