Textual Analysis of Corporate Disclosures: A Survey of the Literature

Feng Li

University of Michigan

feng@umich.edu

February 5, 2011

Forthcoming in the Journal of Accounting Literature

I thank Jason Chen, Yu Huang, Kenneth Merkley, Nemit Shroff, and Jenny Tucker (the Associate Editor) for their comments and Carrie Wood for her editorial help.

1.0 Introduction

Accounting is fundamentally about information. For instance, accountants aggregate hundreds of economic transactions into net income, a summary measure of firm performance. In addition to the numeric information, there is a large amount of unstructured textual information in corporate disclosures.[1] Take Dow Chemical's 2009 Annual Report as an example. The 207-page report contains 5 pages of financial statements and about 25 pages of other tables. The rest of the report consists of text including the CEO and Chairman's Letter to the Stockholders, Management Discussion and Analysis, and the Notes to the Financial Statements, among other things.

Understanding the textual information in corporate disclosures is important for financial accounting research. First, the textual disclosures contain information about the data generating function of the numeric financial data that accounting researchers have been examining for decades. Obviously, two reported sales revenue numbers should have different properties if the revenue recognition policies are different. However, the data generating function differences can also be due to different management incentives. For instance, when the management discussions in 10-K filings are complex and difficult to understand, managers might be obfuscating information and the reported financial data are therefore likely to have lower quality [Li, 2008]. Thus the textual information can provide a very useful context for understanding the financial data and testing interesting economic hypotheses.

Second, managers' communication patterns could reveal certain managerial characteristics and thus have significant implications for understanding corporate decisions. Recent developments in behavioral economics emphasize the cognitive biases

of human beings and the roles of these biases in decision making [Kahneman, 2003]. However, it is difficult for archival researchers to identify direct measures of the behavioral biases of managers or investors. As a communication vehicle for management, textual disclosures can provide a means for researchers to assess managers' behavioral biases and understand firm behavior. For instance, in annual reports, managers tend to refer to themselves more frequently when firm performance is better. This evidence suggests that managers have the self-serving attribution bias, which could affect their investing and financing decisions [Li, 2011].

Third, communication by managers provides researchers a powerful setting in which to understand managers' incentives and private information sets (i.e., see the world from the eyes of the managers) and therefore better understand firm behavior. For instance, Simon [1997] argues that employee communication patterns during critical decision-making could reveal the underlying organization design.

This paper surveys research on textual analysis of corporate disclosures, such as financial statements, earnings releases, and conference call transcripts. Specifically, I focus on research related to earnings quality, stock market efficiency, and corporate financial policies using the textual information in corporate disclosures. Accounting researchers have examined textual disclosures for a long time (See, for example, Cole and Jones [2005] for a review of the research on Management Discussion and Analysis and Jones and Shoemaker [1994] for a review of the research on content analysis.) The prior literature tends to approach these research questions using small-size samples based on manually coded data. In this survey, I focus on the empirical large-sample textual

analysis built upon the latest development in computational linguistics, natural language processing, and statistics.

Two factors likely have contributed to the growing trend in this line of research. First, a large amount of unstructured textual data has recently become available electronically and accessible to researchers. The Edgar filing system and the availability of conference call transcripts, financial analyst reports, and comment letters to the SEC, among other things, provide researchers with rich textual datasets to test different economic hypotheses. Second, there have been significant developments in research in the fields of computational linguistics, text mining, and machine learning in the past two decades [Manning and Schutze, 1999; Jurafsky and Martin, 2000; Mitchell, 2006]. These developments provide accounting researchers with powerful tools to understand corporate disclosures better [Core, 2001]. From a historical perspective, this research area mirrors areas that developed in the late 1960s, when advances in computing technology and the availability of large databases of accounting and stock prices data brought about much of modern empirical financial accounting research.

In Section 2, I briefly discuss the methodology of textual analysis. In Section 3, I review the existing literature on large-sample textual analysis of corporate disclosures. The literature has developed substantially recently, and to better organize the papers, I discuss them by classifying them into different categories based on the economic questions they examine. In Section 4, I provide a summary, discuss the challenges facing the literature, and propose suggestions for future research.

2.0 Methodology for textual analysis

The fundamental problem of understanding textual disclosure is data reduction. The goal of data reduction is to aggregate the information contained in a large amount of text into manageable numeric variables for further analysis. In this section, I briefly discuss some of the methodology issues with textual analysis (for a summary of the related literature, see Manning and Schutze [1999] and Berry [2004], and Hausser [2001]).

2.0.1 Manual versus computer-based content analysis

Prior studies examining corporate disclosures have largely used a manual content analysis approach to understand the texts. For example, Bryan [1997] examines 250 Management's Discussion and Analysis reports [MD&As] and Callahan and Smith [2004] study 71 firms and 420 firm-years. This manual approach has its advantages: the content analysis can be more precise, detailed, and tailored to the specific research setting.

However, the manual content analysis approach has several disadvantages [Core, 2001; Li, 2010a]. First, the cost of manually collecting these data tends to be high. As a result, most studies have small sample sizes, which may limit the scope of the empirical tests. For instance, to study any potential change in the information content of MD&A disclosures over time, researchers need a relatively large panel of data. This not only limits the generalizability of the empirical results, but the small sample size also means that the empirical tests based on these samples tend to have low power. Second, there is also additional difficulty with replication due to subjectivity in the coding process. The difficulty with replication and limited generalizability of the empirical results tend to limit follow-up studies.

Another approach to understanding textual disclosures is to use computer programs to do the content analysis. This approach relies on computer programs to understand text and reduce them into interesting economic variables. The computer-based approach improves the generalizability of the empirical results and could lead to more follow-up research because such studies can be more easily replicated in subsequent research. In addition, the larger sample sizes obtained using this approach significantly increase the power of the empirical tests, which is critical for testing hypotheses such as stock market efficiency with respect to accounting information.

From a practical perspective, large-sample textual analysis of corporate disclosures can also provide financial statement users with useful information that is not easily available otherwise. Even if one argues that in some settings, it may be sufficient to have a small sample size to test economic hypotheses, the large-sample analysis could uncover information for a much broader sample of firms at a lower cost and in a more timely fashion to help investors and other parties make economic decisions. An analogy here is Google as a search engine: even though the page ranking by Google is not as precise as that of a human coder, the fact that it generates reasonably precise ranking efficiently is very useful for internet users.

2.0.2 Dictionary versus statistical approach

There are two general approaches for conducting content analysis using computer program: a rule-based ("dictionary") approach and a statistical approach. The dictionary approach uses a "mapping" algorithm, i.e., the computer program reads the text and classifies the words (or phrases) in the text into different categories based on some pre-defined rules or categories (i.e., dictionary).[2]

The statistical approach, pioneered by computer scientists and mathematicians, relies on statistical techniques to conduct content analysis [e.g., Manning and Schutze, 1999; Jurafsky and Martin, 2000; Mitchell, 2006]. For example, the algorithm may calculate the statistical correlations between some keywords and the document type to classify the documents.

The statistical approach offers several advantages over the dictionary approach [Li, 2010a]. First, few dictionaries exist that are built for the setting of corporate financial statements and thus may not work well for such a setting.[3] The following example from [Li, 2010a] illustrates this point. Consider the sentence "In addition, the Company has experienced attrition of its Medicare and commercial business in 1998 and 1999 and expects additional attrition." According to the General Inquirer,[4] the sentence has 2 or 10.53% positive words ("expect" and "experience") and no negative words, even though it is obvious that this sentence has a negative tone. Second, the simple dictionary-based approach ignores the context of a sentence. For instance, if a sentence is about cost, then the word "increase" should be treated as negative; however, it is likely to be a positive word if the sentence is about "sales."

Third, the dictionary approach ignores any prior knowledge that researchers may have. For example, if most of the sentences that appear in MD&A reports are neutral, then it might be more efficient to classify a random sentence as being of a neutral tone unless there is strong evidence suggesting otherwise. This point is especially important when analyzing managerial disclosure, because managers have incentives to disclose strategically. Lastly, the statistical approach provides a natural way to validate

classification efficiency. The training data collected during the statistical content analysis are human coded and thus could be used to test the effectiveness of the algorithm.

Indeed, several papers have demonstrated that the dictionaries built by researchers from other fields (e.g., psychologists) may not be suitable for doing content analysis for corporate disclosures [Henry and Leone, 2010; Li, 2010a; Loughran and McDonald, 2011]. Of course, this does not mean that the results in the existing research based on those dictionaries [Kothari, Li, and Short, 2009; Davis, Piger, and Sedor, 2008] are invalid; rather, it suggests that those results hold *despite* the relatively noisy dictionaries used. However, because any empirical tests are joint tests of the methodology and the economic hypotheses, it could well be the case that on many other occasions, researchers do not find any significant empirical results because they have used dictionaries not tailored to the specific disclosure settings that they are examining.[5] Future researchers would find it much more appealing either to use tailored dictionaries or rely on a statistical approach for the content analysis.

2.0.3 Variables of interest from textual information

Conceptually, there are at least three disclosure characteristics that are potentially interesting to researchers [Li, 2010a]: the amount of disclosure, the tone, and the transparency (or readability). The amount of disclosure is relatively easy to measure as it typically involves the length or the size of the file [Peterson, 2008; Li, 2008; Leuz and Schrand, 2009; You and Zhang, 2009; Miller, 2010; Lee, 2010; Merkley, 2011]. However, these papers that analyze the length of a document (or a section of the document) often treat it as a measure of complexity or transparency of the disclosure rather than the amount of disclosure. The truth is perhaps somewhere in between: the

length of disclosures is likely to capture the level or the amount of disclosure as well as the complexity of disclosure. Thus, future researchers using disclosure length as a measure of complexity or transparency need to control for the underlying business or operational complexity.

To help categorize the literature, Table 1 tabulates some of the papers that I will discuss in the next section based on the dependent and independent variables of these studies. Most of these papers focus on either the tone or the complexity of disclosures and their implications for earnings or stock prices.

3.0 Recent research on textual disclosures

In this section, I discuss some papers written in the last ten years (published or working papers) on textual disclosures. Whenever appropriate, I will also discuss potential future research opportunities.

3.0.1 Information content of corporate textual disclosures

A natural question with respect to corporate textual disclosures is whether they have information content. The alternative hypothesis is that these disclosures are boilerplate generic disclosures and not informative [SEC, 2003; Bloomfield, 2008]. A challenge for researchers examining this question is to convince the readers that the hypothesis is not a "straw man." In other words, some readers may believe that there is likely to be at least some information content in textual disclosures and that the hypotheses tested lack tension. Therefore, researchers need to clearly layout the factors that potentially make the disclosures informative and those that are likely to cause them to be less informative.

Both economic theories and institutional factors need to be considered to build the tensions when testing the hypothesis that textual disclosures have information content. Take the example of management discussions in 10-K and 10-Q filings. In 1980, the Securities and Exchange Commission [SEC] mandated that public companies include in their annual reports a section for Management's Discussion and Analysis of Financial Condition and Results of Operations [MD&A]. The MD&A is intended to assess an enterprise's liquidity, capital resources, and operations and is an important component of the 10-K filings.

Li [2010a] studies whether the forward-looking statements in the MD&As are informative about future performance. He argues that ex ante whether MD&A disclosures are truly informative remains an open empirical question because of the following two sets of reasons. On the one hand, consistent with the SEC's intention, the MD&A is one of the most read and most important components of the financial statements [Tavcar, 1998; Knutson, 1993; Rogers and Grant, 1997]. Furthermore, the safe harbor provisions of the Private Securities Litigation Reform Act of 1995 encourage more forward-looking information and should make MD&A disclosures more informative. On the other hand, the MD&A might not be informative because companies have concerns over proprietary costs [Verrecchia, 1983] and uncertainties about the judicial interpretation of safe harbor protection. Also, the MD&As are not required to be audited [Hufner, 2007] and the SEC worries that MD&As mainly include substantial boilerplate disclaimers and disclosures, generic language, and immaterial detail [SEC, 2003]. These two sets of arguments build the tension for the hypothesis to test the information content of MD&A disclosures. Of

course, the economic theories and institutional factors are likely to be different when analyzing different types of disclosures.

The literature has used both the correlations with current or future fundamentals and the associations with contemporaneous market reactions to assess the information content of textual disclosures. Most papers have found that the textual disclosures are informative with respect to both fundamentals and market reactions. First, the level of optimism in earnings releases is positively associated with the market's short-term response to the announcement [Henry, 2008; Demers and Vega, 2008; Davis, Piger, and Sedor, 2008] and a similar relation holds for 10-K SEC filings [Loughran and McDonald 2009; Feldman, Govindaraj, Livnat, and Segal, 2010]. For instance, using Diction, a dictionary software for content analysis, Davis, Piger, and Sedor [2008] find evidence that suggests that managers use optimistic and pessimistic language in earnings press releases to provide investors with information about expected future firm performance and that the market responds to these disclosures. Specifically, they find a positive (negative) association between levels of optimism (pessimism) in earnings press releases and future return on assets. Moreover, higher unexpected optimism (pessimism) is also found to be associated with higher (lower) returns around the earnings press release.

On the other hand, Li and Ramesh [2009] find that the significant market reaction surrounding 10-Q filings is limited to filings that release earnings information for the first time. They also find that a similar market reaction is obtained in 10-K reports only when they are filed around calendar quarter-ends. Overall, the evidence in Li and Ramesh [2009] indicates limited reaction from the market around 10-K and 10-Q filings dates, suggesting that there is limited information content in the filings that is incremental to

what the market already knows. How can we reconcile the finding that there is little market reaction to the filings in Li and Ramesh [2009] with the findings in Loughran and McDonald [2009] and Feldman, Govindaraj, Livnat, and Segal [2010] that market reactions around filings are significantly associated with the content in the filings? One potential reason is that Li and Ramesh [2009] focus on the *on average* effect around the filings, while the other papers examine the market reaction conditional on the tone of the disclosures.[6]

Several papers also find that the optimism in corporate disclosures is associated with *future* firm performance. Using a Bayesian statistical learning approach, Li [2010a] finds that when managers are more optimistic when discussing future events in MD&As, future earnings and liquidity are indeed much better, even after controlling for stock returns and other predictors of future performance. This suggests that management discussions are informative with respect to future firm performance. Interestingly, Li [2010a] does not find a positive relation between MD&A optimism and future performance if several dictionaries (the General Inquirer, LIWC, and Diction) are used to do content analysis, suggesting that these dictionaries may not work with corporate disclosures.

Brown and Tucker [2011] confirm the informativeness of MD&As using a change specification. They introduce a measure for the degree to which the MD&A differs from the previous disclosure. They document that firms with larger economic changes modify the MD&A more than those with smaller economic changes and that the magnitude of stock price responses to 10-K filings is positively associated with the MD&A

modification score. This evidence further supports the argument that MD&A has information content.

Some papers also validate the information content of textual disclosures using the corresponding numeric data related to the textual information as a benchmark. Levine and Smith [2006] analyze a large sample of Critical Accounting Policy disclosures from SEC filers and examine the extent to which CAP disclosures correlate with existing financial statement information, provide new information, and correlate with existing measures of accounting quality. They find that disclosures about critical accounting policies can be directly traceable to a specific balance sheet account. For instance, firms that cite accounts receivable as a critical accounting policy have, on average, higher accounts receivable as a percentage of total assets and a higher time-series variance of scaled accounts receivable than firms that do not. Similar results hold for the other balance sheet accounts that they examine.

While it is interesting to show that on average textual disclosure is informative, future research is likely to be more focused on the cross-sectional variations or variations over time in the information content of the textual disclosures. For instance, the information content of the MD&A is likely to be a function of the firm's performance, litigation risk, and information environment based on economic theories on disclosures. Merkley [2011] finds that as current performance decreases, firms provide more R&D-related disclosure. His results are based on an analysis of within-firm variation and year-to-year changes. The results also indicate that this relation is more pronounced for firms that place more importance on R&D and for firms with higher outside monitoring.

Evidence on the cross-sectional variations in corporate textual disclosures' information content can build on the existing "on average" results and test economic theories about disclosure behavior better. This is especially important because of the evidence in Li and Ramesh [2009] that the average market reaction around corporate filings is limited, suggesting that there is perhaps substantial variation in the information content of the filings. Therefore, it is interesting to explore the cross-sectional variations in the information content of these disclosures. From the perspective of the practitioners, the empirical results documented here also provide practitioners with guidance on the circumstances under which they can benefit more from reading corporate disclosures more carefully.

3.0.2 The implications of corporate textual disclosures for earnings quality

As discussed in the previous section, the textual information can be considered as the data generating function of the numeric data. The second line of research on textual disclosures links them to the numeric accounting data by testing economic hypotheses using textual disclosures as a context. In this section, I survey research that examines earnings quality in the context of textual disclosures made by managers.

Virtually every financial accounting text has discussions on earnings quality. Yet measuring or understanding earnings quality is a challenging task [Dechow and Schrand, 2004].[7] Prior research on earnings quality generally relies on one of two approaches: studying the properties of accounting numbers or extracting information from stock prices [Sloan [1996], Dechow and Dichev [2002], and Ecker, Francis, Kim, Olsson, and Schipper, 2006].

Examining the textual disclosures or corporate disclosure decisions by managers could prove to be a fruitful area to study earnings quality for two reasons. First, managers make many decisions based on future profitability, and arguably have more precise and complete information about their firm's profitability than do other stakeholders. Therefore, to the extent that information asymmetry exists between managers and outsiders, the earnings quality inferred from managerial disclosures and decisions can be incremental or even superior to existing empirical measures. Second, earnings quality is a function of management incentives and the textual communications by managers can provide a means for researchers to assess these incentives.

Li, Lundholm, and Minnis [2011] is an example of studying earnings quality based on management's views on firms' competitive environment extracted from textual disclosures. An important aspect of consideration when assessing a firm's profitability and earnings quality is its competitive environment. Standard microeconomic theory argues that when a firm faces strong competition, its abnormal profits mean revert faster. Hence, the earnings quality of such firms is lower as compared to firms facing little competition. Li, Lundholm, and Minnis [2011] test this hypothesis using a very simple textual analysis of the firm's 10-K filing to gauge its competitive environment. They find that a firm's profit mean reverts faster when the frequency of references to competition in the 10-K is higher, suggesting that the information about competition from a firm's 10-K report contains information about earnings quality.

Another contribution of this paper is that it proposes and validates a *firm-specific* measure of competition that could be used to test other interesting economic hypotheses. The most common measures of competition used in the prior literature, the Herfindahl

index and the four-firm concentration ratio, focus on the distribution of production across firms within the industry. However, these measures are only defined at the industry level, while there is potentially considerable variation in competition within an industry.

Examining the implications of annual report readability for earnings quality, Li [2008] studies earnings quality by exploiting possible management incentives based on textual information. The study is motivated by the "management obfuscation hypothesis," i.e., if markets react less completely to information that is less easily extracted from public disclosures, then managers have more incentive to obfuscate information when firm performance is poor [Bloomfield, 2002]. This "management obfuscation hypothesis" suggests that the earnings quality is lower when managers are not forthcoming in the disclosure. Li [2008] hypothesizes that the positive earnings of firms with more complex annual reports are less persistent and the negative earnings of such firms are more persistent in the immediately following years. Using the Fog index from the computational linguistics literature [Gunning, 1952] to measure the readability of corporate annual reports, he finds strong evidence consistent with this hypothesis.

For future research, studies linking textual disclosures to the numeric accounting data can be extended along two dimensions. First, additional accounting items (other than earnings) could be examined in the context of textual disclosures. For example, accruals are computed with estimation errors [Dechow and Dichev, 2002]. The level of estimation errors can be potentially assessed using the textual disclosures to evaluate the quality of accruals. Second, it may be fruitful to examine the Notes to the Financial Statements and the Critical Accounting Policy sections of corporate filings more carefully, as these sections contain more direct information about the data generating functions of financial

data. For instance, if there are significant year-to-year changes in the Critical Accounting Policies, how would earnings properties change and how would financial analysts deal with these issues?

3.0.3 Valuation of textual information and test of market efficiency

Testing stock market efficiency with respect to textual information is potentially a very fruitful area for research. This is because the information processing cost for textual disclosures is likely to be higher compared with numeric data and the stock market may not be as efficient with respect to the textual information when investors have limited attention [Hirshleifer and Teoh, 2003].

Information extracted from textual disclosures can be used to test market efficiency in two ways: as the main variable or as a contextual variable. In the first case, the research question is: does the stock market understand and fully price the information contained in the textual corporate disclosures? In the second case, the research question is: how does textual information contained in corporate disclosures affect the market's understanding of other information?

Several papers examine the implications of textual information for stock price drift. Post–earnings announcement drift, or PEAD, initially proposed by Ball and Brown [1968] and later more carefully examined in Bernard and Thomas [1989, 1990], is the tendency for a stock's cumulative abnormal returns to drift in the direction of an earnings surprise for several weeks (even several months) following an earnings announcement. One of the most widely accepted explanations for the effect is investor under-reaction to earnings announcements. It is a natural question, then, to examine whether the stock market also under-reacts to information in the textual disclosures. Given the relatively

higher information processing cost compared with that for numeric data (such as earnings), variables extracted from the textual disclosures seem more likely to be mispriced.

You and Zhang [2009] document that the market underreaction to 10-K filings is stronger for firms with longer 10-K filings. Feldman, Govindaraj, Livnat, and Segal [2010] take the analysis one step further by exploring whether the information contained in the MD&A section of 10-Qs and 10-Ks is associated with stock price drifting. They classify words into positive and negative categories to measure the tone change in the MD&A section relative to prior periodic SEC filings. They find that tone change signals help predict the subsequent quarter's earnings surprise. Consistent with the hypothesis that the market under-reacts to this information, they show that management's tone change adds significantly to portfolio drift returns in the 2-day window after the SEC filing date through 1 day after the subsequent quarter's preliminary earnings announcement, beyond financial information conveyed by accruals and earnings surprises.

Li [2010b] also directly examine whether the stock market fully understands the information content of textual disclosures. Specifically, he focuses on the disclosure in 10-K filings about risk and uncertainty and finds that the market does not fully understand the implications of firms' risk disclosure for future profitability. Overall, the evidence in Li [2010b] and Feldman, Govindaraj, Livnat, and Segal [2010] support the argument that the market may not fully understand the implications of textual disclosures.

Several papers examine two capital market anomalies, the post-earnings announcement drift and the accrual anomaly, by exploring textual information. Lee

[2010] tests and finds that less of the earnings-related information is incorporated into the firm's stock price during the three days following the 10-Q filings for firms with longer or less readable 10-Qs. She also finds that there is greater information asymmetry during the 10-Q filing window for firm-quarters with quarterly reports of lower readability. This evidence supports the notion that the transparency of corporate textual disclosures affects how the stock market processes earnings information. Note that unlike Feldman, Govindaraj, Livnat, and Segal [2010], which focuses on the price drift based on MD&A tone sorting, Lee [2010] is about the implications of textual information for the earnings-based price drift.

Following Sloan [1996], a large literature has emerged examining the possible explanations for the accrual anomaly. Both accruals and corporate textual disclosures can be considered signals from managers that contain information about future firm performance. Information in the textual corporate disclosures could therefore potentially help mitigate the accrual anomaly. Several papers examine this question empirically. Li [2010a] examines the implications of MD&As for the mispricing of accruals. To the extent that the MD&A tone provides a more direct prediction about future outlook than do accruals, the information in the MD&A tone about future performance is more salient. He distinguishes between situations where managers "warn" or do not "warn" investors about the future performance implications of the accruals in MD&As. Specifically, he finds that the tone in MD&As mitigates the mispricing of accruals. When managers "warn" about the future performance implications of accruals (i.e., the MD&A tone is positive (negative) when accruals are negative (positive)), accruals are not associated with future returns.

Li, Lundholm, and Minnis [2011] also examine the implications of textual information for the accrual anomaly. One hypothesis about the accrual anomaly is that this anomaly arises because the market fails to fully account for diminishing marginal returns to investment, and is surprised when future earnings change in response to changes in asset growth. The results in Li, Lundholm, and Minnis [2011] take the argument one step further. They find that the mispricing of changes in net operating assets (i.e. accruals) is largest when a firm faces stronger competition as measured using managers' disclosures about competition in 10-K filings.

To summarize, the availability of the large amount of textual corporate disclosures adds additional dimensions to the examination of market efficiency. Future research can focus on other types of valuation implications of the textual information in corporate disclosures. For example, the asset pricing literature finds that returns to the momentum trading strategy are a function of trading volume [Lee and Swaminathan, 2000]. Given that disclosures are likely to be correlated with a firm's information environment, it is interesting to examine how disclosure characteristics interact with the momentum strategy.

3.0.4 Textual disclosure and firms' information environment

There is a large literature in accounting that examines the role of information intermediaries such as financial analysts in the capital market and how they interact with the corporate financial reporting environment [Kothari, 2001; Beyer, Cohen, Lys, and Walther, 2011]. Recently, the manners in which investors and these information intermediaries use the textual information disclosed by managers become the focus of several papers. These papers examine how textual disclosures are associated with

investors' trading behavior, analyst following properties, and other information environment factors.

A hypothesis that is (often jointly) tested in many papers is that more complex disclosures increase the information processing cost for investors, especially for small investors [Grossman and Stiglitz, 1980; Bloomfield, 2002]. Miller [2010] estimates the complexity of a 10-K using four measures (two measures related to length and two measures related to readability) and, consistent with the information processing cost hypothesis, finds evidence that more complex 10-Ks are associated with lower trading volume and that this effect of the complexity of the 10-K on trading activity is mostly driven by small investors.

Using different designs and measures of readability, Loughran and McDonalds [2010] also find a relationship between improved 10-K readability and increased small investor trading. In addition to the more standard measure of readability from the computational linguistics literature (the Fog Index and the Flesch Score), Loughran and McDonald construct their own readability score based on specific examples provided by the SEC during the plain English initiative. The evidence in Miller [2010] and Loughran and McDonalds [2010] is consistent with the hypothesis that the complexity of corporate disclosures affects small investors disproportionately.

Several studies hypothesize and find that financial analysts' behavior is different for firms with different textual disclosures. Lehavy, Li, and Merkley [2011] examine the effect of the readability of 10-K filings on the behavior of sell-side financial analysts. They find that the number of analysts following a firm, the amount of effort incurred to generate their reports, and the informativeness of their reports are greater if a firm has

less readable 10-Ks. Additionally, they find that less readable 10-Ks are associated with greater dispersion, lower accuracy, and greater overall uncertainty in analyst earnings forecasts. Overall, these results are consistent with the prediction of an increasing demand for analyst services for firms with less readable communication and a greater collective effort by analysts for firms with less readable disclosures.

Consistent with the results in Lehavy, Li, and Merkley [2011], Kravet and Muslu [2010] find that increases in risk disclosures are associated with increases in the number of earnings forecasts, more dispersed and divergent forecast revisions, and increases in trading volume. They also document that increases in risk disclosures are also associated with increased forecast accuracy.

However, Brown and Tucker [2011] document that analyst earnings forecast revisions are unassociated with the year-to-year changes in MD&As, even though the changes in MD&As are positively and significantly associated with the stock market reaction. They conclude that this evidence suggests that financial analysts do not use MD&A information. Two factors could potentially explain the different results in Brown and Tucker [2011] as compared to those in Kravet and Muslu [2010]. First, the changes in disclosures may not necessarily affect the level of analyst forecasts; rather, they are more likely to affect the second moment of the forecasts (e.g., dispersion), which is examined by Kravet and Muslu. The other possible reason for the different results is that Brown and Tucker focus on the changes in the entire MD&A, whereas Kravet and Muslu examine the risk-related disclosures specifically.

Taking a different approach, Kothari, Li, and Short [2009] study a broad range of information environment measures, including the cost of capital, and link them to the

textual nature of the disclosures. They find that when there are favorable disclosures, the firm's risk (as proxied by the cost of capital, stock return volatility, and analyst forecast dispersion) declines. On the other hand, unfavorable disclosures are accompanied by increases in each of the risk measures used in their study. Note that unlike the other papers discussed in this section, Kothari, Short, and Li [2009] focus on the tone of the disclosures, rather than the readability or transparency.

Finally, Muslu, Rahhakrishnan, Subramanyam, and Lim [2010] hypothesize and find that more firms make more forward-looking disclosures in the MD&A sections of 10-K filings when their stock prices have poor informational efficiency (i.e., when these stock returns poorly reflect future earnings information). They also find that the greater levels of forward-looking MD&A disclosures improve the informational efficiency of stock prices for such firms. In summary, current research finds significant evidence that firms' information environment is impacted by their textual disclosures and this effect is incremental to those from the numeric financial data.

3.0.5 Textual disclosure and litigation

Another interesting area of research is to examine the implications of textual disclosures for litigations. This is because in their complaints plaintiffs frequently include qualitative statements from corporate disclosures among the specific statements cited as misleading, and courts have judged qualitative statements to be material, allowing them to survive the defendant's motion to dismiss [Rogers, Van Buskirk, and Zechman, 2010]. Therefore, linking textual characteristics of disclosures to litigation or litigation risk a company faces could shed light on how firms behave in a litigious environment.

Building upon the evidence in Skinner [1994] and Kasznik and Lev [1995] that firms

are more likely to preempt large, negative earnings surprises than other types of earnings news, Nelson and Pritchard [2007] investigate the disclosures of "cautionary language" and their association with litigation risk. Their findings indicate that firms facing greater litigation risk disclose more cautionary language. This evidence is consistent with managers attempting to reduce expected litigation costs by altering their disclosure choices.

Rogers, Van Buskirk, and Zechman [2010] examine the relation between disclosure tone and shareholder litigation to determine whether managers' use of optimistic language increases litigation risk. Using both general-purpose and context-specific text dictionaries (the Diction software and dictionaries from Henry [2008] and Loughran and McDonald [2011]) to quantify tone, they find that plaintiffs target more optimistic statements in their lawsuits and that sued firms' earnings announcements are abnormally optimistic when compared to similar firms. They also document that the litigation risk is even higher when managers are both optimistic and engaging in abnormal selling.

3.0.6 Implications of corporate textual disclosures for organizational design and corporate financial policies

Textual disclosures can also reveal considerable insights on larger economic questions. Organizational theories [e.g., Simon, 1997] suggest that employee communication patterns during critical decision-making reveal most clearly the underlying patterns of the organization structure, which is typically not readily available to researchers.

A couple of recent papers explore the textual disclosures by managers to answer questions related to corporate organization design and financial policies. Li, Minnis,

Nagar, and Rajan [2010] analyze the information contained in conference call transcripts to test economic theories on the allocation of authority within organizations. Theories of the firm such as Aghion and Tirole [1997] distinguish *formal* authority from *real* authority: a manager could be formally responsible for a decision, but in reality may acquiesce to her better-informed subordinate. Li, Minnis, Nagar, and Rajan [2010] propose that the extent to which CEOs communicate in earnings conference calls can serve as a measure of their real authority over top management. The empirical results indicate that the real authority measure based on conference call transcripts is distinct from formal authority measures of CEOs, and is significantly associated with the theoretical organizational factors proposed by Aghion and Tirole [1997].

Another promising venue for large-sample textual analysis in accounting research is measuring management (behavioral) characteristics and examining their implications for corporate decisions. While there is pervasive evidence that individuals exhibit different behavioral biases in lab experiments or surveys, relatively little *direct* evidence based on archival data exists on whether corporate executives have these biases and whether theses biases have economic consequences due to data limitations. Li [2011] provides such evidence using textual disclosures by studying managers' self-serving attribution bias [SAB] and its implications for corporate financial policies. The SAB refers to individuals taking responsibility for successful outcomes but blaming circumstances or other persons for unsuccessful outcomes [Libby and Rennekamp, 2010]. Li [2011] finds that managers tend to use more first-person pronouns (relative to second- and third-person pronouns) in the MD&A of the 10-K filings when firm performance is better. The evidence in Li [2011] also supports the argument that managers with more

SAB are more overconfident and their firms tend have make less optimal investment decisions, have higher leverage, are more likely to repurchase stocks, and are less likely to issue dividends. Collectively, the evidence suggests that managers have self-serving attribution bias and this bias has implications for corporate policies.

To summarize, relatively little research has been conducted utilizing the textual information in disclosures to answer more general economic research questions. With more detailed knowledge about institutional features and corporate disclosure practices, accounting researchers have a comparative advantage in working in this area and can contribute to the economics and management literature.

4.0 Challenges and future research

In this section, I discuss some of the challenges facing the literature in the large-sample textual analysis of corporate disclosures and propose some possible opportunities for future research. Significant progress has been made in the literature on testing interesting economic hypotheses using the newly available textual data. In this survey, I have mostly discussed papers written in the last ten years that examine the textual corporate disclosures using large samples. In this process, I have omitted papers on large-sample analysis of other textual disclosures, such as financial analysts' reports [De Franco, Vasari, Vyas, and Wittenberg-Moerman, 2010; Huang, Zang, and Zheng, 2010], media [Miller, 2006; Core, Guay, and Larcker, 2008; Soltes, 2009; Bushee, Core, Guay, and Hamm, 2010], and internet message boards [Antweiler and Frank, 2004].

However, several challenges exist for researchers who want to advance this literature. First, much of the existing research has a strong methodology flavor. This is not surprising because the textual disclosure is a newly available data source and

conducting large-sample empirical analysis using this data requires some new methodology. Because the empirical inferences are based on joint tests of the methodology and the economic hypotheses examined in the papers, not surprisingly, existing research often devotes a significant amount of effort to developing the methodology. A consequence of this emphasis on methodology is that some of the hypotheses are not well developed or sometimes have a "straw man" flavor. Future research is likely to benefit more from developing hypotheses that are more closely tied to economic theories, from both classical economics and behavioral and experimental economics.

Second, like many other empirical studies in the accounting literature, the large-sample textual analysis literature also raises concerns about endogeneity. The endogeneity concern could be related both to omitted variables and reverse causality issues. For instance, Bloomfield [2008] argues that bad news may be inherently more difficult to communicate and hence the annual reports of poorly performing firms are more complex. This alternative explanation could change the interpretation of results in Li [2008]. Furthermore, many of the studies discussed in the previous section are basically association studies. For example, the effect of disclosure readability on analyst following properties documented in Lehavy, Li, and Merkley [2011] is not established as a causality effect. Future research needs to devote more efforts to mitigating these endogeneity concerns for better empirical identification. While it may be difficult to find good instruments, it seems important for researchers to conduct robustness tests using a change specification whenever appropriate.

Third, for some studies, especially for studies that attempt to measure managerial

characteristics using the corporate filings, there could be some concerns about who actually writes the corporate disclosures. Much of corporate textual disclosure is the output of the collaborative efforts of managers, auditors, lawyers, and public relations staff. Therefore, measures of management characteristics based on these disclosures need to be examined carefully to ensure that the results are not driven by considerations such as strategic disclosure decisions. For instance, the result in Li [2011] that managers use first-person pronouns more often in the MD&A section of 10-Ks when firm performance is better may not necessarily be driven by their self-serving attribution bias. Rather, it is possible that the public relations people in these companies want to make the management look better to investors by referring to them more often during good performance periods. To mitigate these concerns, it is useful to triangulate the empirical evidence using different sources of textual disclosures. For example, the statements made by managers in the Q&A section of the conference calls are more spontaneous and thus less likely to be influenced by the staged preparation.

Throughout the review in previous sections, I have discussed possible opportunities for future research. Below I suggest some more specific ideas:

- The interactions between different textual disclosure channels. With the exception of Davis and Tama-Sweet [2009], most existing papers focus on one specific textual disclosure channel. However, the different disclosure channels (e.g., corporate filings, conference calls, and earnings releases) are likely to be influenced by similar economic factors and are jointly determined.

- Management incentives and the features of the corporate textual disclosures. Much of current research focuses on the informativeness of these disclosures

without considering the incentives of managers due to compensation, contracting, or capital market considerations. Exploiting these management incentives to understand the large-sample textual disclosures in future research can not only test more interesting hypotheses, but also tie the research to the more established research on disclosures in the accounting literature.

- Cross-country differences in the characteristics of corporate textual disclosures and their implications. The way that management communicates with investors is likely to be a function of the culture and institutional features of the country that the firm is from. Exploring these cross-country differences and linking them to textual disclosures could provide evidence on the determinants of these disclosures and how investors respond to them differently. A natural place to explore these variations is the firms that cross-list in the U.S., since researchers have access to the disclosures made by these firms and it is also relatively easy to collect financial data for these firms.

- Corporate governance and the informativeness of textual disclosures. How do ownership structure, board composition, and other governance affect corporate disclosures? What are the roles of disclosures in corporate governance? Large-sample analysis of corporate disclosures could help answer these questions posed by theory papers such as Hermalin and Weisbach [2011].

- Bankruptcy prediction and textual information. There is a large literature on bankruptcy prediction using both accounting and stock price data. Whether the textual information contained in the financial statements and other disclosures have predictive power in bankruptcy forecasting incremental to existing variables is interesting. Given the evidence from existing research that these textual disclosures often have incremental

information content in predicting future performance, it seems likely that they can also predict bankruptcy.

- Management turnover and changes in disclosures. Current research typically examines the determinants and consequences of disclosures using cross-sectional data. One criticism of this approach is that omitted firm characteristics may drive the results. Examining change in disclosures around management turnover could provide change-based evidence that is not subject to this concern.

- Improvement in the large-sample textual analysis methodology. There is certainly space for improving the accuracy and efficiency of the textual analysis of corporate disclosures. For instance, most of the current papers focus on individual words or phrases. It might be useful to conduct more modeling based on n-gram analysis [Manning and Schutze, 1999; Jurafsky and Martin, 2000]. Of course, as discussed previously, the danger here for researchers is that the study becomes a pure methodology paper in the process of pursuing improvement in methodology. In addition, how much additional benefit researchers can get from refinement in the methodology is uncertain and remains an empirical question.

ANNOTATED BIBLIOGRAPHY

1. Feldman, R., S. Govindaraj, J. Livnat, and B. Segal. 2010. "Management's Tone Change, Post Earnings Announcement Drift and Accruals." *Review of Accounting Studies.* Volume 15, Number 4, December 2010, pp. 915-953(39).

This paper examines the information content of management discussions in corporate filings. The authors classify words into positive and negative categories to measure the tone of management discussions. They find that tone change is significantly associated with both short-window return around the filing date and the drift returns in the post-filing period and this association is incremental to that due to accruals and earnings surprises. The evidence in this paper suggests that management discussions are informative and the market under-reacts to this information.

2. Kothari, S.P., X. Li, and J. E. Short. "The Effect of Disclosures by Management, Analysts, and Financial Press on the Equity Cost of Capital: A Study Using Content Analysis." *The Accounting Review* 84 (2009): 1639–70.

This paper studies the implications of disclosures by managers, financial analysts, and the media for firms' information environment. They find that when there are favorable disclosures, the firm's risk (as measured by the cost of capital, stock return volatility, and analyst forecast dispersion) declines. This evidence suggests that textual disclosures affect firms' risk and information environment.

3. Li, F. 2008. Annual Report Readability, Current Earnings, and Earnings Persistence. *Journal of Accounting and Economics* 45: 221–47.

This paper hypothesizes that annual report readability captures management obfuscation: when a firm's annual report is more complex, its managers might be obfuscating information. The author relies on the Fog index and the length of 10-K filings to capture the readability of 10-Ks. The paper finds that (1) when a firm has poor performance, its 10-K tends to be harder to read; (2) when a firm's 10-K is more complex, its profit is less persistent. These results support the obfuscation hypothesis.

4. Li, F, 2010. The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach, *Journal of Accounting Research*. Vol. 48 No. 5 December:1049-102.

The author examines the information content of the forward-looking statements in the Management Discussion and Analysis section of the 10-Ks and 10-Qs. It relies on a Bayesian machine learning approach to conduct content analysis. The results indicate that when managers are more optimistic in these forward-looking statements, future performance is better, suggesting that the management discussions have information content. Furthermore, when the implications of management discussions about future

performance are consistent with those of accruals, the accruals are less likely to be mispriced.

5. Lehavy, R., F. Li, and K. Merkley. 2011. The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts. *The Accounting Review*, May.

This paper studies how 10-K readability is associated with sell-side financial analyst following and the properties of their earnings forecasts. Using the measure of readability from Li [2008], the authors document that analyst following, the amount of effort incurred to generate their reports, and the informativeness of their reports are greater for firms with less readable 10-Ks. Furthermore, less readable 10-Ks are associated with greater dispersion, lower accuracy, and greater overall uncertainty in analyst earnings forecasts. The results in the paper suggest that less readable management communication leads to an increasing demand for analyst services and a greater collective effort by analysts.

REFERENCES

Antweiler, W. and M. Z. Frank, 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *Journal of Finance* 59, 1259-1294.

Ball, R. and P. Brown. 1968. An empirical evaluation of accounting income numbers, *Journal of Accounting Research*, Autumn 1968, pp. 159-178

Bernard, V. L., and J.K. Thomas, 1989. Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium? *Journal of Accounting Research*, 27, 1-36.

Bernard, V. L., and J.K. Thomas, 1990. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. *Journal of Accounting and Economics*, 13(4), 305-340

Berry, M.W., 2004. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer-Verlag, New York.

Beyer, A., D. Cohen, T. Lys, and B. Walther, 2011. The Financial Reporting Environment: Review of the Recent Literature, *Journal of Accounting and Economics*.

Bloomfield, R.J., 2002. The "incomplete revelation hypothesis" and financial reporting. *Accounting Horizons* 16. 233–243.

Bloomfield, R.J., 2008. Discussion of "Annual Report Readability, Earnings and Earnings Persistence," *Journal of Accounting and Economics*. August 2008. Vol. 45, Issues 2-3, 248-252.

Brown, S., and J. Tucker. 2011. "Large-sample Evidence on Firms' Year-over-Year MD&A Modifications." *Journal of Accounting Research*.

Bryan, S. H. "Incremental Information Content of Required Disclosures Contained in Management Discussion and Analysis." *The Accounting Review* 72 (1997): 285–301.

Bushee, B., J. Core, W. Guay, and S. Hamm., 2010. The Role of the Business Press as an Information Intermediary, *Journal of Accounting Research* 48: 1-19.

Callahan, C. M., and R. E. Smith. "Firm Performance and Management's Discussion and Analysis Disclosures: An Industry Approach." Working paper, University of Arkansas-Fayetteville, 2004.

Choudhary, P., K. J. Merkley, and J. D. Schloetzer. 2010. "Timeliness of Form 10-K, Firm Financial Per formance and Information Uncertainty." Working paper, Georgetown University.

Cole, C. J., and C. L. Jones. “Management Discussion and Analysis: A Review and Implications for Future Research.” Journal of Accounting Literature 24 (2005): 135–74.

Core, J. E. 2001. A Review of the Empirical Disclosure Literature: Discussion. *Journal of Accounting and Economics.* 31: 441–56.

Core, J., W. Guay., and D. Larcker. 2008. The Power of the Pen and Executive Compensation, *Journal of Financial Economics* 88: 1-25.

Davis, A. K., J. Piger, and L. M. Sedor. 2008. Beyond the Numbers: Managers' Use of Optimistic and Pessimistic Tone in Earnings Press Releases. *SSRN eLibrary*.

Davis, A.K., and I. Tama-Sweet., 2009. Managers’ use of language across alternative disclosure outlets: Earnings press releases versus MD&A. Working paper. University of Oregon.

De Franco, G., P. Vasari, D. Vyas, and R. Wittenberg-Moerman, 2010. A Debt Analyst’s View of Debt-Equity Conflicts of Interest. Working paper, University of Toronto.

Dechow, P. M., and I. D. Dichev. 2002. “The Quality of Accruals and Earnings: The Role of Accrual Estimation Errors.” *The Accounting Review* 77 (2002): 35–59.

Dechow, P.M., and C.M. Schrand. 2004. *Earnings Quality*, first ed. Research Foundation of CFA Institute, Charlottesville, Virginia.

Demers, E. A., and C. Vega. 2008. Soft Information in Earnings Announcements: News or Noise? *SSRN eLibrary*.

Ecker, F.; J. Francis; I. Kim; P. Olsson; and K. Schipper. 2006. “A Returns-Based Representation of Earnings Quality.” *The Accounting Review* 81 (2006): 749–80.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal. 2010. “Management’s Tone Change, Post Earnings Announcement Drift and Accruals.” *Review of Accounting Studies.* Volume 15, Number 4, December 2010 , pp. 915-953(39).

Grossman, S. and J. Stiglitz. 1980. On the impossibility of informationally efficient markets, *American Economic Review* 70 (June): 393-407.

Gunning, R. 1952. *The technique of clear writing.* McGraw-Hill International Book Co; New York, NY.

Hausser, R., 2001. *Foundations of Computational Linguistics.* Springer-Verlag. New York.

Henry, E. “Are Investors Influenced by How Earnings Press Releases Are Written?” *Journal of Business Communication* 45 (2008): 363–407.

Henry, E., and A. J. Leone. “Measuring Qualitative Information in Capital Markets Research.” Working paper, University of Miami, 2010.

Hermalin, B.E., and M.S. Weisbach. 2011. Information Disclosure and Corporate Governance, *Journal of Finance*, forthcoming.

Hirshleifer, D. and S. Teoh. 2003. Limited Attention, Financial Reporting and Disclosure. *Journal of Accounting and Economics*: 337–86.

Huang, A. A., Zang, and R. Zheng. 2010. Informativeness of Text in Analyst Reports: A Naïve Bayes Machine Learning Approach. Working paper, HKUST.

Hufner, B. 2007. The SEC’s MD&A: Does It Meet the Informational Demands of Investors? A Conceptual Evaluation. *Schmalenbach Business Review* 59: 58–84.

Jones, M.J., Shoemaker, P.A., 1994. Accounting narratives: a review of empirical studies of content and readability. *Journal of Accounting Literature* 13, 142.

Jurafsky D., and J. H. Martin, 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ: Prentice Hall

Kahneman, D. 2003. Maps of Bounded Rationality: Psychology for Behavioral Economics, *The American Economic Review, 93(5)*, pp. 1449-1475, December 2003

Kasznik, R., and B. Lev. 1995. “To Warn or Not to Warn: Management Disclosures in the Face of an Earnings Surprise.” *The Accounting Review* 70 (1995): 113-134.

Knutson, P. H. *Financial Reporting in the 1990s and BEYOND*. Charlottesville, VA: Association for Investment Management and Research, 1993.

Kothari, S., 2001. Capital markets research in accounting. *Journal of Accounting & Economics* 31, 105–231.

Kothari, S.P., X. Li, and J. E. Short. “The Effect of Disclosures by Management, Analysts, and Financial Press on the Equity Cost of Capital: A Study Using Content Analysis.” *The Accounting Review* 84 (2009): 1639–70.

Kravet T. and V. Muslu, 2010. Informativeness of Risk Disclosures in Corporate Annual Reports. Working paper, University of Texas-Dallas.

Lawrence, A.N. 2011. Individual investors and financial disclosure. Working paper, University of Toronto.

Lee, Y.J., 2010, The Effect of Quarterly Report Readability on Information Efficiency of Stock Prices, Working paper, National Taiwan University.

Lee, C.M., and B. Swaminathan. "Price Momentum and Trading Volume." *Journal of Finance* 55 (2000): 2017-2070.

Lehavy, R., F. Li, and K. Merkley. 2011. The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts. *The Accounting Review*, May.

Leuz, C., and C. Schrand, 2009. Disclosure and the Cost of Capital: Evidence from Firms' Responses to the Enron Shock. Working paper, University of Chicago.

Levine, C. and M. Smith, Critical Accounting Policy Disclosures (August 2006). Available at SSRN: http://ssrn.com/abstract=927486

Li, F. 2008. Annual Report Readability, Current Earnings, and Earnings Persistence. *Journal of Accounting and Economics* 45: 221–47.

Li, F, 2010a. The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach, *Journal of Accounting Research*. Vol. 48 No. 5 December:1049-102.

Li, F. 2010b. "Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?" Working paper, University of Michigan.

Li, F, 2011. Managers' Self-Serving Attribution Bias, Overconfidence, and Corporate Financial Policies. Working paper, University of Michigan.

Li, F., R. Lundholm, and M. Minnis. 2011. The impact of perceived competition on the profitability of investments and future stock returns. Working paper, University of Michigan.

Li, F., M. Minnis, V. Nagar, and M. Rajan. 2010. Formal and Real Authority in Organizations: An Empirical Assessment. Working paper, University of Michigan.

Li, E., and K. Ramesh. "Market Reaction Surrounding the Filing of Periodic SEC Reports." *The Accounting Review* 84(2009): 1171-1208.

Libby, R., and K. Rennekamp, 2010. Self-serving attribution bias, overconfidence and the initiation of voluntary disclosures. Working paper, Cornell University.

Loughran, T. and B. McDonald. 2010. Measuring readability in financial text. Working paper, The University of Notre Dame.

Loughran, T. and B. McDonald. 2011. When is liability not a liability? Textual analysis,

dictionaries, and 10-Ks. *Journal of Finance*.

Manning, C., and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Merkley, K., 2011, More Than Numbers: R&D-related Disclosure and Firm Performance, Working paper, University of Michigan.

Miller, B, 2010. The Effects of Reporting Complexity on Small and Large Investor Trading. *The Accounting Review*.

Miller, G.S., 2006. The press as a watchdog for accounting fraud. *Journal of Accounting Research*, 44(5): 1001-1033.

Mitchell, T. M. "The Discipline of Machine Learning." Working paper, Carnegie Mellon University CMU-ML-06-108, 2006.

Muslu, V., S. Rahhakrishnan, K. R. Subramanyam, and D. Lim. 2010. "Forward-Looking Disclosures in the MD&A and the Financial Information Environment" Working paper, University of Texas at Dallas and University of Southern California.

Nelson, K. K., and A. C. Pritchard. 2007. Litigation Risk and Voluntary Disclosure: The Use of Meaningful Cautionary Language. *SSRN eLibrary*.

Peterson, K. 2008. Accounting Complexity and Misreporting: Manipulation or Mistake, Working paper. University of Oregon.

Rogers, R. K., and J. Grant. "Content Analysis of Information Cited in Reports of Sell-Side Financial Analysts." *Journal of Financial Statement Analysis* 3 (1997): 17–30.

SEC. 2003. Commission Guidance Regarding Management's Discussion and Analysis of Financial Condition and Results of Operations; Interpretation; Rule.

Simon, H. 1997. *Administrative Behavior*. 4th Ed. The Free Press. New York, NY.

Skinner, D. 1994. "Why Firms Voluntarily Disclose Bad News." *Journal of Accounting Research* 32: 38-60.

Sloan, R. 1996. Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings? *Accounting Review* 71: 289–315.

Tavcar, L. R. "Make the MD&A More Readable." *CPA Journal Online* (1998). http:// nysscpa.org/cpajournal/1998/0198/newsViews/0198NV7.htm (1.12.2005).

Soltes, E. F. 2009. News Dissemination and the Impact of the Business Press. Ph.D. diss., The University of Chicago Graduate School of Business.

Verrecchia, R. E. 1983. Discretionary disclosure. *Journal of Accounting and Economics* 5, 365-380.

You, H. and X. Zhang. 2009. Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting Studies*. 14 (4) 559-586.

Table 1: Sample Papers on the Implications of Corporate Disclosures

	Independent variable		
Dependent Variable	Tone	Transparency / readability	Other
Future earnings (or earnings quality)	Li [2010a]	Li [2008]	
Market pricing	Davis, Piger, and Sedor [2008] Henry [2008] Li [2010a] Feldman, Govindaraj, Livnat, and Segal [2010]	Lee [2010]	You and Zhang [2009] Li [2010b] Brown and Tucker [2011] Li, Lundholm, and Minnis [2011]
Analyst behavior	Kothari, Li, and Short [2009]	Lehavy, Li, and Merkley [2011]	Kravet and Muslu [2010] Brown and Tucker [2011]
Information environment	Kothari, Li, and Short [2009]		Muslu, Rahhakrishnan, Subramanyam, and Lim [2010]
Cost of capital	Kothari, Li, and Short [2009]		
Litigation	Rogers, Van Buskirk, and Zechman [2010]	Nelson and Pritchard [2007]	
Other		Li [2011]	Li, Minnis, Nagar, and Rajan [2010]

ENDNOTES

[1] Unstructured information here refers to information that either does not have a pre-defined data model and/or does not fit well into relational tables. In this sense, textual information is likely to be "unstructured" even though it may be physically organized. For example, the Notes to the Financial Statements in corporate filings tend to be well organized and numbered. However, from the perspective of information processing, these notes are still unstructured in the sense that they are of high-dimension in nature and need to be further processed for empirical analysis. Unstructured textual information typically has irregularities and ambiguities that make it difficult to understand using traditional computer programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents.

[2] The Diction software, the General Inquirer, and the Linguistic Inquiry and Word Count (LIWC) software are often used in content analysis. The Diction is distributed by Professors Roderick P. Hart and Craig Carroll, professors of Communication at the University of Texas (Austin) and Lipscomb University, respectively. The General Inquirer is published by Harvard psychologist Philip J. Stone. The Linguistic Inquiry and Word Count (LIWC) software is created by University of Texas (Austin) psychologist James W. Pennebaker.

[3] The few exceptions are Henry and Leone [2010] and Lougrhan and McDonald [2011].

[4] The online version of the General Inquirer dictionary can be accessed at http://www.webuse.umd.edu:9090/

[5] The selection bias in the publication process is perhaps the reason that we do not see more of such papers.

[6] For example, Choudhary, Merkley, and Schloetzer [2010] find that the timeliness of 10-K filings is associated with market measures of firm information uncertainty suggesting that there is cross-sectional variation in the information content of 10-K filings.

[7] There is no consensus on the definition of earnings quality. I define earnings quality as the closeness of reported earnings to the "permanent earnings" following Dechow and Schrand [2004], who use earnings persistence to operationalize this concept.