JOURNAL OF INFORMATION SYSTEMS
Vol. 33, No. 1
Spring 2019
pp. 145–180

American Accounting Association
DOI: 10.2308/isys-51979

Text Data Sources in Archival Accounting Research: Insights and Strategies for Accounting Systems' Scholars

Michael Chuancai Zhang
Dan N. Stone
Hong Xie
University of Kentucky

ABSTRACT: This paper reviews the emerging computer-aided text analysis (CATA) accounting literature through proposing a model of the corpus linguistic research production process, followed by analysis of the main text archival data sources in published papers in the Top Six accounting journals from 2010 to 2016. Reviewed papers appear in a 5 × 5 matrix that includes five categories of text data (i.e., SEC filings, conference call transcripts, earnings press releases, financial analyst reports, and other sources) and five categories of text measures (i.e., tone, readability, similarity, firm characteristics and environment, and other measures). A brief review of the CATA literature published in two AIS journals is followed by a summary of the tools and KS (knowledge and skills) observed in the reviewed research. Finally, we offer implications by discussing four issues related to CATA accounting research. We conclude that the emerging CATA accounting research offers unique opportunities for knowledgeable AIS scholars.

Keywords: text data sources; corpus linguistics; accounting research; accounting systems.

I. INTRODUCTION, PURPOSE, AND SCOPE

Do the words, sentences, and paragraphs in accounting reports have information value? Historically, scholars analyzed the numbers in accounting and auditing reports. A growing literature explores the information value of the text "data" in accounting reports and other market-relevant data sources (e.g., conference calls). Advances in technology and data availability create possibilities for examining the text content of accounting-related reports using computer-aided text analysis (CATA). This paper reviews the 46 CATA accounting papers published in six accounting journals (*Journal of Accounting Research* [*JAR*]; *Journal of Accounting & Economics* [*JAE*]; *The Accounting Review* [*TAR*]; *Contemporary Accounting Research* [*CAR*]; *Accounting, Organizations and Society* [*AOS*]; and *Review of Accounting Studies* [*RAS*]) from 2010 to 2016. In supplemental analysis, it summarizes the 13 CATA papers that are published in two accounting information systems (AIS) journals (*Journal of Information Systems* [*JIS*] and *Journal of Emerging Technologies in Accounting* [*JETA*]) and compares the AIS research to that published in the traditional "top" accounting journals over comparable periods (2010–2016).[1] This paper seeks to inform AIS scholars of the state of CATA accounting research, its primary data sources and methods, and some possibilities for research foci and strategies for AIS scholars to contribute to CATA accounting research. This effort reflects an assumption that many AIS scholars have knowledge and skills (KS) that can be adapted and extended to CATA to enhance the quality and validity of CATA accounting research.

We define text data as systematically collected material consisting of written, printed, or electronically published words, typically either purposefully written or transcribed from speech (Badie, Berg-Schlosser, and Morlino 2011, 525). We include

The authors thank Pinky Ruesli and Russell Williamson for conversations about and advice on this topic and paper; Jacob M. and Anna Rose and the reviewers for comments and suggestions on previous drafts; and Mary B. Curtis (University of North Texas) for suggesting relevant papers and data sources. The authors also thank the University of Kentucky, the Gatton College of Business, and the Von Allmen School of Accountancy for financial support.

This paper summarizes a large, diverse body of sophisticated methods and practices. All errors are our own.

Editor's note: Accepted by Jacob M. Rose.

Submitted: December 2016
Accepted: November 2017
Published Online: November 2017

[1] Despite a large literature investigating academic accounting journals, little evidence exists (beyond surveys) of research quality differences across accounting journals. Nevertheless, we refer to the six accounting journals that surveys often rank highest as "Top Six."

spoken word data (e.g., earnings conference call audio analysis) as text data (Hobson, Mayew, and Venkatachalam 2012; Mayew and Venkatachalam 2012) while noting that spoken word data include added information and complexity (i.e., vocal cues). Given our interest in promoting scholars' *direct* engagement with and analysis of text, we omit research that reports only dichotomous counts, i.e., of the presence or absence of text but does not directly analyze text data. Hence, while interesting, we omit Prokofieva (2015), who dichotomously reports on whether firms tweet, and Du and Jiang (2015), who report the presence or absence of firms on seven social media sources. However, we include Fisher and Nehmer (2016), who conduct a "design science" analysis that maps the text of FASB and IFRS standards related to inventory to XBRL tags. We define corpus CATA accounting as the systematic study of text data that are relevant to accounting-related reports and practices.

Our choice of focus, journals, and review period derive from existing reviews of portions of the CATA accounting and related literature, including Fisher, Garnsey, Goel, and Tam (2010), Li (2010a), and Loughran and McDonald (2016; hereafter, LMcD).[2] Fisher et al. (2010) conduct an exemplary review of the text analytics literature in accounting, which considers data extraction and content analysis methods across 163 accounting and business sources (academic and professional). Of the 35 papers that are cited in the "lessons learned" portion of the review, 17 are working papers and 18 are published papers appearing between 1994 and 2009. Herein, we review research published in the Top Six journals during 2010–2016, i.e., after the Fisher et al. (2010) review. We contribute by reviewing, in depth, the published literature in the Top Six accounting journals.

Li (2010a) reviews the text analysis literature (n = 22 papers) investigating corporate disclosures (i.e., financial statement and SEC filings). Many of the papers reviewed in Li (2010a) were, at the time of the review, working papers that were later published. Li (2010a) summarizes the methods (circa 2010) for text analysis, research results related to the information content and transparency (readability) of corporate text disclosures, and their implications for earnings quality, market efficiency, financial policies, and litigation. The present review includes 19 papers, now published, that overlap with Li's (2010a) review, 16 of which were working papers (i.e., unpublished) when Li (2010a) was published.

LMcD (2016) conduct a review with a "selective focus" on text analytics research methods in finance and accounting. They discuss some common text analytics methods, highlight challenges in applying these methods in accounting and financial research, and provide suggestions for future research. To summarize, existing reviews of relevant literature focus on "lessons learned" through 2010 (Fisher et al. 2010), on corporate text disclosure research through 2010 (Li 2010a), and on "selected" aspects of text analytic research methods (LMcD 2016).

The scope of the present review is unique in its focus on recent research *published* in the *top-ranked accounting journals* beginning in 2010. Now that a sufficient quantity of literature appears in top accounting journals, we propose consideration of "what is uniquely ours" (Kinney 2001)—to accounting scholars and scholarship—in the emerging CATA accounting research literature. For example, which research projects, some of which were included in previous reviews as working papers, met the acceptance hurdle for publication in the Top Six accounting research journals? What are emerging trends and techniques in Top Six journal research in this focal area? In addition, we explore more general issues related to the "crisis of validity" in science (e.g., the use of p-values and related issues of validity [Dyckman 2016; Stone 2018]). Consideration of these issues is fundamental to building a credible CATA accounting and AIS literature.

This review also offers insights for AIS scholars regarding "what *should* be uniquely ours" in CATA accounting research. The rapid evolution of CATA technologies (Chen and Wojcik 2016; Mayer-Schönberger and Cukier 2013) requires ongoing learning from scholars to produce quality research. Many AIS scholars have expertise and knowledge in data management, harvesting, cleaning, mining, and analysis, and in internal control principles and systems develop practices in AIS. As will be argued herein, the emerging CATA accounting literature needs these skills to build and enhance its scientific credibility. Hence, unique opportunities exist for AIS scholars to partner with non-AIS accounting scholars (e.g., those with deep financial, taxation, or managerial accounting expertise) and scholars with unique technology knowledge (e.g., of specialized machine-learning software) to produce high-quality CATA accounting scholarship. Such scholarship will advance our knowledge of the role of language and communication in accounting practices, institutions, and reporting.

Hence, the principal focus of this review, i.e., on the text data sources used in published accounting research, differs from that of previous research. We also focus on AIS scholars' potential contributions to CATA accounting scholarship. Unique opportunities and hazards exist in available text data sources. One goal of our review is to highlight unrecognized opportunities in the sources, i.e., to paraphrase William R. Kinney, to consider what "big potatoes" remain and what "potatoes" are harvested in existing research. To extend Kinney's metaphor, there may also be deceptively attractive but "blighted potatoes" in the research field, which will waste scholars' time on unproductive tools or projects (e.g., data, software tools, answered research

[2] Because its primary focus is finance, we omit Kearney and Liu's (2014) review of sentiment in text analytics research. In it, they consider published research and working papers between 2006 and 2013. Kearney and Liu's (2014) review includes 28 papers, which consist of 12 working papers, and nine, four, and three papers published in finance, accounting, and other business disciplines, respectively.

questions). Hence, we seek to, metaphorically, identify strategies and resources that point AIS scholars toward the remaining "big golden potatoes"—which maximize the contributions of AIS scholars to this emerging literature. Accordingly, we offer signposts for AIS scholars to direct them toward "golden," and away from superficially attractive but "blighted," big potatoes.

We considered but decided against writing a detailed literature review of CATA research published in the *Journal of Information Systems* (*JIS*). There are many recent, high-quality reviews of the *JIS* literature, including a text mining examination of key research themes (Guan, Levitan, and Goyal 2018), an analysis of *JIS* research topics and methods (Myers, Snow, Summers, and Wood 2016), and several other relevant and important review themes and topics (e.g., Schmidt, Wood, and Grabski 2016; Moffitt, Richardson, Snow, Weisner, and Wood 2016; Janvrin and Wood 2016). The AIS literature is exemplary in the frequency and quality of published literature reviews (e.g., see, Neely and Cook 2011; Masli, Richardson, Sanchez, and Smith 2011; Kauffman, Lee, Prosch, and Steinbart 2011; Lee, Welker, and Odom 2009; Hutchison, Lee, and White 2004). Hence, the present paper is not a detailed review of the CATA AIS literature. It does however include a summary assessment of CATA research published in *JIS* and *JETA* between 2010 and 2016. In addition, we document and synthesize the research published in the top accounting journals beginning in 2010, and compare it to research published in a similar period in two top AIS journals.

Our analysis includes considering the unique and compelling attributes of text data sources, with a goal of directing the scarce and valuable attention of the community of AIS scholars. The promise of this endeavor is echoed in the following observations about text data:

> The steps in this process are not dissimilar to those found in any empirical research design, but the systematic use of textual data may present special challenges in the stages related to identifying the data to be collected and preparing the data for analysis. In the modern era of near-ubiquitous electronic textual content, this problem stems not from a lack of availability of textual data or tools for working with them but, rather, from the opposite: The staggering quantity and variety of textual data created real challenges in selecting what texts to analyze. Despite such challenges, however, textual data remain one of the most promising and one of the least explored sources of systematic information about the political and social world. (Badie et al. 2011, 526)

Given the unique challenges of text data, the guiding maxim of our paper is to "find and follow the analysis of text data" in accounting research. This focus on text data is motivated by multiple considerations. First, text data sources are important determinants of research themes, variables, and the application scope of CATA (Fisher et al. 2010). Accordingly, understanding text data sources can enrich future research themes in accounting by highlighting existing sources, variables, themes, and scope. Second, published research from text data sources have implications for managers and accounting practice. For example, research demonstrates that CATA can help firms better communicate with investors (e.g., Lee 2016; Tan, Wang, and Zhou 2015; De Franco, Hope, Vyas, and Zhou 2015; Hales, Kuang, and Venkataraman 2011; Tan, Wang, and Zhou 2014). As a result, CATA research may improve the quality of information provided to markets. In addition, it can inform standard-setters' efforts to guide and structure firms' text disclosures (e.g., Lundholm, Rogo, and Zhang 2014). However, to date, no comprehensive review summarizes the text data sources, and papers, in publications in top accounting journals. This review, by also considering published CATA AIS research, should also direct AIS researchers' efforts to contribute to the CATA accounting literature.

We focus on archival accounting research data sources, since this is the largest and fastest growing body of corpus linguistic scholarship in accounting; accordingly, we restrict the sample to archival research, since the data sources of experimental, interview, and survey research are human participants, not archival text sources. While relevant and important to this emerging body of literature, the experimental, interview, and survey research is deserving of its own review. Herein, our primary focus is on corpus linguistic (i.e., archival) data.

Our intent is primarily description (theory building) and not prediction (theory testing); hence, we pose research questions rather than hypotheses:

RQ1: Which text data source papers appear in published research in top accounting journals between 2010 and 2016?

RQ2: How often do top accounting journals publish text analysis papers?

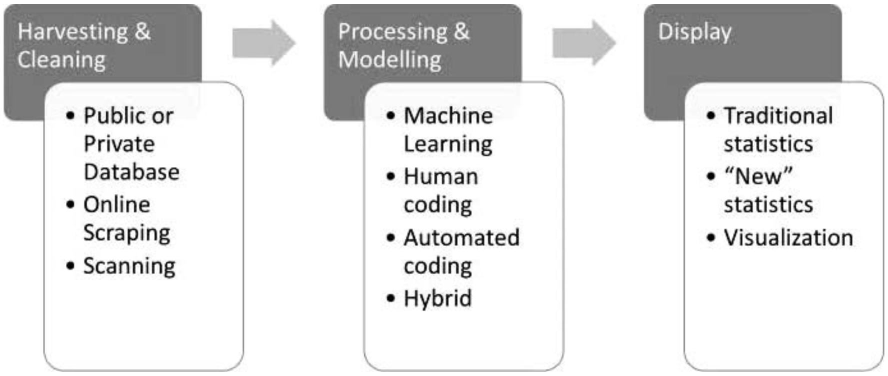
RQ3: What available text data sources, constructs, and variables are represented, and expected but unrepresented, in this literature?

RQ4: Are text data source samples "Big Data"? What are the implications of these sample sizes for managing text data?

RQ5: What are the main topics and research questions of published CATA accounting research?

RQ6: What insights from the existing literature can guide AIS scholars in producing accounting-relevant, high-quality CATA research?

FIGURE 1
Model of the CATA Research Process



Source: Adapted from Gartner Reports (2015).

This paper proceeds as follows. Section II introduces a model of the CATA research process and presents the selection criteria, sampling, and analysis methods. Section III presents the results of the review, which include the 5 × 5 matrix that structures and classifies the literature and the primary text variables analyzed in the papers. It also presents a summary review of the CATA AIS research (published in *JIS* and *JETA*). Section IV contains limitations of the present endeavor and discusses implications and directions for AIS-related CATA research. Section V concludes the paper.

II. THEORY AND METHOD

A "Theory" of CATA Research Production

A starting point for analyzing CATA research is to model the process of linguistic corpus creation, analysis, and reporting. The proposed model, illustrated in Figure 1 (partially adapted from Gartner Reports [2015]), proposes a three-stage process that is, at an abstract level, a core concept in AIS; i.e., input, processing, and output. We omit the critical, pre-text production phase of CATA accounting research, which includes a combination of either theorizing (in theory testing research) or speculating about (in theory generating research) the relationship between actors and variables in the analyzed text corpus. We are agnostic as to the desirability of theory generating (i.e., data mining) versus hypothesis testing in CATA research. In the reviewed literature, important contributions are evident in both hypothesis testing (e.g., Campbell, Chen, Dhaliwal, Lu, and Steele 2014; Davis, Piger, and Sedor 2012; Huang, Zang, and Zheng 2014a; Li, Minnis, Nagar, and Rajan 2014) and data mining (e.g., Guan et al. 2018; Larcker and Zakolyukina 2012; Liu and Moffitt 2016; Purda and Skillicorn 2015; Shirata and Sakagami 2008; Shirata, Takeuchi, Ogino, and Watanabe 2011; see, also, Debreceny and Gray 2011) research. In addition, we observe both hypothesis testing and data mining research in our sample of CATA accounting research published in the Top Six accounting journals.

Linguistic corpus research production begins with the harvesting and cleaning of text data, a formidable process that can require sophisticated knowledge of technology, extensive use of trained labor (others' or one's own), or both. The fundamental problem of text data is their messiness, i.e., their complexity and absence of structure. For example, imagine a goal of streaming the present manuscript's text into a structured database. Doing so would require that which is required in text analytic research projects, i.e., the creation of a data structure that contextualizes the data, and creates the possibility of unpacking the desired meaning (out of many possible meanings) that the data contain, in unstructured form. To date, many of the tools available for creating such structure require programming knowledge and unique hardware resources (e.g., Hadoop and a server; Gartner Reports [2016]). Hence, the extent of the supply of knowledge of these technologies has been a barrier to CATA research in accounting. More recently, accessible tools (e.g., RStudio, NVivo 11, LIWC2015, and WordStat in Stata) automate and facilitate aspects of data harvesting, cleaning, and analyses. Nevertheless, the challenges to generating a clean, structured data set from streamed "Big Data" text files are daunting. As with aspects of AIS data (e.g., text fields), the challenges include ensuring that the desired data are valid, complete, consistent, private (when needed), protected, and accessible in analyzable form.

Corpus linguistic methods link to conceptual frameworks, i.e., to the use of theory testing versus theory generating (e.g., data mining) approaches. Theory testing approaches may operationalize conceptual constructs using, for example, word lists

TABLE 1

CATA Research Sample
Selection Process

Summary of Sample Selection Process		No. of Papers
Initial Text Analysis Paper Search Results		58
Less:		
Experimental research:	Elliott, Rennekamp, and White (2015); Tan et al. (2015); Tan et al. (2014); Rennekamp (2012); Hales et al. (2011).	(5)
Literature review:	Li (2010a); Loughran and McDonald (2016); Miller and Skinner (2015).	(3)
Field (qualitative) study:	Griffith, Hammersley, and Kadous (2015); Scott and Orlikowski (2012).	(2)
Archival (no text analysis):	Drake, Roulstone, and Thornock (2012).	(1)
Discussion:	Dyer, Lang, and Stice-Lawrence (2016).	(1)
Final Sample:		46

Sources: Search of Top Six accounting journals (*AOS, CAR, JAR, JAE, RAS,* and *TAR*) 2010–2016.

that capture the unique language of financial tone, uncertainty, or the threat of litigation (e.g., LMcD 2011). Alternatively, data mining approaches—which are often atheoretical—may reveal unexpected relationships, e.g., to detect fraud (Larcker and Zakolyukina 2012). Machine-learning approaches often combine human labor with machine learning, for example, by having humans code a test sample of cases, which are then used to train (e.g., in a Naïve Bayesian model) a machine algorithm. The classifications of the resulting model are then tested using an out-of-bag (OOB) sample (e.g., Huang et al. 2014a). Emerging possibilities for analyzing and reporting promise more nuanced representations of the complexities of the Big Data samples that are often present in corpus linguistic research.

We next consider the main sample construction and harvest.

Method: Sample Selection and Text Data Sources

Selection Criteria and Sample of Papers

Our search strategy included keyword and snowball (i.e., in the present context, using identified papers to find additional relevant papers; Huck [2012]) sampling procedures. Most papers were obtained using searches on 24 keywords[3] in the databases for six leading accounting journals (*JAR, JAE, TAR, CAR, AOS,* and *RAS*) from 2010 to 2016. Reading the harvested papers yielded, via snowball sampling, additional papers. At least two authors read each paper identified in the initial search processes and agreed on its inclusion or exclusion based on the definition of text data (see the second paragraph of this paper). Table 1 summarizes the sample selection and winnowing process. The initial sample yielded 58 Top Six journal CATA accounting papers. Excluding false positives in the sample (i.e., 12 papers that consists of five experiments, three literature reviews, two field studies, one archival paper with no text data, and one discussion of a published paper) yielded 46 archival CATA accounting papers.

Coding, Analysis, and Additional Data Sources

Following best practices in the manual content analysis literature (Saldana 2016), we implemented a multi-phase content coding process, which included exploratory and confirmatory phases. The development of the coding categories, and the classification of papers within these categories, co-evolved in the exploratory phase. Exploratory analysis began with one author reading and coding each paper as to data source and primary text constructs and variables. After this, a concurring author independently repeated this process, with periodic meetings with the first coder to refine the taxonomy and the position of each paper within it. In the confirmatory phase of this process, the two coding authors reviewed and agreed upon the coding categories and the locations of all papers. Finally, we repeated this process with a third author. Coding of the primary text constructs and variables required similar, multiple coding iterations, with discussion, to define appropriate conceptual

[3] Key word search terms were generated by initial review of the accounting and text analytics research literature. The final search terms, which were applied to the article "title" and "abstract" fields, were sentiment, textual, tone, language, readability, linguistic, MD&A, content analysis, press release, conference call, social media, Twitter, Facebook, Instagram, LinkedIn, YouTube, Google, Google+, Pinterest, Ryze, Talkbiznow, Affluence, Quora, and press.

categories and their related operational measures. Because the taxonomy for the coding co-evolved, across coding rounds, with the coding of the paper, we do not report standardized coding metrics.

Across the multiple rounds of coding, each paper was categorized in the coding process into a cell in a 5 (data sources) × 5 (text variables) matrix. Analysis included identifying the authors, journal, sample period, research questions, and major results for each paper. To assess the relative frequency of text analysis papers published in journals, we obtained data on the total number of referred published papers in each of the top accounting journals from Summers and Wood (2016).[4]

III. RESULTS

Words as Indicators of Research Foci

The 46 included papers contain a nested sample of 469,634 (23,284 unique) words. Using stemmed word forms, i.e., a way to "normalize" words that share a common root (Baker and Egbert 2016) yields 17,866 unique words. The words used in natural languages (speech or text) follow Zipf's Law, i.e., the frequency of a word in a text source is inversely proportional to its statistical rank (Wolfram MathWorld; see http://mathworld.wolfram.com/ZipfsLaw.html) for about the most common 1,000 words in a document. Hence, the most frequent word will occur about twice as often as the second most frequent word, three times as often as the third most frequent word, etc. Figure 2 illustrates Zipf's Law using the distribution of words in the sampled Top Six journal CATA papers. One implication of Zipf's Law is the relative importance of more frequent (and the unimportance of infrequent) words in documents as indicators of deeper constructs and meanings.

One sampled paper (i.e., Suddaby, Saxton, and Gunz 2015) uses data visualization (e.g., Dilla, Janvrin, and Raschke 2010; Perdana, Robb, and Rohde 2015); given that data visualization methods facilitate presenting CATA results (Cao 2016), we expected more use in the sample. To illustrate one common approach to CATA data visualization, Figure 3 is a word cloud (created using NVivo 11) of the most frequent 100 words in our sample papers. In the figure, larger and more prominently displayed words occur more often in the sampled texts (see note to Figure 3 for generating parameters). A review of the most frequent words in the reviewed literature suggests a strong financial accounting and reporting focus within the reviewed research (i.e., disclosure*, earnings, accounts, reports, using, informs, managing, firm*). Accordingly, we infer that this literature's primary focus is on financial accounting topics with a comparative neglect of managerial (zero papers), auditing (zero papers), taxation (one paper; Law and Mills [2015]), and AIS (zero papers) topics. One inference that might be made from this result is that top accounting journals are uninterested in nonfinancial accounting research, or that financial accounting topics are the "low hanging fruit" of the CATA accounting literature; another might be that nonfinancial accounting authors do not submit corpus linguistic research to top accounting journals. Either inference suggests a monolithic focus—on financial accounting topics—in published Top Six accounting journal CATA research.

Top Six Publications—Distribution of CATA Accounting Papers by Journal and Year

Table 2 presents the distribution of the text analytic accounting papers by journal and year. Archival text analytic papers are about 2.4 percent of the peer-reviewed papers published in these journals. Over the sampled period, a mixed linear model with time (continuous variable) and journal (categorical variable) as predictors, and the percentage of published archival text analytic papers as the dependent measure, indicates no effect of journal ($F(5, 35) = 1.92$; $p = 0.33$) and no effect for year ($F(1, 35) = 2.130$; $p = 0.153$). The most text analytic papers are published in 2015 (11 papers), followed by 2016 (nine papers).[5] *TAR* has published the most such papers (15), while *JAR* has published the most papers as a percentage of total papers published (3.8 percent).

Are Text Analytic Samples "Big Data"?

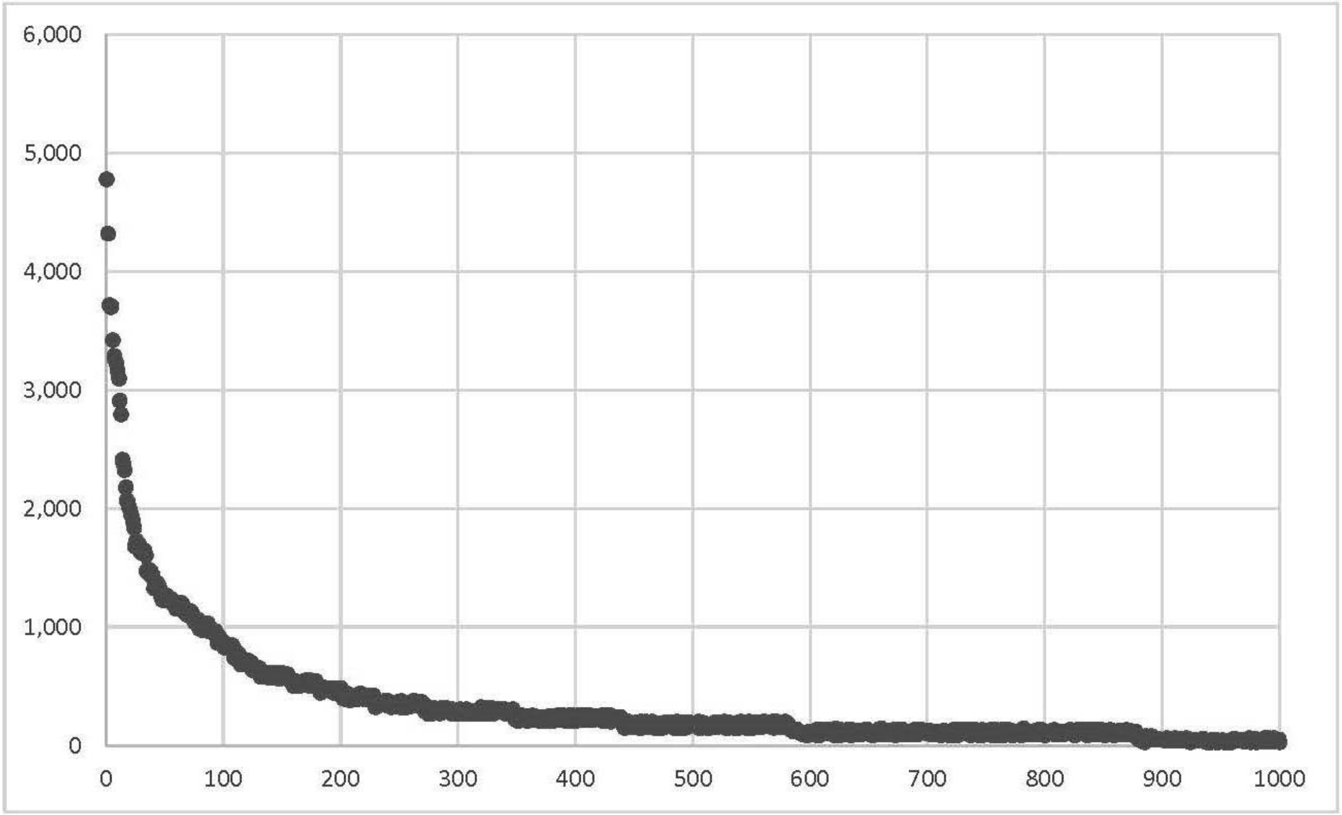
An operational, definitional question regarding CATA accounting samples, which is important in identifying appropriate tools and methods for their analysis, is whether such samples are "Big Data" (Guan et al. 2018; Moffitt and Vasarhelyi 2013; Vasarhelyi, Kogan, and Tuttle 2015).[6] One important realization regarding text data is that sample size depends on the level of analysis that one chooses. For example, consider Suddaby et al. (2015): At a firm-level of analysis, the sample size = 4, i.e., the Big 4 accounting firms. Both U.S. and Canadian Big 4 firms are included; hence, at another level of analysis, the sample is 8 (i.e., 2 × 4—U.S. and Canadian Big 4 accounting firms). But the authors collect about 7,000 Twitter tweets, 590 Facebook

[4] We thank Professor David A. Wood (Brigham Young University) for providing these data.

[5] Our searches also revealed forthcoming, prepublication CATA papers in these journals.

[6] Gartner Research defines Big Data as: "high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. See, also, https://en.wikipedia.org/wiki/Big_data

FIGURE 2
Illustration of Zipf's Law
Distribution of 1,000 Most Frequent Words in Reviewed Top Six Accounting Literature

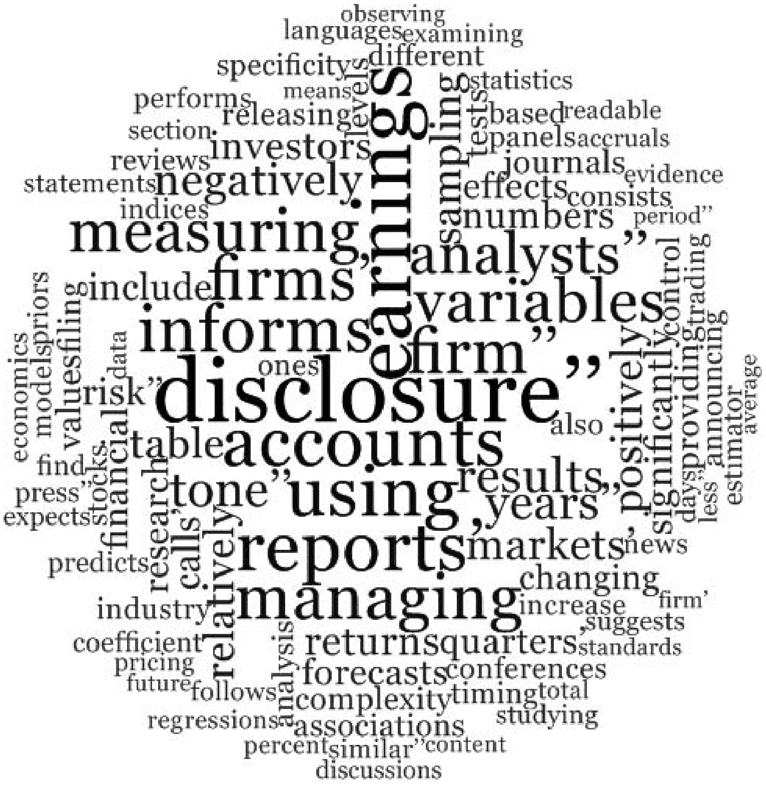


pages, and 33,561 firm webpages. Further, the Twitter messages contain 115,927 words, the Facebook pages 9,418 words, and the sampled webpages about 9,000,000 words, respectively. Accordingly, at the firm-level analysis Suddaby et al. (2015) is not Big Data; at the tweet/webpage level of analysis it approaches Big Data; and at the word level, it is Big Data. Although not as extreme, similar level-of-analysis observations could be made about most of the papers in our sample, although some (e.g., Cho, Roberts, and Patten 2010) have small sample sizes (n = 190 firms). Further, as several authors note (Fisher et al. 2010; LMcD 2016), the management of unique text "big" data sets requires KS in tools for managing large data sets (e.g., Hadoop, AWS [Amazon Web Services]).

Data size, i.e., big or "little," is an important determinant of the feasibility of analysis method. For example, (labor intensive) manual content analyses is restricted to small samples (García Osma and Guillamón-Saorín 2011). In contrast, CATA, e.g., automated content analysis, is less accurate but more scalable and affords greater statistical power (Fisher et al. 2010). In addition, the growing sample sizes in social science research include the risk that scholars will report pragmatically trivial results, with ill-defined constructs, but claim *statistically* significant results (Combs 2010).

Within the present sample, we observe a shift, relative to the literature reviewed in Fisher et al. (2010), toward more sophisticated, complex, automated content analysis methods and models (e.g., Naïve Bayes, VSM—vector space model, Random Forest model, SVR—support vector regression) and away from traditional manual content analysis methods. Within the sample, however, many papers retain a mixed, partially automated content analysis method (e.g., Huang, Teoh, and Zhang 2014b) and some adopt mostly manual analysis methods (García Osma and Guillamón-Saorín 2011). An important concern regarding the (lack of) replicability of this literature, is that descriptions of content analysis methods—manual and automated—are often sparse, insufficient to enable replication, and inconsistent with best practices. Hence, reporting data handling methods, perhaps in online appendices, is important to advancing the replicability, validity, and credibility of this literature.

FIGURE 3
Word Cloud of Most Frequent Words in Top Six CATA Published Papers



Source: Created using NVivo 11.

The parameters for the word cloud were to "stem" (i.e., normalize) related words and omit the following (called "stop") words: "word*," "000," "two," and "2012."

TABLE 2

Distribution of Text Analysis in the Top Six Journal Papers by Journal and Year

	Top Six Accounting Journals														
	JAR		*JAE*		*TAR*		*AOS*		*CAR*		*RAS*		Total		
Year	Text	All	Text	All	Text	All	Text	All	Text	All	Text	All	Text	All	Percentage
2010	2	29	0	27	1	72	1	43	0	31	1	27	5	229	2.18%
2011	1	39	0	34	3	72	1	36	0	55	0	33	5	269	1.86%
2012	2	39	0	45	0	73	0	32	3	45	0	36	5	270	1.85%
2013	1	34	1	36	0	73	0	33	0	53	1	35	3	264	1.14%
2014	0	36	1	31	6	81	0	38	0	45	1	48	8	279	2.87%
2015	2	27	1	37	2	84	1	46	4	64	1	47	11	305	3.61%
2016	1	32	2	42	3	76	0	39	0	45	3	34	9	268	3.36%
Total	9	236	5	252	15	531	3	267	7	338	7	260	46	1,884	2.44%
Percentage	3.81%		1.98%		2.82%		1.12%		2.07%		2.69%		2.44%		

Text = text analysis papers; All = the peer-reviewed papers published in the specific journal and year.

Top Six Journal CATA Accounting Publications: Text Data Sources

Table 3 presents the text data sources of the reviewed papers. Perhaps because of its public (universal) access, the SEC's EDGAR database is the modal data source (present in 54.3 percent of papers), while conference call transcripts account for about 22 percent of the observed data. Earnings press releases and analyst reports account for another 15 percent of the data, while four remaining sources each account for 2.2 percent of the text data sources. Overall, 60.9 percent ($(25 + 1 + 1 + 1)/46$) of the papers use public (39.1 percent private) data sources. The frequency of using public versus private data sources does not differ in our sample (sign test, one-tailed, $p = 0.151$).

The Top Six Journal CATA Accounting Literature

Appendix A summarizes the authors, journal, text data, time, text variables, and research questions in the surveyed research. Tables 13 through 18 present the papers, organized by data source, as follows (in order from most to least observed): Table 13—data from SEC 10-K filings; Table 14—data from both 10-K and other SEC filings; Table 15—data from conference call transcripts (including audio files); Table 16—data from earnings press releases; Table 17—data from "other" text sources; and Table 18—data from analyst reports

Organizing Framework and the Top Six Journal CATA Accounting Modal Paper

The matrix organization of our review, shown in Tables 4 and 5, includes text data variables as columns and the specific data sources as rows. Table 4 presents and organizes the 28 papers that use SEC text data; Table 5 presents and organizes the 21 papers that use non-SEC text data. Because some papers use both SEC and non-SEC text data, the sum of papers in these two tables is more than 46. The columns in Table 4 identify five categories of text variables: (1) tone, (2) readability, (3) similarity, (4) firm characteristics and environment (such as risk and competitive environment), and (5) other measures, while the rows are the specific SEC documents analyzed in this literature (i.e., 10-K, 10-Q, 20-F, 8-K, and 6-K).[7]

The modal sampled paper with SEC data analyzes 10-K or 10-Q filings (23 out of 28 papers or 82.1 percent) (see Table 4). Diversity exists, however, in the text analysis conducted on 10-Ks and 10-Qs, with at least two papers including at least one of the variable categories in each column of Table 4. The modal category of text variable analyzed in SEC data is tone (in 11 of 28 papers). Table 5, which summarizes the non-SEC data source papers, includes four of the five text variables in Table 4 as columns (omitting firm characteristics and environment), and four rows of data sources, i.e., conference calls (audio or transcript), earnings press releases, analyst reports, and other sources. Greater diversity of both sources and measures is evident in the papers using non-SEC data sources. Regarding measures, most papers using non-SEC data sources analyze tone or "other measures," where the "other measures" are unique to individual papers. Regarding sources, most papers using non-SEC data analyze conference calls (ten out of 21 papers), but diversity exists in the measures chosen in the analysis of these conference calls. For the conference call data papers ($n = 10$), three papers examine tone, one examines similarity, and six examine "other measures" of text variables. Hence, the modal paper among these ten papers, and in all non-SEC data sources, analyzes a unique "other measure" of text.

The Columns—Text Variables in CATA Accounting Research

We next consider the columns in Tables 4 and 5, which present the analyzed linguistic variables (1) tone, (2) readability, (3) similarity, (4) firm characteristics and environment (Table 4 only), and (5) other measures.

Tone (sentiment or emotion) measures. Tone is the most analyzed construct in the research (in 21 papers).[8] The tone columns of Tables 4 and 5 include the 21 papers that analyze tone as a primary variable, and exclude those in which tone is presented in supplemental analyses (Hollander, Pronk, and Roelofsen 2010; Purda and Skillicorn 2015) or only the tone of risk disclosures is analyzed (Campbell et al. 2014; Kravet and Muslu 2013). Accessible resources, including software (e.g., DICTION, LIWC [Linguistic Inquiry and Word Count]) and financial word lists (e.g., LMcD 2011, 2014), facilitate research among scholars interested in studying tone. Accordingly, most of the reviewed papers use standardized word lists to identify positive and negative tone as the analysis variable. Exceptions to this standard approach include:

[7] Firms must file annual and quarterly reports with the SEC in Form 10-K and Form 10-Q, respectively. Foreign firms, if less than 50 percent of their voting shares are traded on the U.S. exchange, must file annual reports with the SEC in Form 20-F. Firms must file Form 8-K with the SEC to notify shareholders of certain important events such as (1) signing, amending, or terminating material definitive agreements not made in the ordinary course of business, (2) bankruptcies or receiverships, (3) material impairments, (4) change of auditors, (5) senior officer appointments and departures, (6) earnings announcement press releases, and (7) many other events. Foreign firms whose shares are traded on the U.S. exchange are required to file Form 6-K with the SEC to notify shareholders of certain important events.

[8] Tone is often called "sentiment" in the finance (e.g., see, Kearney and Liu 2014) and natural language processing literatures and "emotion" in linguistics and psychology research (e.g., see, Pennebaker 2004).

TABLE 3

Top Six Journal Papers
Text Data Sources Description

No. of Papers	Percentage	Data Category	Data Source	Source Link	Availability
25	54.3%	SEC Filings: 10-K; 10-Q; 8-K; 6-K; 20-F	EDGAR	https://www.sec.gov/edgar.shtml	Public Domain
10	21.7%	Conference Call Transcripts	Thomson Reuters StreetEvents	https://www.refinitiv.com/en/products/eikon-trading-software	Library Database
			Factiva	https://global.factiva.com	Library Database
			Seeking Alpha	http://seekingalpha.com/earnings/earnings-call-transcripts	Public Domain
			FactSet	http://www.factset.com	Library Database
4	8.7%	Earnings Press Releases	Factiva (Dow Jones Business News and Press Release Newswire)	https://global.factiva.com	Library Database
3	6.5%	Analyst Reports	Thomson ONE Analytics	https://www.thomsonone.com/DirectoryServices/2006-04-01/Web.Public/Login.aspx?brandname=www.thomsonone.com&version=3.7.9.18833&protocol=0	Library Database
			Thomson ONE Banker (Investext)	http://banker.thomsonone.com/	Library Database
1	2.2%	International Financial Annual Reports	Bureau van Dijk (Osiris-Global Reports Database)	https://www.bvdinfo.com/en-gb	Library Database
1	2.2%	Management Earnings Forecasts	Factiva (Dow Jones Business News and Press Release Newswire)	https://global.factiva.com	Library Database
1	2.2%	Social Media	Twitter, Facebook, Firms' Web Pages	Internet	Public Domain
1	2.2%	Annual Results Press Releases of Spanish Companies	CNMV	https://www.cnmv.es/portal/home.aspx	Public Domain
46	100%				

TABLE 4

Distribution of Text Analytic Papers by Data Sources and Variables
SEC Sources

		Text Variables				
Text Data Sources	**No.**	**Tone Word Lists/ Naïve Bayes**	**Readability Fog, Flesch-Kincaid, Flesch Reading Ease; Document Length**	**Similarity See Table 6**	**Firm Characteristics and Environment See Table 7**	**Other Measures See Table 8**
10-K	18	Cho et al. (2010); Law and Mills (2015); Mayew, Sethuraman, and Venkatachalam (2015); Merkley (2014); Frankel et al. (2016)[a]; Davis and Tama-Sweet (2012)[a]	Lehavy et al. (2011); Filzen and Peterson (2015); Merkley (2014); Miller (2010); Guay et al. (2016); Lawrence (2013)	Li, Lundholm, and Minnis (2013); Peterson et al. (2015); Brown and Knechel (2016); Brown and Tucker (2011)	Hope et al. (2016); Li et al. (2013); Kravet and Muslu (2013); Campbell et al. (2014)	Frankel et al. (2016); Cho et al. (2010); Merkley (2014)
10-K; 10-Q	3	Feldman, Govindaraj, Livnat, and Segal (2010); Li (2010b)			Purda and Skillicorn (2015)	
8-K	4	Henry and Leone (2016); Rogers, Van Buskirk, and Zechman (2011); Segal and Segal (2016)	Bozanic and Thevenot (2015)[a]	Bozanic and Thevenot (2015)[a]		
10-K; 20-F	1		Lundholm et al. (2014)[a]			Lundholm et al. (2014)[a]
10-Q	1		Lee (2012)			
8-K; 6-K	1		Lundholm et al. (2014)[a]			Lundholm et al. (2014)[a]
Total	28	11	9	5	5	4

[a] Some papers include multiple data sources. For example, Frankel et al. (2016) includes both 10-K and conference call transcripts (audio); Davis and Tama-Sweet (2012) include both 10-K and earnings press releases. Some papers include multiple text data sources and variables. Accordingly, the total number of papers listed by text data sources does not equal the total number of papers listed by text variables.

TABLE 5

Distribution of Text Analytic Papers by Data Sources and Variables Non-SEC Sources

		Text Variables			
Text Data Sources	No.	Tone Word Lists/Naïve Bayes	Readability Fog, Flesch-Kincaid, Flesch Reading Ease; Document Length	Similarity See Table 6	Other Measures See Table 9
Conference Calls (Audio or Transcript)	10	Davis, Ge, Matsumoto, and Zhang (2015); Allee and DeAngelis (2015); Larcker and Zakolyukina (2012)		Lee (2016)	Brochet et al. (2016); Hobson et al. (2012); Kimbrough and Louis (2011); Hollander et al. (2010); Li et al. (2014); Frankel et al. (2016)[a]
Earnings Press Releases	4	Davis et al. (2012); Huang et al. (2014b); Davis and Tama-Sweet (2012)[a]			Kimbrough and Louis (2011); Kimbrough and Wang (2014)
Analyst Reports	3	De Franco et al. (2015); De Franco et al. (2014)	Huang et al. (2014a); De Franco et al. (2015)		Bozanic and Thevenot (2015)[a]
Other Sources	4	Baginski, Demers, Wang, and Yu (2016); García Osma and Guillamón-Saorín (2011)	Lang and Stice-Lawrence (2015)	Lang and Stice-Lawrence (2015)	Lang and Stice-Lawrence (2015); Suddaby et al. (2015); García Osma and Guillamón-Saorín (2011)
Total	21	10	3	2	11

[a] Indicates that the paper includes multiple data sources. For example, Frankel et al. (2016) includes both 10-K and conference call transcripts (audio); Davis and Tama-Sweet (2012) include both 10-K and earnings press releases. As one paper may use one or more textual data sources and one or more textual variables, the total number of papers by textual data sources is not equal to that by textual variables. Readability measures are used in 32 percent (= 9/28) of papers that analyze SEC data sets (Table 4) and 14 percent (= 3/21) of papers that analyze non-SEC data sets.

1. Four papers that use a Naïve Bayes (Bayesian) machine-learning approach (De Franco, Vasvari, Vyas, and Wittenberg-Moerman 2014; Henry and Leone 2016; Huang et al. 2014b; Li 2010b) within which an artificial intelligence (AI) algorithm is trained using test coding in a pilot sample. Classifications are then tested in a hold-out (out-of-bag [OOB]) sample.
2. García Osma and Guillamón-Saorín (2011), following a previous investigation, i.e., Brennan, Guillamón-Saorín, and Pierce (2009), manually code disclosure tone as positive or negative.
3. Henry and Leone (2016) adopt multiple word lists and a Naïve Bayes approach to triangulate and validate these alternative approaches in their analysis of SEC 8-Ks (n = 64,357). They find low correlations among four-word list measures of tone ($0.295 \leq$ tone ≤ 0.522), which suggests that existing standardized word lists may not capture a common conceptual construct for tone in financial documents. However, the results indicate that all included tone measures predict abnormal returns on unexpected earnings. Results also indicate that simple word list measures of tone predict future earnings about equally as well as the complex and less replicable Naïve Bayes model. Henry and Leone (2016) also find no improvement in predictive validity in their sample when comparing (the more complex) inverse document frequency weighting (idf) of words, as advocated by LMcD (2011), in a comparison with (much simpler) unweighted analysis. Accordingly, the Henry and Leone (2016) analysis suggests that, at least in the context of SEC 8-K documents, simpler approaches (i.e., word lists versus Naïve Bayes and unweighted versus idf weighting) have equal predictive validity, but that sufficient differences exist in word-list-based predictions to warrant additional validation.

Readability measures. Readability is a complex and evolving construct that was designed to assess the grade levels of texts assigned to K–12 students. Perhaps because standard readability formulas are easily calculated, readability measures are common (in about 26 percent = 12/46 papers) in the studied literature. As shown in Table 4 and Table 5, respectively, readability measures are more frequently used in papers that analyze SEC (32 percent = 9/28) than non-SEC (14 percent = 3/21) data sets. The three traditional measures of readability, i.e., Fog, Flesch Reading Ease, and Flesch-Kincaid, are all simple linear combinations of sentence length and word complexity (LMcD 2016); consequently, they are highly correlated. LMcD (2014) find that the Fog index (and hence all traditional readability measures) is a poor measure of readability in financial documents, due to the presence of common "complex" financial words (e.g., financial, company, operations). These limitations motivate a need for revisiting the traditional metrics of readability in CATA accounting research. The alternative measure to the traditional three measures that is observed in the present literature is 10-K submission length as a proxy for readability (LMcD 2016). That even its developers (i.e., LMcD 2016) recognize its inherent confound, i.e., that longer submissions are most likely from complex companies, suggests the need for more valid, i.e., contextually sensitive, measures of readability. Indeed, Filzen and Peterson (2015) relabel LMcD's (2016) measure of "readability" (i.e., document length) as "complexity," further complicating the already problematic construct validity of this metric. That readability and complexity are related, but separate, constructs when applied to financial data, is suggested in the results presented in Guay, Samuels, and Taylor (2016), who report a correlation of 0.49 between these constructs.

That emerging metrics of readability, e.g., Coh-Metrix (McNamara, Graesser, McCarthy, and Cai 2014) and the "Bog" index (StyleWriter n.d.), which are more context sensitive and likely to capture the multiple dimensions of readability, have yet to appear in the accounting literature is both a surprise and an opportunity.[9] Hence, the literature on readability suggests low conceptual and empirical clarity and a need for rigor in defining, refining, and validating constructs. In addition, an important potential contribution is a rigorous analysis of readability constructs in accounting-relevant texts.

Similarity (comparability) measures. Conceptually, similarity is simple: how similar or distinct are two or more objects (e.g., documents)? Operationally, complications exist in the many variations (1) on the objects in this comparison (e.g., Is the comparison an object or a set of objects? What part(s) of the document sections are compared?), and (2) in the (devilish) details of how the comparison is modeled and implemented (e.g., over time, versus across firms, versus both). For example, in a simple comparison of words, called a "lexical" bag of words (BOW) comparison, "the company sued the SEC" is equivalent to "the SEC sued the company" since the words are identical. In a "semantic" comparison, however, the meaning of the phrases differs. "Lexical" comparisons are implementable in large data sets using relatively simple mathematical and statistical models. In contrast, semantic comparisons challenge existing natural language processing methods and are an important and evolving topic of natural language processing research.

These complexities may explain the infrequency in the reviewed literature of similarity metrics, although some software packages now include standardized document similarity metrics (e.g., NVivo 11). We observe seven published papers that use similarity metrics, five using SEC documents (10-Ks and 8-Ks) and two using other sources (see Tables 4 and 5). All use

[9] We note the introduction of a new measure of readability, the "Bog" index, in a paper published after our sample period (Bonsall, Leone, Miller, and Rennekamp 2017).

TABLE 6
Similarity Measures

Paper	Metric	Description
1. Bozanic and Thevenot (2015)	Cosine Similarity	8-K over time: current versus previous quarter press releases.
2. Brown and Knechel (2016, 737)	Vector Space Model (VSM) Similarity	10-K regression residual (predictor: document length) in audited versus unaudited disclosures in "auditor-industry-year reference group."
3. Brown and Tucker (2011)	Vector Space Model (VSM) Similarity	10-K MD&A over time: over time (current versus prior year).
4. Lang and Stice-Lawrence (2015)	Cosine Similarity	Annual reports (1) versus U.S. peer firms, and (2) versus non-U.S. peer firms.
5. Lee (2016)	Cosine Similarity	Function words in management's discussion versus Q&A portions of conference calls.
6. Li et al. (2013)	Cosine Similarity	Hoberg and Phillips's (2010) measure of 10-K product similarity (percent of common words).
7. Peterson et al. (2015)	Vector Space Model (VSM) Similarity	10-K accounting policy footnotes (1) over time: current versus prior-year disclosure, and (2) across firms: target versus within-industry firms.

lexical similarity comparisons, although a semantic approach to similarity is found in a recent issue of the *Journal of Information Systems* (Guan et al. 2018). Table 6 summarizes the similarity measures in the reviewed literature.

Three papers use a vector space model (VSM) implementation of lexical similarity, which was developed for web search algorithms (Salton, Wong, and Yang 1975). Two of these papers include an author with extensive, relevant programming knowledge and expertise (Professor Stephen V. Brown, i.e., see Brown and Knechel [2016] and Brown and Tucker [2011]). These papers examine differing aspects of 10-K disclosures, including year-to-year MD&A disclosure changes (Brown and Tucker 2011), document length in audited versus unaudited "auditor-industry-year reference groups" (Brown and Knechel 2016), and changes in accounting policy footnotes over time and across firms (i.e., in industry groups) (Peterson, Schmardebeck, and Wilks 2015).

Four papers use a similar although simpler approach, i.e., cosine similarity, although in some cases with adjustments for document length to improve data validity and interpretability (Lang and Stice-Lawrence 2015). Lang and Stice-Lawrence (2015) compare financial statement similarity in U.S. versus non-U.S. firms and among firms that must comply with standards. Bozanic and Thevenot (2015) compare quarter-to-quarter disclosure changes in 8-Ks and find that disclosure similarity reduces uncertainty about firm performance. Lee (2016) compares the lexical similarity of management discussion compared with Q&A portions of conference calls as an innovative measure of managers' lack of spontaneity (or "scripted-ness"). Results indicate a negative market reaction to more scripted conference calls. Finally, Li et al. (2013) adopt the 10-K product similarity measure from Hoberg and Phillips (2010) as a supplemental measure of industry competitiveness.

Given evolving and improving technologies and the importance of comparability (a form of similarity) to accounting measurement, future analysis promises improved results and implications from the next generation of CATA accounting research. Accordingly, investments in testing and validating similarity metrics are important contributions to understanding changes in accounting documents.

Firm characteristics and environment measures. The firm characteristics and environment variables are a mix of researcher-developed text variables, software-package-provided word lists, and a few cases of emerging modeling methods, all with a goal of capturing firm characteristics or environment (see Table 7 for a summary). Conceptual constructs in these papers include risk, competition, and the likelihood that management's claims are true. Most papers in this category adopt a lexical BOW method to measuring constructs. For example, Cho et al. (2010) use the DICTION 5.0 software to capture words in the constructs of "optimism" (a measure of tone) and "certainty" in their analysis of environmental disclosures by 190 firms for which both environmental and 10-K data were available. In more recent work, LMcD (2015) provide evidence that DICTION's optimism and pessimism words misclassify words in 10-K documents. Hence, scholars applying the DICTION dictionaries to financial documents might consider using them to triangulate the validity of other word list sources. The application of the DICTION dictionaries to financial documents has been tested in SEC filings. However, the usefulness of the DICTION dictionaries remains uninvestigated in the taxation, auditing, management accounting, and accounting control settings.[10]

[10] The DICTION word lists are identical in versions 5.0 and 7.0 (source: email message from DICTION support staff on December 13, 2016).

TABLE 7

Firm Characteristics and Environment Measures

Paper	Metric(s)	Description (10-K Data Source)
1. Campbell et al. (2014)	Risk Disclosures: financial risk, tax risk, litigation risk, other systematic risk, other idiosyncratic risk, systematic risk, idiosyncratic risk.	Risk-related words.
2. Hope et al. (2016)	Risk Specificity: the number of specific words or phrases conveying specific information relevant to the disclosing firm, divided by the number of total words in the risk factor disclosure section.	10-K: SVR machine modeling using words or phrases.
3. Kravet and Muslu (2013)	Risk Disclosure: the total number of sentences with at least one risk-related keyword. Δ*Risk Disclosure*$_{i,t}$ = the difference between *Risk Disclosure*$_{i,t}$ and *Risk Disclosure*$_{i,t-1}$.	Risk-related sentences that include at least one of these risk-related keywords (where * implies that suffixes are allowed): can/cannot, could, may, might, risk*, uncertain*, likely to, subject to, potential*, vary*/varies, depend*, expos*, fluctuat*, possibly*, susceptible, affect, influenc*, and hedg*.
4. Li et al. (2013)	Competition (management's perceptions of the intensity of the competition they face): net occurrences of competition words/total number words in the 10-K.	Competition words: competition, competitor, competitive, compete, competing.
5. Purda and Skillicorn (2015)	Probability of truth: first, create the word-frequency matrix; second, produce the rank-order of word's predictive ability; finally, use support vector machine (SVM) classification with the top 200 words most predictive of fraud in the MD&A as the input to measure how likely a report is to be truthful.	Top 200 words that are most predictive of fraud in the MD&A.

Three firm characteristic papers take differing models, foci, and levels of analysis to capture the content of 10-K risk-related disclosures. Hope, Hu, and Lu (2016) create an innovative measure called "specificity" to capture firms' risk-related 10-K disclosures. Specifically, they use the named entity recognition (NER) technique found in the Stanford NER tool to identify and extract entity names (e.g., Microsoft, China) in 10-K disclosures. They then demonstrate that the specificity measure correlates with market reactions. Campbell et al. (2014) use a list from previous research and an emerging method called "latent Dirichlet allocation," i.e., an emerging, computation-intensive, cluster analysis model (Chen and Wojcik 2016; Kosinski, Wang, Lakkaraju, and Leskovec 2016) to identify over 300 words as risk indicators in 10-K filings. Kravet and Muslu (2013) parse 10-K filings into sentences (using custom-developed UNIX Perl code), which are then tagged as risk related if they contain at least one risk-related word from a set of 18 "risk-related words" that were developed by "reading 100 randomly selected annual reports."

Hence, three papers published in top accounting journals differ in their approach to measuring risk in a common data set (10-K filings). Across these three papers, the number of risk words differs (from more than 300 to a low of about 18), as does the unit of analysis (word versus sentence) and the modeling approach (previous research list, latent Dirichlet allocation data mining, a human reading of 100 randomly selected reports, SVR modeling). This degree of variability in method, without validation efforts to reconcile the differences in methods and results, is a testament to both innovation and the lack of common scientific methods and practices in the present CATA accounting research. Obviously, opportunities for discovering the implications of and differences in methods and practices exist in this literature.

Li et al. (2013) measure the competitiveness of a firm's environment by counting the number of competition-related words (using a list of five words) in firms' 10-K filings. However, no validation of the word list is offered; no sensitivity analyses are performed on the listed words. Using a similar approach, although for a different construct, Merkley (2014) created a list of R&D-related words by reviewing 10-K filings, consulting industry experts, and adapting from word lists from earlier research. Merkley (2014, Appendix A) reports that the results are insensitive to narrower and broader versions of the list. Finally, Purda and Skillicorn (2015) use a "random forest method," i.e., a machine-learning model based in decision trees and ensemble learning, to identify words in 10-K MD&A filings that are predictive of fraud. Random forest model results are then input to a

TABLE 8

Other Measures
SEC Sources

Paper	Metric(s)	Description
1. Cho et al. (2010)	"Certainty" of environment disclosures (from DICTION).	Attributes of environmental disclosures.
2. Frankel et al. (2016)[a]	MD&A accruals and MD&A future cash flows obtained from the SVR approach.	10-K MD&A and conference calls: data mined single-word and two-word phrases that predict accruals.
3. Lundholm et al. (2014)	Number of numbers: the logarithm of total number of numbers in the MD&A section of the 20-F or 10-K, excluding dates.	20-F or 10-K: numbers in the MD&A section.
4. Merkley (2014)	Narrative R&D disclosure: the total number of R&D-related sentences in the 10-K filing.	R&D-related sentences.

[a] Use both SEC and non-SEC data sources.

support vector machine (SVM) analysis to judge the probability of a truthful 10-K filing. Hence, Purda and Skillicorn (2015) use multiple, emerging natural language methods in pursuit of a critical outcome, i.e., identifying fraud.

Other measures. Text "other measures" are an innovative assortment of mostly researcher-constructed measures with a few examples of emerging modeling methods. Descriptions of the specifics of these measures appear in Table 8 for the four papers that analyze SEC text data, and in Table 9 for the 12 papers that analyze non-SEC text data. One paper in this group (Frankel, Jennings, and Lee 2016) includes text analysis of both SEC and non-SEC (i.e., conference call) data using support vector regression (SVR), a data mining, statistical learning theory model. The SVR model is created using a training (i.e., data mining) sample that is then used to forecast (classify) outcomes in an out-of-box (OOB) sample (Chen and Wojcik 2016). Results indicate that SVR text mining of 10-K MD&A documents and conference call transcripts yield accurate predictions of next-period cash flows. The appendix in Frankel et al. (2016) provides a useful introduction to such models. However, an important remaining contribution is articulating the unique contributions and limitations of SVRs compared to alternative supervised and unsupervised machine-learning models.

Lundholm et al. (2014) create a measure at the edge of what can be considered "text data," i.e., of the "number of numbers" in firms' 10-K filings. They find that, with increasing geographic distance and reporting format differences, firms report more numbers in their 10-K filings. The meaning of this finding is unclear, and construct validity of this measure is suspect, however, since the conceptual construct that the "number of numbers" represents is vague. While perhaps related to firm complexity, or to firms' efforts to convey their operating complexity, limited evidence is offered that would allow for a clearer definition and articulation of the construct. Hence, it is difficult to understand the role and meaning of the conceptual construct that is captured in the "number of numbers" measure.

The remaining papers that use non-SEC text data adopt diverse methods and models (see Table 9). Hobson et al. (2012) are unique in the reviewed accounting literature in their analysis of speech (i.e., indicators of cognitive dissonance identified using proprietary software) in conference call audio files, and are exemplary in their use of an experiment to validate, in the context studied, the main archival data reported. While the barriers to entry for harvesting, cleaning, and analyzing audio files is substantial, scholars who learn these technologies, and who partner with other scholars who possess complementary expertise, may be rewarded with the "big golden potatoes" of exceptional contributions to CATA scholarship.

Two harvested *AOS* papers adopt impression management theories to understand the relationship between corporate governance and Spanish firms' use of impression management in annual earnings press releases (García Osma and Guillamón-Saorín 2011) and North American accounting firms' diverse and evolving use of social media (Suddaby et al. 2015). In a reflection of the editorial board composition and foci of *AOS*, these two papers remain closer to traditional qualitative research practices than any others in the sample. In addition, García Osma and Guillamón-Saorín (2011) use unusual composite variables that are validated in previous research, and Suddaby et al.'s (2015) is the only paper in the sample to use emerging visualization methods. Results reported in García Osma and Guillamón-Saorín (2011) indicate influences of governance structure on annual earnings press release reporting. In addition, Suddaby et al. (2015) provide evidence of an evolving relationship between professional accountants and IT professionals in their analysis of the integration of social media into the professional practices of Big 4 accounting firms.

Bozanic and Thevenot (2015) create a unique measure of within-document lexical diversity, unique words divided by non-unique words, as a measure of information content. Results indicate that this measure is informative to the market, although no

TABLE 9

Other Measures
Non-SEC Sources

Paper	Metric(s)	Description
1. Bozanic and Thevenot (2015)	Linguistic Diversity = unique/non-unique words of earnings press releases.	Within-document comparisons: analyst forecasts.
2. Brochet et al. (2016)	Linguistic Opacity: 1. Non-Plain English 2. Linguistic Errors	1. Non-plain English: sentence length, passive voice verbs, personal pronouns, other (Loughran and McDonald 2014). 2. Linguistic errors: grammatical errors, abnormal article use, frequency of passive voice verbs.
3. Cho et al. (2010)	Environment disclosures: "optimism" and "certainty" scores from DICTION.	Attributes of environmental disclosures.
4. García Osma and Guillamón-Saorín (2011)	Impression management (disclosure tone, emphasis, performance comparisons, and selectivity).	Human assessment of keywords and amounts: annual results press releases of Spanish firms.
5. Hobson et al. (2012)	LVA eX-Sense Pro R voice-based measure of cognitive dissonance.	Conference call audios.
6. Hollander et al. (2010)	Incomplete disclosure = 1 when a question remains unanswered, 0 otherwise.	Questions: conference call transcripts.
7. Kimbrough and Louis (2011)	Historical or forward-looking information based on word counts.	Word counts: conference call transcripts.
8. Kimbrough and Wang (2014)	Seemingly self-serving attributions = 1 if a firm provides a seemingly self-serving attribution in its earnings press release (the earnings announcement press release related to a positive [negative] earnings surprise and had at least one causal statement that linked favorable [unfavorable] factors to internal [external] causes), and 0 otherwise.	Earnings press releases.
9. Lang and Stice-Lawrence (2015)	Boilerplate (redundant) content in annual financial statements.	Lexical bundles (four-word phrases) in target firm versus (1) U.S. peer firms, and (2) non-U.S. peer firms.
10. Li et al. (2014)	Percentage CEO (CFO) text: the ratio of the number of characters spoken by the CEO (CFO) during conference calls to the number of characters spoken by all company executives during conference calls.	Characters (letters) in conference call transcripts.
11. Merkley (2014)	Narrative R&D disclosure: the total number of R&D-related sentences in the 10-K filing.	R&D-related sentences.
12. Suddaby et al. (2015)	Social media word counts.	Word counts in Twitter, Facebook, and firms' web pages.

measures are reported that would allow for an assessment of its relationship to similar measures (e.g., comparability) found in the literature. Lang and Stice-Lawrence's (2015) is one of the few papers in the reviewed literature to adopt the common linguistic practice of examining "bundles" of words rather than examining individual BOW. They examine "tetra grams," i.e., groupings of four words, within internal financial statements to create a complex, non-intuitive measure of the extent to which disclosures are "boilerplate." As with many of the reviewed papers, the procedures and process of identifying boilerplate in Lang and Stice-Lawrence (2015) are reasonable, complex, and difficult to replicate. In addition, no presented evidence links this construct and measure with similar others found in the literature (e.g., Campbell et al. 2014).

Brochet, Naranjo, and Yu (2016) are unique, and exemplary, in their study of the potential for language issues (barriers) to impede non-U.S. firms' communications in conference calls. They create measures of linguistic "opacity" using an assortment of previous measures, some of which are specific to the context studied (e.g., SEC recommendations) and others of which are adopted from the broader linguistic literature (e.g., passive voice verbs, sentence length). In perhaps a demonstration of the

TABLE 10

Distribution of Published Text Analysis Papers by Year
Top Six versus AIS (*JIS, JETA*)

	Journals									
	Top Six Accounting Journals			***JIS* and *JETA***						
				JIS		***JETA***		**Total**		
Year	**Text**	**All**	**%**	**Text**	**All**	**Text**	**All**	**Text**	**All**	**%**
2010	5	229	2.2%	0	11	1	5	1	16	6.3%
2011	5	269	1.9%	0	17	1	5	1	22	4.6%
2012	5	270	1.9%	0	24	0	6	0	30	0%
2013	3	264	1.1%	0	29	1	4	1	33	3.0%
2014	8	279	2.9%	0	29	0	6	0	35	0%
2015	11	305	3.6%	4	26	1	9	5	35	14.3%
2016	9	268	3.4%	0	26	5	28	5	54	9.3%
Total	46	1,884	2.4%	4	162	9	63	13	225	5.8%
Percent		2.4%			2.5%		14.3%		5.8%	

Text = text analysis papers; All = the peer-reviewed papers published in the specific journal and year.

upper limit of human coding of Big Data, two papers (Hollander et al. 2010; Kimbrough and Wang 2014) create dichotomous measures based on an analysis to identify self-serving attributions (Kimbrough and Wang 2014) and incomplete discussion of issues in conference calls (Hollander et al. 2010). In Kimbrough and Wang (2014), two human coders tag all observed self-serving attributions in a sample of over 1,000 conference calls. In Hollander et al. (2010), three coders each tag over 1,000 conference calls. Kimbrough and Louis (2011) use two human coders (an author and research assistant) to manually code a small subsample (n = 20) of stock-for-stock merger announcements found in conference calls and earnings announcements for past (history) and future (forward-looking) content. As with the two studies discussed earlier, agreement between coders is discussed but no agreement data are presented.

Although coder agreement is discussed in three studies, they do not report agreement statistics among coders (Neuendorf 2002; Krippendorff 2004). In addition, the reporting of machine-based coding of the same data set that was coded by humans in these studies would provide useful validation data for future research about the merits and limitations of human and state-of-the-art machine content coding. Li et al. (2014) analyze the percentage of content (characters) in conference calls spoken by CEOs versus CFOs using Fortran code to parse the calls. The online appendix to Li et al. (2014) describes, in detail, the procedures used to generate the data, and supplemental data and analysis tables that provide insight into the validity of the data.

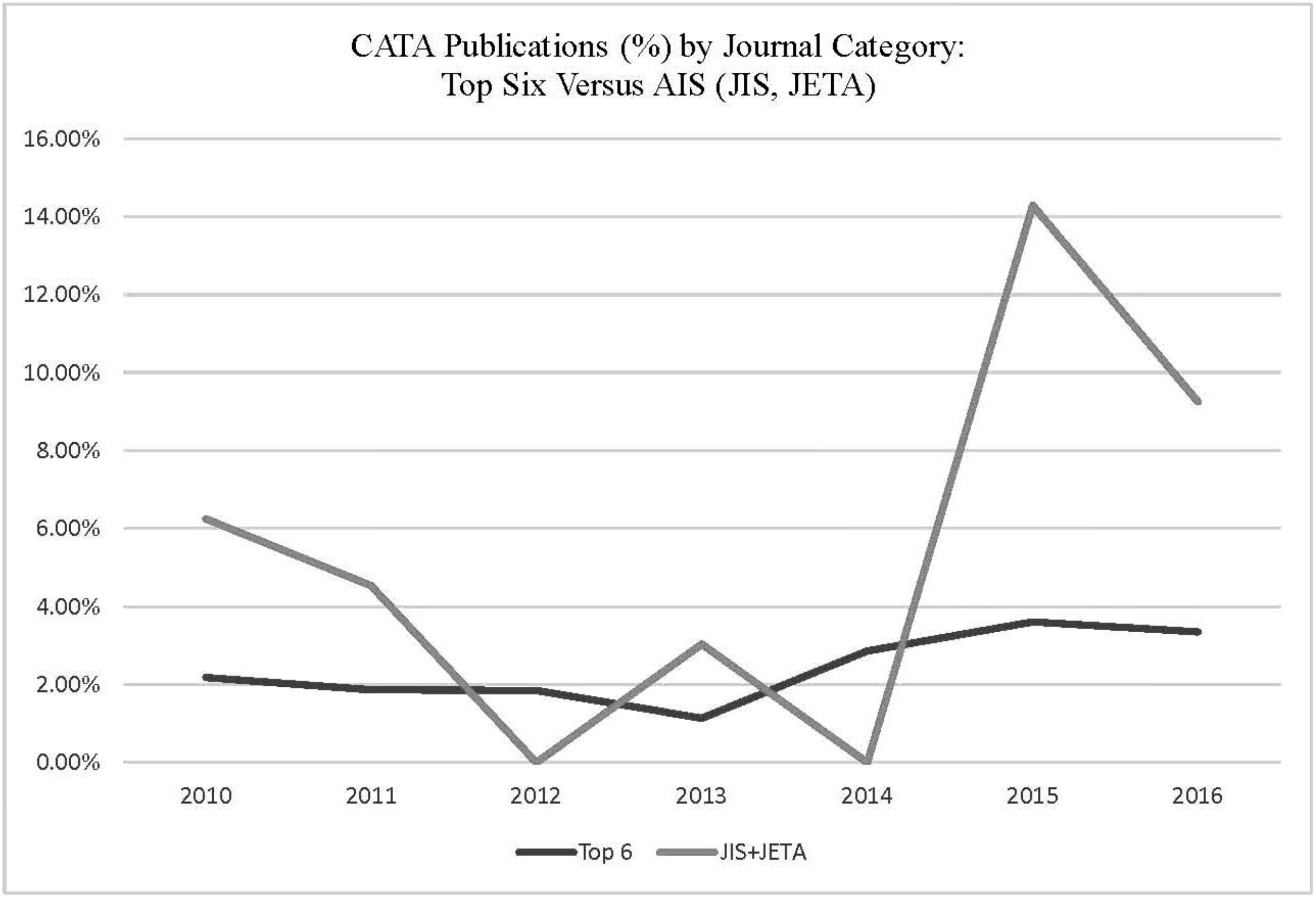
Supplemental Analysis—Comparison of Top Six Journal CATA Accounting and CATA AIS Literatures

Sampling Procedure and Frequency of Publication

A comprehensive comparative analysis of the Top Six accounting versus the accounting information systems (AIS) CATA literatures is worthy of a separate paper. However, we conducted a summary comparison of the CATA research published in the Top Six versus two AIS journals (*JETA* and *JIS*) over the same sample period (i.e., 2010–2016).[11] We used the sampling period and procedures described previously for the two AIS journals. This procedure harvested 13 CATA papers published in *JIS* and *JETA* from 2010–2016 (four in *JIS*, nine in *JETA*). Table 10 presents the distribution of published text analysis papers by year for the Top Six journals versus the AIS (*JIS* and *JETA*) journal sample. The percentage of CATA papers published in AIS journals does not differ from that of the Top Six accounting journals (5.78 percent versus 2.44 percent, respectively; t (6) = 1.68, p = 0.144; n = 7 years). Figure 4 plots the distribution of published papers by year and journal type (i.e., AIS versus Top Six). CATA publications in AIS journals vary more by year than do CATA papers published in Top Six accounting journals (AIS std. dev. = 5.2; Top Six std. dev. = 0.9; untabulated). In addition, AIS journals evidence more recent CATA publications, i.e., 77 percent versus 44 percent of the total number of CATA papers are published in 2015 and 2016 for AIS versus Top Six

[11] We omit the *International Journal of Accounting Information Systems* from our sample. However, even omitting *IJAIS*, our manuscript approaches a book-length manuscript.

FIGURE 4
CATA Publications by Journal Category



journals, respectively. A regression of the number of annual publications on year indicates a marginally positive effect of publication frequency by year for the Top Six journals (the coefficient on year = 0.288; p = 0.085), meaning an increasing number of CATA accounting publications over time, but no significant effect for year in AIS journals (the coefficient on year = 1.0; p = 0.339).

Table 11 presents the text data sources and text variables for published AIS (i.e., *JIS* and *JETA*) papers. This analysis indicates that much CATA AIS research is distinct in its data sources and text variables. We next explore the common and distinct data sources, text variables, and methods of the CATA Top Six and AIS literatures.

AIS Text Data Sources

Three of the 13 published CATA AIS papers use SEC-related data sets, six papers analyze social media, and four papers utilize data sources that are unique to the harvested accounting CATA literatures. Only two of the CATA AIS papers use the primary data source found in CATA accounting Top Six journal publications: SEC data. Goel, Gangolly, Faerman, and Uzuner (2010) conduct a linguistic analysis (using a Naive Bayesian model) for fraud in 10-K filings, a procedure and method that is extended in a paper in our Top Six journal sample, i.e., Larcker and Zakolyukina (2012), in their fraud analysis of conference calls. Pencle and Mălăescu (2016) use SEC data in initial public offering (IPO) filings (S-1 or 424 documents) to develop and validate four dictionaries, i.e., word lists, for categories of CSR activity: social and community, human resources, environment, and human rights. A third CATA AIS paper also uses SEC-related data. Specifically, Liu and Moffitt (2016) analyze SEC comment letters, an SEC data source that is unique in the harvested literature. Liu and Moffitt (2016) analyze whether the

TABLE 11

Text Data Sources and Text Variables in Published AIS Papers

		Text Variables			
Text Data Sources	**No.**	**Tone/Intensity**	**Readability**	**Firm Characteristics and Environment**	**Other**
SEC Filings and Comment Letters	3	Liu and Moffitt (2016): SEC comment letters.		Goel et al. (2010): fraud (10-Ks); Pencle and Mălăescu (2016): CSR (S-1 or 424 filings).	
Social Media	6			Bonsón and Bednárová (2015): YouTube videos of CSR.	Zhou et al. (2015): Corporate use of Facebook and Twitter; Witherspoon and Stone (2013): Yelp reviews of tax preparers; Eschenbrenner et al. (2015): Business objectives in Facebook and Twitter posts; Perdana et al. (2015): Themes in XBRL LinkedIn groups; O'Leary (2015): Crowdsourced tags—accounting and financial information
Other Sources	4	Rich et al. (2016): Government Comprehensive Annual Financial Reports (CAFR).	Boritz et al. (2016): SOX 404 internal control reports.	Shirata et al. (2011): Japanese financial reports.	Fisher and Nehmer (2016): XBRL tagging: GAAP and IFRS taxonomies—inventory.
Total	13	2	1	4	6

strength of linguistic assertions in SEC comment letters, which they label "intensity," predicts the probability of a restatement. Results support their hypothesis.

The six CATA AIS papers that analyze social media data use a diverse set of data sources. Specifically:

1. Witherspoon and Stone (2013) compare the effectiveness of three software tools (i.e., SentiStrength, LIWC2007, and DICTION 6.0) at analyzing sentiment found in 4,063 Yelp reviews of tax preparers.
2. Bonsón and Bednárová (2015) analyze the CSR-relevant videos of 306 European company's official YouTube channels.
3. Eschenbrenner, Nah, and Telaprolu (2015) analyze the Facebook and Twitter use, categorized by business goal, of six large U.S. accounting firms.
4. O'Leary (2015) examines the properties of crowdsourced tags related to accounts receivable on two platforms: Delicious.com and Twitter.
5. Perdana et al. (2015) use machine-learning software (Leximancer) and data mining to investigate the themes present in LinkedIn XBRL groups from 2010 to 2013.
6. Zhou, Lei, J. Wang, Fan, and A. Wang (2015) examine the social media adoption rates of 9,861 public U.S. firms on two platforms: Facebook and Twitter.

Four papers utilize data sources that are unique in the harvested CATA AIS literatures: Shirata et al. (2011) extract text data from Japanese financial reports of bankrupt (n = 90) and nonbankrupt (n = 90) companies. Boritz, Hayes, and Timoshenko (2016) analyze the readability of management and auditor SOX 404 internal control reports (n = 126,271) obtained from Audit Analytics. Fisher and Nehmer (2016) investigate the semantic and syntactic properties of the GAAP and IFRS taxonomies related to inventory. Finally, Rich, Roberts, and Zhang (2016) test the ability of Government Comprehensive Annual Financial Reports (CAFRs) to predict financial reporting delays.

AIS Literature Text Variables, Research Questions, and Methods

Some CATA AIS literature text variables overlap with those found in the Top Six journal CATA accounting literature. Text variables that are common in the Top Six and AIS literature include Liu and Moffitt's (2016) adaptation of LMcD's (2011) measures of language intensity to their application to SEC comment letters and Rich et al.'s (2016) adoption of LMcD's (2011) positive and negative tone words. Similarly, Boritz et al. (2016) utilize the Fog index to measure the readability of internal control reports. The CATA AIS literature also evidences the use of data mining methods and applications, including Goel et. al. (2010), Shirata et al. (2011), and Liu and Moffitt (2016). Unique research questions and methods within the CATA AIS literature include design science mapping of FASB and IASB standards (Fisher and Nehmer 2016), descriptive taxonomy of crowdsourcing tags (O'Leary 2015), use of an IBM product to extract bankruptcy-relevant phrases from corporate reports (Shirata et al. 2011), and harvesting of Yelp reviews to assess the sentiment of taxpayers' linguistic reviews of tax preparers (Witherspoon and Stone 2013).

The reviewed CATA AIS literature is striking for its diversity and innovative research questions. The diversity of this literature is evident in Table 11. At least one paper is found in eight of the 12 possible cells of this table. In addition, nine of the 13 reviewed papers (69.2 percent) are in either "other" column or row. Further, about 46 percent (6/13) of the papers are the only paper found in the cells that they occupy. In short, the CATA AIS literature evidences considerable methodological, topic, and data source diversity. This literature is less than exemplary in its drilling down to focus on a critical topic or topics. To the extent that a topical focus emerges in the CATA AIS literature, it is on social media. However, even within the social media data focus, the CATA AIS literature explores unique-to-each-paper issues. The extent of topical overlap is restricted to two papers that examine social media adoption among public companies (Zhou et al. 2015) and large U.S. accounting firms (Eschenbrenner et al. 2015).

Like the Top Six accounting journal CATA literature, the CATA AIS literature evidences a mixed focus on issues of reliability and validity. For example, Pencle and Mălăescu (2016) are exemplary in their development and validation of four CSR data dictionaries. Similarly, Liu and Moffitt (2016) use human coders to assess the fit of two of LMcD's (2011) word lists to their sample of SEC comment letters. Contrasting these examples, however, are Perdana et al. (2015), who adopt a software tool (i.e., Leximancer), and use it to produce innovative data visualizations, but without presenting evidence of the reliability and validity of the tool in the AIS domain.

The CATA literature in both the Top Six accounting and AIS journals exhibits some evidence of methodological sophistication. Both literatures offer exemplars of validation and reliability, and published papers that show little awareness of or attention to the issues of validation and reliability. The CATA AIS literature evidences more diverse and innovative data sources and a broader set of research questions and topic focal areas within accounting (including accounting systems and taxation). In contrast, the Top Six accounting journal CATA literature has a much stronger focus on financial accounting and evidences larger sample sizes, on average, and some examples of more complex modeling. Both literatures offer exemplars and nadirs of construct validity.

Observed Knowledge and Skills (KS) in CATA Accounting Research

We begin the consideration of the KS required to conduct CATA accounting research by observing the KS found in published research. We conduct automated and manual word searches of the primary and supplemental literature to identify tools in the literature. Table 12 lists the programming languages, software and modeling tools, and computer science labor sources observed in the reviewed literature. Surprisingly, we observe one paper (Li et al. 2014) that uses Fortran, a programming language that we assumed was deceased. We find extensive use of Perl, Python, and R.

The most frequent software in the observed research is database extraction and manipulation tools. In the sample of Top Six accounting journal papers, two CATA packages (i.e., Linguistic Inquiry and Word Count [LIWC] and DICTION), used primarily to do word counts in documents, are penultimate in software use frequency. Six papers use LIWC (the 2007 version), while four use DICTION (version 5.0). Recent use of DICTION has likely declined because of LMcD's (2015) finding that the DICTION sentiment word lists correlate less with market reactions, when applied to SEC 10-K filings. Given the unusual construct validation criterion adopted by LMcD (2015), i.e., that market reaction equates to construct validity, opportunity exists for broadening the scope and quality of validation efforts related to word list constructs used in accounting research. Indeed, many of the observed papers create language-based constructs but provide little or no evidence supporting the validity of these research-created constructs. The validation procedures used in the broader corpus linguistics literature (e.g., see, Pennebaker, Mehl, and Niederhoffer 2003) are largely absent from, but are needed in, CATA accounting research.

Three papers use a VSM to measure similarity (Brown and Knechel 2016; Brown and Tucker 2011; Peterson et al. 2015), and one each uses SVR (Frankel et al. 2016), random forest modeling (Peterson et al. 2015), and the General Inquirer (GI) software (Li 2010b). Across the set of papers, we find evidence of both strong author team member programming expertise

TABLE 12

Computing Tools and Resources in the CATA Accounting Literature

Programming Languages

1. Fortran (Li et al. 2014).
2. Perl (Allee and DeAngelis 2015; Bozanic and Thevenot 2015; Feldman et al. 2010; Henry and Leone 2016; Kravet and Muslu 2013; Lundholm et al. 2014; Mayew et al. 2015; Merkley 2014; Miller, 2010; Rogers et al. 2011).
3. Python (Davis et al. 2015; De Franco et al. 2015; Filzen and Peterson 2015; Peterson et al. 2015; Suddaby et al. 2015).
4. R programming language and suite of tools (Hobson et al. 2012; Larcker and Zakolyukina 2012).

Software and Modeling Tools

1. Database extraction and manipulation (Baginski et al. 2016).
2. DICTION 5.0 (Baginski et al. 2016; Cho et al. 2010; Davis et al. 2012; Davis and Tama-Sweet 2012).
3. General Inquirer (http://www.wjh.harvard.edu/~inquirer/) (Li 2010b).
4. Linguistic Inquiry and Word Count (LIWC) 2007 (Hobson et al. 2012; Huang et al. 2014a; Larcker and Zakolyukina 2012; Lee 2016; Mayew et al. 2015; Purda and Skillicorn 2015).
5. Random Forest Modeling (Purda and Skillicorn 2015).
6. Support Vector Regression (SVR) (Frankel et al. 2016).
7. Vector Space Model (VSM) (Brown and Knechel 2016; Brown and Tucker 2011; Peterson et al. 2015).

Computer Science Labor

1. Hired programmers (Baginski et al. 2016; Davis and Tama-Sweet 2012).
2. Author team programming expertise (Brown and Knechel 2016; Brown and Tucker 2011).

Website Files, Tools, and Resources

1. Brian Miller, Indiana University, Bog Readability Index 10-K data: https://kelley.iu.edu/bpm/activities/bogindex.html.
2. Samuel Bonsall, The Ohio State University, Perl code for cleaning 10-K files in StyleWriter: http://u.osu.edu/bonsall.2/research-data-and-code/.
3. Bill McDonald, University of Notre Dame, dictionaries, scripts, SEC 10-K data links: https://www3.nd.edu/~mcdonald/.
4. Natural Language Processing in Python: http://www.nltk.org/book/.
5. Text Analytic Resources for the humanities: http://digitalhumanities.berkeley.edu/resources/text-analysis-resources.

(Brown and Knechel 2016; Brown and Tucker 2011) and of hired programmers to complete scraping, harvesting, and analysis work.

In summary, we observe multiple tools that are relevant to AIS scholars interested in contributing to CATA accounting research. Perhaps the simplest path to building initial expertise is the use of word list software that also enables user-developed word lists (e.g., LIWC2015, DICTION 7). In addition, several qualitative software packages (e.g., NVivo, ATLAS.ti) include CATA functions, although none are observed in the present samples (Top Six or AIS); also, qualitative software packages can be "user unfriendly." For applications that require coding, Python and R are popular CATA accounting research languages that appear to be replacing Perl (although CATA accounting scholars will encounter Perl code when attempting to replicate accounting scholars' work). Finally, even the paper in our sample that is most tied to the qualitative research tradition (Suddaby et al. 2015) evidences use of sophisticated quantitative CATA methods.

We observe an evolving mix of manual and automated analysis methods, suggesting that while CATA is improving in efficiency and accuracy, manual analysis coding methods, of at least test samples, remain relevant. Particularly encouraging is the observed use of both automated and manual approaches to provide evidence for validating and triangulation methods (e.g., Davis et al. 2012). Of course, human (manual) coding scales poorly to Big Data; hence, Big Data necessitates either machine-learning approaches informed by human coding, or mixed CATA and human coding, where humans code a sample rather than the full data set.

IV. ISSUES IN, AND DIRECTIONS FOR, CATA ACCOUNTING RESEARCH

We next consider four issues, with related implications and directions, for CATA accounting research. We consider both issues that are specific to AIS CATA research and more general issues of CATA accounting research.

Greater Attention to Innovation than Validation

One might expect that accounting scholars would gravitate toward scientific validation work with a paucity of innovation. However, our conclusion regarding the present literatures is strong innovative with less attention to research validity. For example, we observe considerable evidence of the creation of *ad hoc* linguistic measures in the literature but far less attention—bordering on neglect in some cases—of efforts to validate these measures and constructs. Such practices may evidence a rudimentary understanding of the principles of construct validity in science (Shadish, Cook, and Campbell 2002). One reason for this inattention to construct validation may be that, although the sampled research is rooted in economics and finance, the principles of linguistic construct validity and validation derive from psychology, education, and, in some cases, qualitative research (Saldana 2016). Accordingly, an important opportunity in the CATA accounting research is to evidence a deeper understanding of the forms and types of validity applied to CATA accounting data. Such efforts are essential to establishing and enhancing the scientific legitimacy of CATA accountancy research. A necessary complement to building stronger construct validity in the literature is greater attention to cross-citations of published papers where an author team uses a different measurement approach than a previous author team that has published work related to the same construct. While we observe cases of such cross-citation in the reviewed literature, it is, unfortunately, the exception rather than the norm.

Appropriate Use of p-values

The rise and prominence of Big Data has heightened one aspect of the crisis of confidence in science. Specifically, with "Big Data" sample sizes, p-values yield "significant" results that are often meaningless in applied domains (Cumming 2012, 2014; Combs 2010). To date, most research—including in the Top Six CATA journal publications—relies on (largely outdated) null-hypothesis significance testing (NHST) approaches. Many "significant" p-value results in Big Data studies are artifacts of large samples, a corresponding statistical power = 1.0 (Cohen 1969, 1988; Borkowski, Welsh, and Zhang 2001), almost limitless computing power, and sophisticated modeling methods (Mayer-Schönberger and Cukier 2013). This dark nexus of forces enables the production of "p-hacked" articles (Starbuck 2016) where statistical "significance" is a sure result from a lengthy "economic exorcism" of the data (cf. LMcD 2016).

Accordingly, a growing literature (including in accounting [Dyckman 2016; Dyckman and Zeff 2014; Stone 2018]) recognizes the widespread abuse and misuse of p-values. Advances in methods have also resulted in superior means of measuring the size and importance of effects (Cumming 2014, 2012), including emerging methods in data visualization (e.g., Dilla et al. 2010). Because these methods increase the rigor demanded for publication, scholars have resisted their implication. Hence, we find little evidence (except for Suddaby et al. 2015) of these methods in the reviewed literature. While AIS scholars do not have unique expertise in these areas, we note this issue herein because of the growing misunderstanding of the relations between Big Data, computing resources, and NHST statistical tools (including p-values) developed for a bygone era in which data were scarce and sample sizes were small (Mayer-Schönberger and Cukier 2013).

System Development Controls as a Requisite of Research Validity

Controls over the system development process is a core topic in the AIS curriculum, which is covered in every known AIS textbook (e.g., Turner and Weickgenannt 2017; Hurt 2016; Hall 2016; Simkin, Rose, and Strand Norman 2015; Romney and Steinbart 2015). The CATA research process often includes the development of sophisticated information processing systems. That these systems need to be reviewed, tested, and validated seems to be little recognized in the present literature. We find evidence of few system development control processes in the reviewed literature. While research that uses packaged software (e.g., LIWC and DICTION) has the validity of the standardized implementations of computing processes, much of the reviewed research uses custom-developed software, without evidence of independent review, testing, or validation of these researcher-developed systems. That identified errors in custom-developed software and spreadsheets have led to scholars' misrepresenting their results (e.g., see, Herndon, Ash, and Pollin 2014) is alarming, and a potential opportunity for AIS scholars, most of who are trained in systems development models and methods. Accordingly, proposing and demonstrating software development processes that can guide the increasingly sophisticated CATA research software systems that harvest, clean, and generate the data used in CATA research is an important potential contribution of AIS scholars, who have expertise that is critical to improving the integrity of the reviewed literature (cf. Neely and Cook 2011). Opportunities also exist for AIS scholars to partner with colleagues to assist in processes to ensure the integrity and privacy of collected and stored CATA data. For example, some computer science, MIS, and AIS faculty (e.g., Boritz and No 2011; Kauffman et al. 2011; Morris, Kleist, Dull, and Tanner 2014; Raschke, Krishen, and Kachroo 2014; Toy and Hay 2015; Sidgman and Crompton 2016) are experts in issues of data integrity and privacy. Partnerships between such experts and those with CATA expertise may yield superior methods for ensuring data integrity and privacy (when needed) in text data sources.

Opportunities with Emerging Data Sources

While two papers in the sample of CATA AIS literature explore XBRL-related issues (i.e., Perdana et al. 2015; Fisher and Nehmer 2016), implementations of XBRL have, to date, given little attention to text disclosures. Given that AIS scholars have the deepest XBRL-related expertise in the academic accounting community, proposing and testing methods for tagging (even in block form, e.g., see https://xbrl.us/guidance/tagging-table-text-blocks-with-multiple-tags/) text disclosures, and integrating this process with existing CATA tools, represents an unexplored, but important, opportunity for AIS scholars. In addition, we note the potential use of many accounting-relevant data sources (e.g., tax filings, system logs, documents in the public domain, internal corporate, and organizational documents) that are of potential relevance to accounting issues but are not present in the reviewed research. Such are the unexplored opportunities of the reviewed Top Six accounting CATA literature that, to date, few linguistic sources other than those related to financial accounting are present in the reviewed literature. We note, with appreciation, a greater diversity of data sources in the reviewed AIS than Top Six CATA accounting literature.

In addition, the nascent tools for analyzing speech discourse (e.g., in conference calls) promise emergent research opportunities for audio file analysis (Hobson et al. 2012; Mayew and Venkatachalam 2012). Currently, such analysis is of audio transcripts of conference calls using emergent cognitive dissonance software. Recent research demonstrating the value of automated speech analysis as a medical diagnostic tool (e.g., of depression [Cummins et al. 2015] and to identifying speaker expertise [Worsley and Blikstein 2010] and learning [Blikstein and Worsley 2016]) promises future applications in accounting.

V. CONCLUSION

This paper reviews 46 CATA papers published in the six leading accounting journals (*JAR, JAE, TAR, CAR, AOS,* and *RAS*) from 2010 to 2016. We find five main text data sources: SEC filings, analyst reports, conference call transcripts, earnings press releases, and other sources. We also identify five main categories of analyzed text variables: tone, readability, similarity, firm characteristics and environment, and other measures. We review the research sources and related methods for the reviewed studies and propose research directions that are relevant to AIS scholars. In supplemental analysis, we summarize the 13 CATA papers published in two AIS journals (*JIS* and *JETA*) within the same sample period. Our review leads us to speculate that future CATA accounting research will both improve the quality of companies' and standard-setters' text disclosures and advance our knowledge of how text influences accounting reports and users' and analysts' decisions based on these reports.

Several aspects of the present literature lie beyond our scope. Our focus is on the data sources, and potential contributions of AIS scholars, to CATA research. Accordingly, we do not summarize the major substantive findings of the research, nor do we consider, in depth, their non-text variables and data sources. Beyond summarizing the published CATA literature in two AIS journals, we do not review the CATA research published outside of the Top Six accounting journals, nor do we consider research published before 2010. In addition, although important, we do not consider the opportunities for the integration of CATA into the scholarship of teaching AIS, nor do we consider emerging opportunities for the use of text analytics in AIS and professional accounting practice.

The explosion of emerging text analytic sources—coupled with advances in technology—promises great opportunities for entrepreneurial AIS scholars to harvest "golden potatoes" in the emerging "field" of CATA accounting research. But a requisite for such contributions is obtaining the needed KS and partnering with complementary co-authors to form productive teams. Some collaborative teams might include AIS scholars matched with complementary non-AIS accounting, computer science, MIS, IT, communication, and data visualization experts. We anticipate, with excitement, the future contributions of our AIS colleagues to this important literature.

REFERENCES

Allee, K. D., and M. D. DeAngelis. 2015. The structure of voluntary disclosure narratives: Evidence from tone dispersion. *Journal of Accounting Research* 53 (2): 241–274. https://doi.org/10.1111/1475-679X.12072

Badie, B., D. Berg-Schlosser, and L. Morlino. 2011. *International Encyclopedia of Political Science*. Volume 1. New York, NY: Sage.

Baginski, S., E. Demers, C. Wang, and J. Yu. 2016. Contemporaneous verification of language: Evidence from management earnings forecasts. *Review of Accounting Studies* 21 (1): 165–197. https://doi.org/10.1007/s11142-015-9347-6

Baker, P., and J. Egbert. 2016. Triangulating methodological approaches in corpus linguistic research. In *Routledge Advances in Corpus Linguistics*. New York, NY: Routledge.

Blikstein, P., and M. Worsley. 2016. Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics* 3 (2): 220–238. https://doi.org/10.18608/jla.2016.32.11

Bonsall, S. B., IV, A. J. Leone, B. P. Miller, and K. Rennekamp. 2017. A plain English measure of financial reporting readability. *Journal of Accounting & Economics* 63 (2/3): 329–357. https://doi.org/10.1016/j.jacceco.2017.03.002

Bonsón, E., and M. Bednárová. 2015. YouTube sustainability reporting: Empirical evidence from Eurozone-listed companies. *Journal of Information Systems* 29 (3): 35–50. https://doi.org/10.2308/isys-50993

Boritz, J. E., and W. G. No. 2011. E-commerce and privacy: Exploring what we know and opportunities for future discovery. *Journal of Information Systems* 25 (2): 11–45. https://doi.org/10.2308/isys-10090

Boritz, J. E., L. Hayes, and L. M. Timoshenko. 2016. Determinants of the readability of SOX 404 reports. *Journal of Emerging Technologies in Accounting* 13 (2): 145–168. https://doi.org/10.2308/jeta-51593

Borkowski, S. C., M. J. Welsh, and Q. Zhang. 2001. An analysis of statistical power in behavioral accounting research. *Behavioral Research in Accounting* 13 (1): 63–84. https://doi.org/10.2308/bria.2001.13.1.63

Bozanic, Z., and M. Thevenot. 2015. Qualitative disclosure and changes in sell-side financial analysts' information environment. *Contemporary Accounting Research* 32 (4): 1595–1616. https://doi.org/10.1111/1911-3846.12123

Brennan, N. M., E. Guillamón-Saorín, and A. Pierce. 2009. Impression management: Developing and illustrating a scheme of analysis for narrative disclosures—A methodological note. *Accounting, Auditing & Accountability Journal* 22 (5): 789–832. https://doi.org/10.1108/09513570910966379

Brochet, F., P. Naranjo, and G. Yu. 2016. The capital market consequences of language barriers in the conference calls of non-U.S. firms. *The Accounting Review* 91 (4): 1023–1049. https://doi.org/10.2308/accr-51387

Brown, S. V., and W. R. Knechel. 2016. Auditor-client compatibility and audit firm selection. *Journal of Accounting Research* 54 (3): 725–775. https://doi.org/10.1111/1475-679X.12105

Brown, S. V., and J. W. Tucker. 2011. Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research* 49 (2): 309–346. https://doi.org/10.1111/j.1475-679X.2010.00396.x

Campbell, J. L., H. Chen, D. S. Dhaliwal, H. Lu, and L. B. Steele. 2014. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies* 19 (1): 396–455. https://doi.org/10.1007/s11142-013-9258-3

Cao, N. 2016. *Introduction to Text Visualization*. New York, NY: Springer.

Chen, E. E., and S. P. Wojcik. 2016. A practical guide to Big Data research in psychology. *Psychological Methods* 21 (4): 458–474. https://doi.org/10.1037/met0000111

Cho, C. H., R. W. Roberts, and D. M. Patten. 2010. The language of U.S. corporate environmental disclosure. *Accounting, Organizations and Society* 35 (4): 431–443. https://doi.org/10.1016/j.aos.2009.10.002

Cohen, J. 1969. *Statistical Power for the Behavioral Sciences*. San Diego, CA: Academic Press.

Cohen, J. 1988. *Statistical Power for the Behavioral Sciences*. 2nd edition. San Diego, CA: Academic Press.

Combs, J. G. 2010. Big samples and small effects: Let's not trade relevance and rigor for power. *Academy of Management Journal* 53 (1): 9–13. https://doi.org/10.5465/amj.2010.48036305

Cumming, G. 2012. Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. In *Multivariate Applications Series*. New York, NY: Routledge.

Cumming, G. 2014. The new statistics: Why and how. In *Psychological Science*. Thousand Oaks, CA: Sage Publications, Inc.

Cummins, N., S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication* 71: 10–49. https://doi.org/10.1016/j.specom.2015.03.004

Davis, A. K., and I. Tama-Sweet. 2012. Managers' use of language across alternative disclosure outlets: Earnings press releases versus MD&A. *Contemporary Accounting Research* 29 (3): 804–837. https://doi.org/10.1111/j.1911-3846.2011.01125.x

Davis, A. K., J. M. Piger, and L. M. Sedor. 2012. Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research* 29 (3): 845–868. https://doi.org/10.1111/j.1911-3846.2011.01130.x

Davis, A. K., W. Ge, D. Matsumoto, and J. Zhang. 2015. The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies* 20 (2): 639–673. https://doi.org/10.1007/s11142-014-9309-4

De Franco, G., O.-K. Hope, D. Vyas, and Y. Zhou. 2015. Analyst report readability. *Contemporary Accounting Research* 32 (1): 76–104. https://doi.org/10.1111/1911-3846.12062

De Franco, G., F. P. Vasvari, D. Vyas, and R. Wittenberg-Moerman. 2014. Debt analysts' views of debt-equity conflicts of interest. *The Accounting Review* 89 (2): 571–604. https://doi.org/10.2308/accr-50635

Debreceny, R. S., and G. L. Gray. 2011. Data mining of electronic mail and auditing: A research agenda. *Journal of Information Systems* 25 (2): 195–226. https://doi.org/10.2308/isys-10167

Dilla, W., D. J. Janvrin, and R. Raschke. 2010. Interactive data visualization: New directions for accounting information systems research. *Journal of Information Systems* 24 (2): 1–37. https://doi.org/10.2308/jis.2010.24.2.1

Drake, M. S., D. T. Roulstone, and J. R. Thornock. 2012. Investor information demand: Evidence from Google searches around earnings announcements. *Journal of Accounting Research* 50 (4): 1001–1040. https://doi.org/10.1111/j.1475-679X.2012.00443.x

Du, H., and W. Jiang. 2015. Do social media matter? Initial empirical evidence. *Journal of Information Systems* 29 (2): 51–70. https://doi.org/10.2308/isys-50995

Dyckman, T. R. 2016. Significance testing: We can do better. *Abacus* 52 (2): 319–342. https://doi.org/10.1111/abac.12078

Dyckman, T. R., and S. A. Zeff. 2014. Some methodological deficiencies in empirical research articles in accounting. *Accounting Horizons* 28 (3): 695–712. https://doi.org/10.2308/acch-50818

Dyer, T., M. Lang, and L. Stice-Lawrence. 2016. Do managers really guide through the fog? On the challenges in assessing the causes of voluntary disclosure. *Journal of Accounting & Economics* 62 (2/3): 270–276. https://doi.org/10.1016/j.jacceco.2016.08.001

Elliott, W. B., K. M. Rennekamp, and B. J. White. 2015. Does concrete language in disclosures increase willingness to invest? *Review of Accounting Studies* 20 (2): 839–865. https://doi.org/10.1007/s11142-014-9315-6

Eschenbrenner, B., F. Nah, and V. Telaprolu. 2015. Efficacy of social media utilization by public accounting firms: Findings and directions for future research. *Journal of Information Systems* 29 (2): 5–21. https://doi.org/10.2308/isys-50894

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal. 2010. Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies* 15 (4): 915–953. https://doi.org/10.1007/s11142-009-9111-x

Filzen, J. J., and K. Peterson. 2015. Financial statement complexity and meeting analysts' expectations. *Contemporary Accounting Research* 32 (4): 1560–1594. https://doi.org/10.1111/1911-3846.12135

Fisher, I. E., and R. A. Nehmer. 2016. Using language processing to evaluate the equivalency of the FASB and IASB standards. *Journal of Emerging Technologies in Accounting* 13 (2): 129–144. https://doi.org/10.2308/jeta-51621

Fisher, I. E., M. R. Garnsey, S. Goel, and K. Tam. 2010. The role of text analytics and information retrieval in the accounting domain. *Journal of Emerging Technologies in Accounting* 7 (1): 1–24. https://doi.org/10.2308/jeta.2010.7.1.1

Frankel, R., J. Jennings, and J. Lee. 2016. Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting & Economics* 62 (2/3): 209–227. https://doi.org/10.1016/j.jacceco.2016.07.003

García Osma, B., and E. Guillamón-Saorín. 2011. Corporate governance and impression management in annual results press releases. *Accounting, Organizations and Society* 36 (4/5): 187–208. https://doi.org/10.1016/j.aos.2011.03.005

Gartner Reports. 2015. Market guide for text analytics. In *Research Notes*. Stamford, CT: Gartner.

Gartner Reports. 2016. Four data preparation challenges for text analytics. In *Research Notes*. Stamford, CT: Gartner.

Goel, S., J. Gangolly, S. R. Faerman, and O. Uzuner. 2010. Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting* 7 (1): 25–46. https://doi.org/10.2308/jeta.2010.7.1.25

Griffith, E. E., J. S. Hammersley, and K. Kadous. 2015. Audits of complex estimates as verification of management numbers: How institutional pressures shape practice. *Contemporary Accounting Research* 32 (3): 833–863. https://doi.org/10.1111/1911-3846.12104

Guan, J., A. S. Levitan, and S. Goyal. 2018. Text mining using latent semantic analysis: An illustration through examination of 30 years of research at *JIS*. *Journal of Information Systems* 32 (1): 67–86. https://doi.org/10.2308/isys-51625

Guay, W., D. Samuels, and D. Taylor. 2016. Guiding through the fog: Financial statement complexity and voluntary disclosure. *Journal of Accounting & Economics* 62 (2/3): 234–269. https://doi.org/10.1016/j.jacceco.2016.09.001

Hales, J., X. I. Kuang, and S. Venkataraman. 2011. Who believes the hype? An experimental examination of how language affects investor judgments. *Journal of Accounting Research* 49 (1): 223–255. https://doi.org/10.1111/j.1475-679X.2010.00394.x

Hall, J. A. 2016. *Accounting Information Systems*. 9th edition. Boston, MA: Cengage Learning.

Henry, E. 2006. Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting* 3 (1): 1–19. https://doi.org/10.2308/jeta.2006.3.1.1

Henry, E. 2008. Are investors influenced by how earnings press releases are written? *Journal of Business Communication* 45 (4): 363–407. https://doi.org/10.1177/0021943608319388

Henry, E., and A. J. Leone. 2016. Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *The Accounting Review* 91 (1): 153–178. https://doi.org/10.2308/accr-51161

Herndon, T., M. Ash, and R. Pollin. 2014. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics* 38 (2): 257–279. https://doi.org/10.1093/cje/bet075

Hoberg, G., and G. Phillips. 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. *Review of Financial Studies* 23 (10): 3773–3811. https://doi.org/10.1093/rfs/hhq053

Hobson, J. L., W. J. Mayew, and M. Venkatachalam. 2012. Analyzing speech to detect financial misreporting. *Journal of Accounting Research* 50 (2): 349–392. https://doi.org/10.1111/j.1475-679X.2011.00433.x

Hollander, S., M. Pronk, and E. Roelofsen. 2010. Does silence speak? An empirical analysis of disclosure choices during conference calls. *Journal of Accounting Research* 48 (3): 531–563. https://doi.org/10.1111/j.1475-679X.2010.00365.x

Hope, O.-K., D. Hu, and H. Lu. 2016. The benefits of specific risk-factor disclosures. *Review of Accounting Studies* 21 (4): 1005–1045. https://doi.org/10.1007/s11142-016-9371-1

Huang, A. H., A. Y. Zang, and R. Zheng. 2014a. Evidence on the information content of text in analyst reports. *The Accounting Review* 89 (6): 2151–2180. https://doi.org/10.2308/accr-50833

Huang, X., S. H. Teoh, and Y. Zhang. 2014b. Tone management. *The Accounting Review* 89 (3): 1083–1113. https://doi.org/10.2308/accr-50684

Huck, S. W. 2012. *Reading Statistics and Research*. 6th edition. Boston, MA: Pearson.

Hurt, R. L. 2016. *Accounting Information Systems: Basic Concepts and Current Issues*. 4th edition. New York, NY: McGraw-Hill Education.

Hutchison, P. D., T. M. Lee, and C. G. White. 2004. Development of the *Journal of Information Systems (JIS)* from the editors' perspectives. *Journal of Information Systems* 18 (1): 39–51. https://doi.org/10.2308/jis.2004.18.1.39

Janvrin, D. J., and D. A. Wood. 2016. The *Journal of Information Systems* 2015 conference on information technology audit. *Journal of Information Systems* 30 (1): 3–5. https://doi.org/10.2308/isys-10483

Kauffman, R. J., Y. J. Lee, M. Prosch, and P. J. Steinbart. 2011. A survey of consumer information privacy from the accounting information systems perspective. *Journal of Information Systems* 25 (2): 47–79. https://doi.org/10.2308/isys-10091

Kearney, C., and S. Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* 33: 171–185. https://doi.org/10.1016/j.irfa.2014.02.006

Kimbrough, M. D., and H. Louis. 2011. Voluntary disclosure to influence investor reactions to merger announcements: An examination of conference calls. *The Accounting Review* 86 (2): 637–667. https://doi.org/10.2308/accr.00000022

Kimbrough, M. D., and I. Y. Wang. 2014. Are seemingly self-serving attributions in earnings press releases plausible? Empirical evidence. *The Accounting Review* 89 (2): 635–667. https://doi.org/10.2308/accr-50628

Kinney, W. R., Jr. 2001. Accounting scholarship: What is uniquely ours? *The Accounting Review* 76 (2): 275–284. https://doi.org/10.2308/accr.2001.76.2.275

Kosinski, M., Y. Wang, H. Lakkaraju, and J. Leskovec. 2016. Mining Big Data to extract patterns and predict real-life outcomes. *Psychological Methods* 21 (4): 493–506. https://doi.org/10.1037/met0000105

Kravet, T., and V. Muslu. 2013. Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies* 18 (4): 1088–1122. https://doi.org/10.1007/s11142-013-9228-9

Krippendorff, K. 2004. *Content Analysis: An Introduction to Its Methodology*. 2nd edition. Thousand Oaks, CA: Sage.

Lang, M., and L. Stice-Lawrence. 2015. Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting & Economics* 60 (2/3): 110–135. https://doi.org/10.1016/j.jacceco.2015.09.002

Larcker, D. F., and A. A. Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50 (2): 495–540. https://doi.org/10.1111/j.1475-679X.2012.00450.x

Law, K., and L. F. Mills. 2015. Taxes and financial constraints: Evidence from linguistic cues. *Journal of Accounting Research* 53 (4): 777–819. https://doi.org/10.1111/1475-679X.12081

Lawrence, A. 2013. Individual investors and financial disclosure. *Journal of Accounting and Economics* 56 (1): 130–147. https://doi.org/10.1016/j.jacceco.2013.05.001

Lee, C.-C., R. B. Welker, and M. D. Odom. 2009. Features of computer-mediated, text-based messages that support automatable, linguistics-based indicators for deception detection. *Journal of Information Systems* 23 (1): 5–24. https://doi.org/10.2308/jis.2009.23.1.24

Lee, J. 2016. Can investors detect managers' lack of spontaneity? Adherence to predetermined scripts during earnings conference calls. *The Accounting Review* 91 (1): 229–250. https://doi.org/10.2308/accr-51135

Lee, Y.-J. 2012. The effect of quarterly report readability on information efficiency of stock prices. *Contemporary Accounting Research* 29 (4): 1137–1170. https://doi.org/10.1111/j.1911-3846.2011.01152.x

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* 86 (3): 1087–1115. https://doi.org/10.2308/accr.00000043

Li, F. 2010a. Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature* 29: 143–167.

Li, F. 2010b. The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research* 48 (5): 1049–1102. https://doi.org/10.1111/j.1475-679X.2010.00382.x

Li, F., R. Lundholm, and M. Minnis. 2013. A measure of competition based on 10-K filings. *Journal of Accounting Research* 51 (2): 399–436. https://doi.org/10.1111/j.1475-679X.2012.00472.x

Li, F., M. Minnis, V. Nagar, and M. Rajan. 2014. Knowledge, compensation, and firm value: An empirical analysis of firm communication. *Journal of Accounting & Economics* 58 (1): 96–116. https://doi.org/10.1016/j.jacceco.2014.06.003

Liu, Y., and K. C. Moffitt. 2016. Text mining to uncover the intensity of SEC comment letters and its association with the probability of 10-K restatement. *Journal of Emerging Technologies in Accounting* 13 (1): 85–94. https://doi.org/10.2308/jeta-51438

Loughran, T., and B. McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66 (1): 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., and B. McDonald. 2014. Measuring readability in financial disclosures. *The Journal of Finance* 69 (4): 1643–1671. https://doi.org/10.1111/jofi.12162

Loughran, T., and B. McDonald. 2015. The use of word lists in textual analysis. *Journal of Behavioral Finance* 16 (1): 1–11. https://doi.org/10.1080/15427560.2015.1000335

Loughran, T., and B. McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54 (4): 1187–1230. https://doi.org/10.1111/1475-679X.12123

Lundholm, R. J., R. Rogo, and J. L. Zhang. 2014. Restoring the Tower of Babel: How foreign firms communicate with U.S. investors. *The Accounting Review* 89 (4): 1453–1485. https://doi.org/10.2308/accr-50725

Masli, A., V. J. Richardson, J. M. Sanchez, and R. E. Smith. 2011. The business value of IT: A synthesis and framework of archival research. *Journal of Information Systems* 25 (2): 81–116. https://doi.org/10.2308/isys-10117

Mayer-Schönberger, V., and K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt.

Mayew, W. J., and M. Venkatachalam. 2012. The power of voice: Managerial affective states and future firm performance. *The Journal of Finance* 67 (1): 1–43. https://doi.org/10.1111/j.1540-6261.2011.01705.x

Mayew, W. J., M. Sethuraman, and M. Venkatachalam. 2015. MD&A disclosure and the firm's ability to continue as a going concern. *The Accounting Review* 90 (4): 1621–1651. https://doi.org/10.2308/accr-50983

McNamara, D. S., A. C. Graesser, P. M. McCarthy, and Z. Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

Merkley, K. J. 2014. Narrative disclosure and earnings performance: Evidence from R&D disclosures. *The Accounting Review* 89 (2): 725–757. https://doi.org/10.2308/accr-50649

Miller, B. P. 2010. The effects of reporting complexity on small and large investor trading. *The Accounting Review* 85 (6): 2107–2143. https://doi.org/10.2308/accr.00000001

Miller, G. S., and D. J. Skinner. 2015. The evolving disclosure landscape: How changes in technology, the media, and capital markets are affecting disclosure. *Journal of Accounting Research* 53 (2): 221–239. https://doi.org/10.1111/1475-679X.12075

Moffitt, K. C., and M. A. Vasarhelyi. 2013. AIS in an age of Big Data. *Journal of Information Systems* 27 (2): 1–19. https://doi.org/10.2308/isys-10372

Moffitt, K. C., V. J. Richardson, N. M. Snow, M. M. Weisner, and D. A. Wood. 2016. Perspectives on past and future AIS research as the *Journal of Information Systems* turns thirty. *Journal of Information Systems* 30 (3): 157–171. https://doi.org/10.2308/isys-51495

Morris, B. W., V. Kleist, R. B. Dull, and C. D. Tanner. 2014. Secure information market: A model to support information sharing, data fusion, privacy, and decisions. *Journal of Information Systems* 28 (1): 269–285. https://doi.org/10.2308/isys-50705

Myers, N., N. Snow, S. L. Summers, and D. A. Wood. 2016. Accounting institution citation-based research rankings by topical area and methodology. *Journal of Information Systems* 30 (3): 33–62. https://doi.org/10.2308/isys-51343

Neely, M. P., and J. S. Cook. 2011. Fifteen years of data and information quality literature: Developing a research agenda for accounting. *Journal of Information Systems* 25 (1): 79–108. https://doi.org/10.2308/jis.2011.25.1.79

Neuendorf, K. A. 2002. *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage Publications.

O'Leary, D. E. 2015. Crowdsourcing tags in accounting and finance: Review, analysis, and emerging issues. *Journal of Emerging Technologies in Accounting* 12 (1): 93–115. https://doi.org/10.2308/jeta-51195

Pencle, N., and I. Mălăescu. 2016. What's in the words? Development and validation of a multidimensional dictionary for CSR and application using prospectuses. *Journal of Emerging Technologies in Accounting* 13 (2): 109–127. https://doi.org/10.2308/jeta-51615

Pennebaker, J. W. 2004. Expressive writing and the regulation of emotion over time. *Psychophysiology* 41: S23.

Pennebaker, J. W., M. R. Mehl, and K. G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology* 54 (1): 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Perdana, A., A. Robb, and F. Rohde. 2015. XBRL diffusion in social media: Discourses and community learning. *Journal of Information Systems* 29 (2): 71–106. https://doi.org/10.2308/isys-50996

Peterson, K., R. Schmardebeck, and T. J. Wilks. 2015. The earnings quality and information processing effects of accounting consistency. *The Accounting Review* 90 (6): 2483–2514. https://doi.org/10.2308/accr-51048

Prokofieva, M. 2015. Twitter-based dissemination of corporate disclosure and the intervening effects of firms' visibility: Evidence from Australian-listed companies. *Journal of Information Systems* 29 (2): 107–136. https://doi.org/10.2308/isys-50994

Purda, L., and D. Skillicorn. 2015. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research* 32 (3): 1193–1223. https://doi.org/10.1111/1911-3846.12089

Raschke, R. L., A. S. Krishen, and P. Kachroo. 2014. Understanding the components of information privacy threats for location-based services. *Journal of Information Systems* 28 (1): 227–242. https://doi.org/10.2308/isys-50696

Rennekamp, K. 2012. Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research* 50 (5): 1319–1354. https://doi.org/10.1111/j.1475-679X.2012.00460.x

Rich, K. T., B. L. Roberts, and J. X. Zhang. 2016. Linguistic tone of municipal management discussion and analysis disclosures and future financial reporting delays. *Journal of Emerging Technologies in Accounting* 13 (2): 93–107. https://doi.org/10.2308/jeta-51618

Rogers, J. L., A. Van Buskirk, and S. L. C. Zechman. 2011. Disclosure tone and shareholder litigation. *The Accounting Review* 86 (6): 2155–2183. https://doi.org/10.2308/accr-10137

Romney, M. B., and P. J. Steinbart. 2015. *Accounting Information Systems*. 13th edition. Boston, MA: Pearson.

Saldana, J. 2016. *The Coding Manual for Qualitative Researchers*. 3rd edition. Los Angeles, CA: SAGE.

Salton, G., A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18 (11): 613–620. https://doi.org/10.1145/361219.361220

Schmidt, P. J., J. T. Wood, and S. V. Grabski. 2016. Business in the cloud: Research questions on governance, audit, and assurance. *Journal of Information Systems* 30 (3): 173–189. https://doi.org/10.2308/isys-51494

Scott, S. V., and W. J. Orlikowski. 2012. Reconfiguring relations of accountability: Materialization of social media in the travel sector. *Accounting, Organizations and Society* 37 (1): 26–40. https://doi.org/10.1016/j.aos.2011.11.005

Segal, B., and D. Segal. 2016. Are managers strategic in reporting non-earnings news? Evidence on timing and news bundling. *Review of Accounting Studies* 21 (4): 1203–1244. https://doi.org/10.1007/s11142-016-9366-y

Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Shirata, C. Y., and M. Sakagami. 2008. An analysis of the "going concern assumption": Text mining from Japanese financial reports. *Journal of Emerging Technologies in Accounting* 5 (1): 1–16. https://doi.org/10.2308/jeta.2008.5.1.1

Shirata, C. Y., H. Takeuchi, S. Ogino, and H. Watanabe. 2011. Extracting key phrases as predictors of corporate bankruptcy: Empirical analysis of annual reports by text mining. *Journal of Emerging Technologies in Accounting* 8 (1): 31–44. https://doi.org/10.2308/jeta-10182

Sidgman, J., and M. Crompton. 2016. Valuing personal data to foster privacy: A thought experiment and opportunities for research. *Journal of Information Systems* 30 (2): 169–181. https://doi.org/10.2308/isys-51429

Simkin, M. G., J. M. Rose, and C. Strand Norman. 2015. *Core Concepts of Accounting Information Systems*. 13th edition. Hoboken, NJ: Wiley.

Starbuck, W. H. 2016. 60th anniversary essay: How journals could improve research practices in social science. *Administrative Science Quarterly* 61 (2): 165–183. https://doi.org/10.1177/0001839216629644

Stone, D. 2018. The "new statistics" and nullifying the null: Twelve actions for improving accounting research quality and integrity. *Accounting Horizons* 32 (1): 105–120. https://doi.org/10.2308/acch-51949

StyleWriter. n.d. *StyleWriter Version 4*. Available at: http://www.stylewriter-usa.com/stylewriter-editing-readability.php

Suddaby, R., G. D. Saxton, and S. Gunz. 2015. Twittering change: The institutional work of domain change in accounting expertise. *Accounting, Organizations and Society* 45: 52–68. https://doi.org/10.1016/j.aos.2015.07.002

Summers, S. L., and D. A. Wood. 2016. *Accounting Research Rankings*. Available at: http://www.byuaccounting.net/rankings/univrank/rank_university.php?sortorder=ranking6&qurank=All

Tan, H.-T., E. Wang, and B. Zhou. 2014. When the use of positive language backfires: The joint effect of tone, readability, and investor sophistication on earnings judgments. *Journal of Accounting Research* 52 (1): 273–302. https://doi.org/10.1111/1475-679X.12039

Tan, H.-T., E. Wang, and B. Zhou. 2015. How does readability influence investors' judgments? Consistency of benchmark performance matters. *The Accounting Review* 90 (1): 371–393. https://doi.org/10.2308/accr-50857

Toy, A., and D. C. Hay. 2015. Privacy auditing standards. *Auditing: A Journal of Practice & Theory* 34 (3): 181–199. https://doi.org/10.2308/ajpt-50932

Turner, L., and A. Weickgenannt. 2017. *Accounting Information Systems: Controls and Processes*. 3rd edition. Hoboken, NJ: John Wiley & Sons, Inc.

Vasarhelyi, M. A., A. Kogan, and B. M. Tuttle. 2015. Big Data in accounting: An overview. *Accounting Horizons* 29 (2): 381–396. https://doi.org/10.2308/acch-51071

Witherspoon, C. L., and D. N. Stone. 2013. Analysis and sentiment detection in online reviews of tax professionals: A comparison of three software packages. *Journal of Emerging Technologies in Accounting* 10 (1): 89–115. https://doi.org/10.2308/jeta-50747

Worsley, M., and P. Blikstein. 2010. *What's an Expert? Using Learning Analytics to Identify Emergent Markers of Expertise through Automated Speech, Sentiment and Sketch Analysis*. Available at: https://pdfs.semanticscholar.org/eab4/e746e941b20fd01a5a0a52200b2f25634a90.pdf

Zhou, M., L. Lei, J. Wang, W. Fan, and A. G. Wang. 2015. Social media adoption and corporate disclosure. *Journal of Information Systems* 29 (2): 23–50. https://doi.org/10.2308/isys-50961

APPENDIX A
Summaries of Surveyed Literature

TABLE 13
Summary: Data from SEC 10-K Filings

Author	Journal	Text Data	Period	Text Variables	Research Questions
Brown and Knechel (2016)	*JAR*	33,293 firm-years for the business description; 31,213 for MD&A; and 22,026 for the financial footnotes.	1997–2009	Similarity (VSM) in "auditor-industry-year reference groups."	Do auditor-client compatibility, as proxied by the similarity of a client with all other clients of the same auditor in the same industry and same year, affect auditor switching and audit quality?
Brown and Tucker (2011)	*JAR*	28,142 firm-years.	1997–2006	Similarity (VSM) within firms over time.	How do year-to-year MD&A modifications influence investors' reaction and analysts' response?
Campbell et al. (2014)	*RAS*	9,076 firm-year 10-K filings.	2005–2009	Firm Characteristics and Environment (financial risk, tax risk, litigation risk, other systematic risk, and other idiosyncratic risk).	Are the length and content of risk factor disclosures in 10-K filings mandated by the SEC associated with pre-disclosure proxies for firm risk? Do these mandatory risk factor disclosures have information content?
Cho et al. (2010)	*AOS*	190 firms.	2002	Tone (DICTION: optimism) and Other Measures (DICTION: certainty).	Do self-serving biases influence firms' environmental disclosures in 10-K filings?
Davis and Tama-Sweet (2012)	*CAR*	16,141 firm-quarter observations.	1998–2003	Tone (DICTION; Henry 2006; LMcD 2011).	How do managers use language strategically across alternative disclosure outlets in earnings press releases versus MD&A disclosures in 10-K filings?
Filzen and Peterson (2015)	*CAR*	34,616 firm-years.	1994–2008	Readability (document length relabeled as "complexity").	Are firms with greater financial statement complexity, as proxied by the length of their accounting policy disclosures in 10-K filings, more likely to just meet or beat analyst earnings forecasts?
Frankel et al. (2016)	*JAE*	71,847 observations with 296,329 unique words and phrases in the MD&A.	1994–2013	Tone and Other Measures (MD&A accruals estimated using support vector regression [SVR]).	Are support vector regressions (SVRs) useful in assessing the content of unstructured, qualitative disclosures by relating MD&A-based SVR accrual estimates (MD&A accruals) to actual accruals?
Guay et al. (2016)	*JAE*	72,366 firm-years.	1995–2012	Readability (complexity of financial statements proxied by multiple measures of readability and length).	Do managers use voluntary disclosure to mitigate the negative effect of complex financial statements on a firm's information environment?
Hope et al. (2016)	*RAS*	14,865 firm-year risk factor disclosures (Item 1A in the 10-K report).	2006–2011	Firm Characteristics and Environment (specificity of risk disclosures proxied by names of competitors).	Whether and how do investors and analysts benefit from more specific risk factor disclosures in 10-K filings?

(continued on next page)

TABLE 13 (continued)

Author	Journal	Text Data	Period	Text Variables	Research Questions
Kravet and Muslu (2013)	*RAS*	28,110 firm-years (4,315 firms).	1994–2007	Firm Characteristics and Environment (risk).	How are changes in firms' text risk disclosures in 10-K filings associated with changes in stock market reaction and analyst activity around the 10-K filings?
Law and Mills (2015)	*JAR*	Unrecognized tax benefit (UTB) tests: 5,418 firm-years (2,340 firms), 2007–2011. Tests of tax havens and effective tax rates (ETRs): 12,515 firm-years (3,286 firms), 1993-2011.	1994–2011	Tone as a proxy for financial constraints.	Are financial constraints, measured based on the use of negative words in 10-K filings, associated with firms' aggressive tax planning strategies in the concurrent and future periods?
Lawrence (2013)	*JAE*	91,228 account-firm-year observations with 1,555 separate firm-year observations.	1994–1996	Readability (Fog; length).	Are individuals' shareholdings and returns increasing in firms with clear and concise financial disclosures? Do personal traits and financial disclosures interact to influence individuals' shareholdings?
Lehavy et al. (2011)	*TAR*	33,704 firm-years.	1995–2006	Readability (Fog).	How does annual report readability affect analyst following and the properties of analyst earnings forecasts?
Li et al. (2013)	*JAR*	33,492 firm-years.	1995–2009	Firm Characteristics and Environment (competition) and cosine similarity.	Is the measure of competition, based on management's discussion of competition in 10-K filings, related to future operating performance?
Mayew et al. (2015)	*TAR*	460 bankrupt firms.	1995–2012	Tone (LMcD 2011).	Can text disclosures in 10-K MD&A predict going concern issues?
Merkley (2014)	*TAR*	22,482 firm-years.	1996–2007	Tone and Readability (R&D disclosure quantity, detail, tone, and readability).	How does earnings performance firms' narrative R&D disclosure characteristics (e.g., disclosure quantity, detail, tone, and readability) and are such R&D disclosures informative to market participants?
Miller (2010)	*TAR*	12,771 firm-year 10-K filings with 3,809 unique firms.	1994–2006	Readability (complexity proxied by readability and length).	What are the effects of financial reporting complexity on investors' trading behavior?
Peterson et al. (2015)	*TAR*	44,772 firm-years (4,688 firms).	1994–2012	Similarity (VSM).	Is accounting consistency (over time and across firms), proxied by the text similarity of accounting policy footnotes in 10-K filings, related to earnings quality, information asymmetry, analyst behavior, and stock return synchronicity?

TABLE 14

Summary: Data from Both 10-K and Other SEC Filings

Author	Filing	Journal	Text Data	Period	Text Variables	Research Questions

Bozanic and Thevenot (2015)	8-K	*CAR*	1,838 earnings release reports (160 firms).	2004–2012	Readability (Fog);
					Similarity (VSM);
					Other Measures (diversity).
						How do the 8-K filing disclosure readability, text similarity, and lexical diversity affect analysts' information environment?

Feldman et al. (2010)	10-K; 10-Q	*RAS*	153,988 firm-quarters.	1994–2007	Tone (LMcD 2011).	Do narrative tone changes in MD&A sections of 10-Q and 10-K filings have incremental information content?
Henry and Leone (2016)	8-K	*TAR*	63,357 firms-quarters earnings releases.	2004–2012	Tone (FD wordlist; DICTION wordlist; GI wordlist; LMcD 2011 wordlist).	Do word frequency tone measures based on domain-specific wordlists better predict market reaction to earnings announcements relative to word frequency tone measures based on general wordlists or the Naïve Bayesian machine-learning tone measure? Does inverse document frequency weighting outperform equal weighting?
Lee (2012)	10-Q	*CAR*	60,161 firm-quarters (5,372 firms).	2001–2007	Readability (Fog; length).	How does quarterly report readability of 10-Q filings affect the information efficiency of stock prices?
Li (2010b)	10-K; 10-Q	*JAR*	145,479 firm-quarters.	1994–2007	Tone (Naïve Bayes).	Do the forward-looking statements (FLS) in the MD&A section of 10-K and 10-Q filings have information content?
Lundholm et al. (2014)	10-K; 20-F; 8-K; 6-K	*TAR*	37,344 U.S. firm-year MD&A in the 10-K and 3,499 foreign firm-year MD&A in the 20-F from 45 different countries; 21,976 U.S. firm-year earnings press releases in the 8-K and 1,582 foreign firm-year earnings press releases in the 6-K.	2000–2012	Readability (Fog index); Other Measures (number of numbers excluding dates).	How do foreign firms cross-listed in the U.S. manage the readability of text and the use of numbers in their annual filings and earnings press releases to communicate with U.S. investors?
Purda and Skillicorn (2015)	10-K; 10-Q	*CAR*	4,895 reports.	1994-2006	Firm Characteristics and Environment (probability of truth).	Is the probability-of-truth, a text-based fraud detection measure developed by the authors, an effective predictor of fraud? How does the probability-of-truth measure compare with alternative measures of fraud detection?
Rogers et al. (2011)	8-K	*TAR*	628 earnings announcements for the 165 sued firms and 625 announcements for the 165 matched firms.	2003–2008	Tone (DICTION; Henry 2008; LMcD 2011).	Does managers' use of optimistic language increase litigation risk?
Segal and Segal (2016)	8-K	*RAS*	167,470 non-earnings 8-K reports filed by 5,685 firms.	2005–2013	Tone (LMcD 2011)	Do firms report negative news when investor attention is low, and do they bundle positive and negative news?

TABLE 15

Summary: Data from Conference Calls (Transcripts and Audio Files)

Author	Journal	Text Data	Period	Text Variables	Research Questions
Allee and DeAngelis (2015)	*JAR*	33,428 conference calls (3,345 firms).	2004–2014	Tone (average reduced frequency [ARF]).	What is the role of tone dispersion in earnings conference calls, a measure of narrative structure, in voluntary financial disclosure?
Brochet et al. (2016)	*TAR*	11,305 conference calls (4,512 firm-years).	2002–2010	Other Measures (linguistic complexity [non-plain English, SEC 1998 Plain English Initiative] and linguistic errors [the use of erroneous expressions]).	How do language barriers affect the capital market reaction to information disclosures?
Davis et al. (2015)	*RAS*	2,098 conference calls (225 firms, and 121 individual managers).	2002–2009	Tone (DICTION; Henry 2008; LMcD 2011).	How does the manager-specific component of optimism affect the tone of earnings conference calls?
Frankel et al. (2016)	*JAE*	55,869 quarterly earnings conference call transcripts with 18,232 firm-years.	2002–2013	Tone and Other (conference call accruals estimated using support vector regression [SVR]).	Are support vector regressions (SVRs) useful in assessing the content of unstructured, qualitative disclosures by relating conference-call-based SVR accrual estimates to actual accruals?
Hobson et al. (2012)	*JAR*	1,572 streamed audios from quarterly earnings conference calls.	2006–2007	Other Measures (cognitive dissonance).	Are vocal markers of cognitive dissonance useful for detecting financial misreporting?
Hollander et al. (2010)	*JAR*	1,194 earnings-related conference calls.	2004	Other Measures (incomplete disclosure proxied by at least one question per call remaining unanswered).	Do managers suppress information that is in demand in conference calls and, if so, under what conditions?
Kimbrough and Louis (2011)	*TAR*	294 conference calls for 147 treatment-control pairs of cash-financed mergers and 40 conference calls for 20 treatment-control pairs of stock-for-stock mergers.	2002–2006	Other Measures (historical or forward-looking information based on word counts).	What are the determinants and consequences of acquirers' decisions to supplement merger announcement press releases with conference calls?
Larcker and Zakolyukina (2012)	*JAR*	29,663 conference calls.	2003–2007	Tone (LIWC, WordNet).	Can the linguistic features of CEOs and CFOs in earnings conference calls help identify financial misreporting?
Lee (2016)	*TAR*	40,820 earnings conference calls (2,863 firms).	2002–2011	Similarity (cosine similarity measure) between Q&A and management discussion portions of conference calls.	Can investors detect managers' adherence to predetermined scripts during earnings conference calls?
Li et al. (2014)	*JAE*	17,419 earnings conference calls.	2003–2007	Other Measures (Percentages of CEO [CFO] text).	Do external communication patterns reveal the location of knowledge within the management team?

TABLE 16

Summary: Data from Earnings Press Releases

Author	Journal	Text Data	Period	Text Variables	Research Questions
Davis et al. (2012)	*CAR*	23,017 firm-quarter observations.	1998–2003	Tone (DICTION).	How does the language of quarterly earnings press releases affect firm performance and market reaction?
Davis and Tama-Sweet (2012)	*CAR*	16,141 firm-quarter observations.	1998–2003	Tone (DICTION; Henry 2006; LMcD 2011).	How does management strategically communicate financial results to capital markets by mapping language to disclosure outlets?
Huang et al. (2014b)	*TAR*	14,475 firm-year observations.	1997–2007	Tone (LMcD 2011).	Whether and when do firms manage the tone of words in earnings press releases, and how do investors react to tone management?
Kimbrough and Wang (2014)	*TAR*	1,790 firm-quarters earnings press releases related to 98 firms (random selection).	1999–2005	Other Measures (seemingly self-serving attribution).	Do investors calibrate their responses to earnings announcements that contain seemingly self-serving attributions? How do investors assess the plausibility of seemingly self-serving attributions in unstructured real-world settings?

TABLE 17

Summary: Data from Other Text Sources

Author	Journal	Text Data	Text Data Sources	Period	Text Variables	Research Questions
Baginski et al. (2016)	*RAS*	1,764 firm-years (750 firms).	Management earnings forecasts.	1997–2006	Tone (LMcD 2011).	Does information in the disclosure itself provide a contemporaneous verification of the credibility of language and thus strengthens its market pricing?
García Osma and Guillamón-Saorín (2011)	*AOS*	243 publicly listed Spanish companies, out of which 142 release Annual Results Press Releases (ARPR), while 92 of them do not.	Annual Results Press Releases of Spanish companies.	2005–2006	Tone and Other Measures (impression management as proxied by disclosure tone, emphasis, performance comparisons, and selectivity).	What is the role of internal corporate governance mechanisms in limiting managerial self-serving disclosure practices in press releases?
Lang and Stice-Lawrence (2015)	*JAE*	87,608 firm-year observations (42 countries).	International financial annual reports.	1998–2011	Readability (Fog, length); Similarity (cosine similarity). Other Measures (boilerplate).	What are the determinants and consequences of international annual report text disclosure quality?
Suddaby et al. (2015)	*AOS*	7,050 tweets, 590 Facebook pages, 25,921 firm web pages, 5 interviews.	Social media (Twitter, Facebook, firms' webpages).	2010–2011	Other Measures (social media word counts).	How is the domain of accounting expertise reconstituted in new social media—Facebook, LinkedIn, and Twitter—in Big 4 accounting firms?

TABLE 18

Summary: Data from Analyst Reports

Author	Journal	Text Data	Period	Text Variables	Research Questions
De Franco et al. (2015)	*CAR*	356,463 equity analysts' reports (4,014 unique firms and 2,334 distinct analysts).	2002–2009	Tone (Naïve Bayes); Readability (length).	What are the determinants of and trading volume reaction to analyst report readability?
De Franco et al. (2014)	*TAR*	11,052 debt analysts' reports (39,121 conflict-event extractions).	2002–2007	Tone (Naïve Bayes).	How does the tone of sell-side debt analysts' discussions about debt-equity conflict events affect the informativeness of debt analysts' reports in debt markets?
Huang, Zang, and Zheng (2014a)	*TAR*	363,952 equity analysts' reports.	1995–2008	Readability (Fog, Flesch-Kincaid, Flesch Reading Ease; length).	Does the text in analysts' reports have information content?