
This study presents a methodology for utilizing TAPOR, a natural language processing tool, to perform preliminary analysis on interview transcripts. In order to test the framework we utilized data from a previous study involving ethnographic interviewing techniques. The case study is comprised of 20 transcripts from interviews conducted with scholars at the University of North Carolina at Chapel Hill.

Findings suggest that utilizing a textual analysis tool as a preliminary method of transcript analysis could be valuable in reducing the amount of staff time spent analyzing interview transcripts prior to having data that indicate transcripts contain valuable information that is relevant to the scope of research interests. Adopting this methodology as a part of the transcript analysis process could provide valuable information to researchers earlier in the research process by identifying key thematic areas of interest prior to engaging in extended qualitative analysis coding sessions.

Headings:

College and university libraries -- Open access publishing

College and university libraries -- Scholarly publishing

Natural Language Processing

Research techniques/Evaluation
EVALUATING ALTERNATE AVENUES: ASSESSING THE EFFECTIVENESS OF AUTOMATIC TEXT PROCESSING AS A PRECURSOR TO HUMAN TRANSCRIPT ANALYSIS

by
Amber N. Welch

A Master’s paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Information Science.

Chapel Hill, North Carolina
November 2011

Approved by

_______________________________________
Stephanie W. Haas
Introduction

Qualitative interview data could provide libraries and library administrators with valuable information for service assessment, but can be lengthy and difficult to analyze. Uncovering unique or unexpected patterns in transcripts prior to investing significant staff time and energies to analysis could increase the use of this type of data. Qualitative interviews are rich sources of information that “present human behavior in a way that takes not only the physical and social context into account, but also the actors’ intentionality” (Schultze & Avital, 2011, p. 3).

Transparency in patron intentionality as revealed through qualitative interviews could lead to assessment based on knowledge about researcher processes and opinions that remain largely unexplored in traditional survey-based methods of inquiry. Increasing the efficiency with which library administrators can engage with interview data and extract patterns and themes could lead to increased use of qualitative research methods in the academic library environment. This would increase the array of research methodologies that are considered viable options for library assessment and institutional evaluation.

The goal of this study is to test a methodological framework for data analysis that could be applied at any institution wanting to research a specific topic but not necessarily having the human capital necessary to invest large amounts of time and energy to data analysis. Data for the case study utilized to test this method was generated by a previous study utilizing ethnographic interviewing techniques with a moderately sized cohort of
faculty and instructor authors at a research institution. Interview questions prompted reflective consideration of the decision-making processes for determining a final publication venue. In addition to publication venue, there were also discussions about the nature of collaborative research and writing projects, funding and access to research funding, open access publishing, and resources provided by the on-campus library system.

Surveys and personal, brief interactions with institutional affiliates (including authors, instructors, professors, and students) can provide some insight into publication processes; however, interviews that are semi-structured, and thereby fertile for open ended dialogue, will give libraries and librarians deeper access to behavioral processes. Awareness of these processes may highlight areas where library services intersect with the workflow of academic researchers. Knowledge about this intersection could provide fruitful information for service improvement and development. There are several impediments to the use of interviewing techniques in libraries, including time associated with recruitment, scheduling and conducting interviews, as well as time needed for transcribing and coding interview transcripts.

In this study, we examine the use of Textual Analysis Portal (TAPOR)\(^1\), as a preliminary investigative method for transcript analysis. If this method is deemed effective, using natural language processing tools to uncover patterns in interview transcripts could further qualitative analysis options for those concerned with assessment of university provided resources in general, or, for more targeted resource assessment and evaluation by reducing the amount of time spent analyzing transcripts. Although the data used for this case study focuses on open access, author attitudes and processes, and

\(^1\) TAPOR, available at http://taporware.ualberta.ca/
library provided resources, the goal of this study is to test the reliability of preliminary
textual analysis methods, not to provide significant findings related to author behavior.

Academic librarians are increasingly shouldered with the responsibility of
navigating the field of scholarly communications and maintaining awareness of new
trends that emerge in the field of open access. In light of this, the data used to test our
method is timely and relevant. Understanding faculty and author perspectives and
opinions on open access publishing is an example of information that could be valuable
in reinterpreting academic library service models.
Literature Review

A 2010 study by Jingfeng Xia compiled data to examine faculty attitudes towards open access publishing. Xia used normalized questions from multiple studies and created constructs related to the open access publishing process. The three constructs that were created examined awareness, action, and attitude of authors towards open access publishing and open access journals. This longitudinal study applied time series analysis to a collection of data in order to provide a snapshot view of the evolution of attitudes towards open access during a key time frame in the development of the open access movement, 1990-2010. Naturally, findings reflect an increase in open access publications; however, the aim of the study was to provide open access (OA) advocates with information that would inform them of concerns that were prevalent among researchers, despite increasing use of open access publishing. It was expected that this information would assist advocates in addressing these concerns in future open access advocacy campaigns.

Our study reflects a similar goal in testing a methodological framework for data analysis at an aggregate level. Xia notes “to synthes[ze] and normalize[e] across surveys is difficult… however, sacrifices like this have been alleviated by an analysis at the aggregate level to provide a whole picture of scholars’ attitudes and behaviors (Xia, 2010, p. 622). This sentiment reflects the theoretical basis for pursuing the case study to test our methodology, which explores analysis of qualitative interviews using natural language processing tools, with a case study focusing on scholar attitudes’ towards open
access publishing. Utilizing natural language process tools for preliminary analysis of transcripts can reduce time commitment burdens on researchers by providing contextual information about interview content, allowing researchers to determine if more in depth content analysis of transcripts is warranted.

Transcript data from interviews with institutionally affiliated authors provide a starting point for testing the reliability of utilizing automatic textual analysis as a precursor to human evaluation. There are two areas of primary importance to this case study: open access and scholarly communications. Research on open access and scholarly communications has significantly increased since 2000, and generally provides a wide range of perspectives on the topic, including those from scholars, librarians, and publishers. This is an area of high interest among libraries, librarians, and researchers in general, with new research continually being produced. Natural language processing is a more established field, and has research dating to the 1950’s when Alan Turing developed the Turing Test. The Turing Test explores human-computer interactions and the ability of a machine to impersonate a human in conversation. Natural language processing (NLP) covers a wide range of technologies, many of which focus on the personification of machines, and their ability to recreate human language and reasoning, but for the purpose of this study we will only provide a limited historical framework of the NLP field as it intersects with textual analysis.

**Open Access and the Changing Role of Libraries**

The term "open access" has grown to include any publication that is freely available on the web. There are often distribution limitations put into effect, or authors’ fees required to help cover the cost of distribution. Open access is an international
movement, led by librarians, scholars, and students, and could be considered a grassroots initiative. Research examining the use of open access as an alternative, or complement to, traditional publication methods is plentiful, and can largely be divided into broad areas of interest. These areas include examination of library responses to open access publishing patterns, including changes in collection development; librarians and their role within the open-access movement as agents of communication; and the field of scholarly communications, which is increasingly overlapping with librarianship. In addition to exploring the intersection of libraries and open access, there is a more specific focus that can be taken with respect to author and scholar attitudes’ towards open access publishing.

Subscriptions and submissions to open access journals continue to increase. This is largely due to a desire to widen academic readership and to provide immediate access to publications to not only members of the academic community, but also to those who are unaffiliated with an institution of higher learning.

Scholarly communication is an umbrella term for a complex array of related issues, including authors' rights, copyright, access to information, peer review, and publishing, all of which have a direct impact on libraries, universities, and faculty. Driven by years of journal price increases, dwindling serials budgets, and the potential for new distribution channels… academic libraries are responding by creating scholarly communication programs. (Courtois & Turtle, 2008, p. 160)

Among studies examining the nature of the library and the librarian’s role within the open access movement there are very few studies examining the correlation between author attitudes’ towards open access and how those are affected by institutionally provided resources. Current studies primarily focus on the changing role of the librarian to include not only management of publication budgets and subscriptions as they relate to collection development, but also discuss the librarian as advocate for more sustainable and reliable methods of publication harvesting and institutionally-driven changes relating
to journal subscription practices and institutional repository utilization. Librarians often work with scholars to assist and encourage them to deposit their work in institutional repositories. In addition, they manage subscriptions based on budget allocations, and advocate for subscriptions and publication harvesting methods that will provide users with long-term, reliable access to information at fiscally sustainable prices.

Studies focusing on author attitudes include those that are survey based, although rarely do they inquire about the nature of scholar interactions with the library. A 2009 UK based study by Sally Morris and Sue Thorn comprehensively evaluated attitudes towards open access as a format, and focused on decision-making processes related to publication venue. However, of 1,400 participants, there were merely 8 mentions of the library, and they were add-in comments. The commenters noted that they expected a change in the libraries as they relate to the publishing and scholarly communication system, in that they expected libraries to evolve in the future from providing e-only content to “ceasing to exist” (Morris & Thorn, 2009, p. 229). Ian Rowlands and David Nicholas, also UK based researchers, conducted similar survey-based studies that reported on faculty authors’ attitudes towards open access in 2005 and 2006, respectively. Their studies found a significant disconnect among researchers understanding of cost absorption by libraries. Participants in their studies appreciated the fast access to information that digital content provided them, but were generally unaware of the libraries role in providing this access, and were strongly opposed to any author pay models associated with open access publishing (Rowlands & Nicholas, 2006).

In institutions where there are strong researcher-librarian ties, either via a scholarly communications officer, or through librarians embedded in academic
departments, collaborations may include conversations about open access publishing, institutional repository development, subsidized author fees through institutional membership, as well as education about author rights and copyright. These interactions are changing the nature of the scholarly communication landscape as it relates to the role of libraries and librarians. This is evidenced through organizations that support open access initiatives and have a strong librarian presence, such as SPARC, the Scholarly Publishing and Resource Coalition (www.arl.org/sparc), which was developed by the Association of Research Libraries (ARL), and provides information about international open access initiatives. Self-publication, blogging, open access publishing, and storage of publications on personal websites have increased the need for librarians and those involved in the field of scholarly communication to inform researchers about their rights and options, as well as any legal responsibilities of which they should be aware prior to sharing their publications with the larger academic community. Organizations such as SPARC assist libraries in providing these services.

Each stage of the OA movement engages researchers in a decision-making process. Decisions made by researchers will impact the development of OA. This is particularly critical at the stage of OA content accumulation through self-archiving of digital items. Unless a mandate policy is enforced, self-archiving in institutional repositories is a voluntary task for researchers. (Xia, 2011, p. 7)

Access to academic research has historically been limited to a very small population of individuals who pay for access to journals through institutional affiliations as researchers or students, and is now accessible to a much larger constituency due not only to open access initiatives, but also to mandated access policies. The National Institutes of Health (NIH) implemented a public access policy in 2008 for all publications receiving NIH funding (publicaccess.nih.gov), and the National Science Foundation
(NSF) began requiring that all grant funded projects have a data management plan, effective in early 2011 (http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928). Both of these policies have impacted the workflow of academic librarians.

“The integration of open access materials into normal ongoing library operations requires, as other Internet resources do, additional staff time and effort, even though the materials themselves are free. Libraries are no longer simply consumers of scholarly information. A growing number of libraries have become digital publishers, primarily offering free/open access journals and institutional repositories” (Bailey, 2008, p. 371).

Changing attitudes towards transparency in research have resulted in changes to data management by scholars, and therefore to libraries and librarians in respect to their responsibilities to provide appropriate support in this arena. “Librarians remain caught in the maze of complexities and frustrations inherent in the current system and proposed resolutions. However, as highly knowledgeable mediators of the scholarly communication world, librarians are uniquely positioned to further the success of any possible solution” (Albert, 2006, p. 259).

Author attitudes surrounding open access publishing are largely unexamined within the scope of the academic library, but are invaluable sources of information for their ability to shed light on methods and processes that authors utilize in choosing a publication venue. Awareness of author workflows could provide library stakeholders with valuable data for analysis of services provided; however, analyzing lengthy interview transcripts can be burdensome to full-time librarians who have service and collection development responsibilities. Qualitative analysis of interviews could be utilized as one method of obtaining information about author needs as they relate to the library. “The aim of discourse analysis is not only to identify interpretative repertoires, but to point out the power and influence of particular narratives and to analyze their
potential societal and institutional functions and effects” (Talja, 1999, p. 474). The burden of discourse analysis could be alleviated by the utilization of natural language processing and automatic textual analysis tools.

The length and density of qualitative interviews has made them an unpopular choice for service assessment in the academic library environment, where librarians often utilize online or in-person surveys to obtain information about user needs. There are many options for using natural language processing tools to increase the speed with which researchers can engage with lengthy transcripts. NLP tools can automatically summarize conversations, providing a concise presentation of content; provide word-use statistics; and provide information about co-occurring terms, which may indicate areas of topical significance. Although there are many additional NLP capabilities, we will focus on these tools during our case study.

A recent study conducted by Sujin Kim with the University of Kentucky in 2009 examined cancer patient blogs utilizing the TAPOR (http://taporware.ualberta.ca/) tool to compare user-defined and software-generated tags. Although their study utilized both SPSS and TAPOR, where TAPOR was used for list word distribution and SPSS for co-occurrence (262), a review of their methods for content analysis provided inspiration and structure for the analysis of interview transcripts in this study. Their study utilized 485 blog posts that were mined for subject concepts and blogger-defined tags, and assigned software tags using standard vocabularies. Their research questions were broad, such as “what are the descriptive characteristics of blog posts”, which were answered by a simple frequency analysis.
TAPOR (Texual Analysis Portal for Research) was created by researchers at McMaster University Faculty of the Humanities and is a set of textual analysis tools that allows users to choose from several options for analyzing text documents. TAPOR can be utilized by non-experts, and can provide information on co-occurring patterns, compare word distribution, display graphical depictions of results—such as word clouds, and provide general statistical information about text, such as word counts across the document as a whole. The tool allows users to choose from HTML, plain text, and XML tool options. The variety of tools TAPOR provides allow for the deconstruction of multiple hypotheses by exposing larger thematic elements present in documents as revealed through frequency analysis, co-occurrence, and key word finders.

Lightweight textual processing tools can be valuable for exposing the main topics in lengthy text documents. NLP summarization tools can present patterns, action items, and deliverables from meetings, in addition to areas of speaker-to-speaker and speaker-to-topic tension. Discourse analysis at the speaker level allows summarization tool users to extract the essence of dialogue as opposed to only topics and associated speakers. This occurs through modeling, classification, and prediction of dialogue acts, as well as the associated speakers. This analysis output can expose the relationships that develop during the course of a conversation. Analyzing conversations, speaker-to-speaker tendencies, and individual speaker tone, rhythm, and pitch, has evolved significantly during the last twenty years. Between 1990 and 2010 research on the topic of natural language technologies and conversational analysis, such as one may experience in a meeting, a one-to-one conversation, or as is presented in a corpus containing extracted conversation sequences, such as those that from interview transcripts, has evolved from a linguistics
based focus to a deeper examination of human personality as it is uncovered in subjective ways.

**Early NLP Efforts**

Early efforts to analyze and classify conversational dialogue utilized dialogue act (DAs) models. A dialogue act “can be thought of as a tag set that classifies utterances according to a combination of pragmatic, semantic, and syntactic criteria” (Stolcke et al., 2000, p. 340). Parsing through these conversational units such as greeting, statement, question, backchannel (non-speaker acknowledgement of understanding), agreement, and disagreement allow for further investigation into the nature of control shifts, test prediction, investigation of personality characteristics (introvert, extrovert), and analysis of topic change dynamics as executed during a conversation.

Gaining a deeper understanding of the characteristics of these conversational elements will lead to the creation of machine agents that engage in human dialogue in ways that allow for human-machine reasoning during the course of discussions. In the 1990 piece, *Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation*, Walker et al. conclude that despite best efforts to predict discourse based on control shift “there are other levels of structure in discourse that are not captured by the control rules, e.g. control shifts do not always correspond with task boundaries. There can be topic shifts without change of initiation, change of control without a topic shift” (Walker, 1990, p.76).

Dialogue act modeling in the late 1990’s and early 2000’s focused on automatic tagging and labeling of discourse structure, which “indicates how the utterances of a discourse group together into segments and how those segments are related to each other”
This modeling was largely motivated by a desire to develop trustworthy summarization tools and assist in the refinement of human-computer dialogue systems. Prosodic analysis, or the examination of pitch, tempo, and stress in speech, is a strong element in both the 1998 study *Dialog Act Modeling for Conversational Speech* by Stolcke and Shriberg and a 2000 study in which Stolcke also participated. Prosody will remain an important part of the dialogue surrounding conversational analysis for the following ten years.

**Deeper Investigation of Prosody and Sentiment**

The segmentation of meeting topics builds directly on participant involvement and is encouraging for those hoping to find a more complete model for analyzing the structure of meetings from both a speaker based and task or topic based perspective to gain deeper insight into the essential nature of conversations as they occurred. Stolckes’ investigation into prosodic features and meeting participation and involvement led Wrede and Shriberg to examine the functionality of utterances in their 2003 study of meeting hot spots. A meeting hot spot can include areas of high back and forth discourse, disagreement, or the generation of important information in a condensed period of time. The context of DAs is revisited in this study in respect to speaker identity and meeting type. “Involvement categories (non-involved, disagreeing, amused, and other) show statistically significant associations with independently annotated DAs. While some results were expected, such as the correlation between jokes and amusement, or backchannels and non-involvement, others were surprising … and increased or decreased involvement accordingly” (Wrede & Shriberg, 2003, p. 184).
“Silences, overlaps, and speaker changes are indicative of topic shifts” (Galley & McKeown, 2003, p. 562). Word distribution allows for segmentation of meeting transcripts into topically relevant chunks. Analyzing the correlation between words and topic boundaries—the number of word occurrences near a boundary, and the number of appearances overall, lead to deeper insight into the inherent nature of dialog. “Gaps are silences not attributable to any party, a silence that follows a question or in the middle of somebody’s speech [is] a pause and any other silences a gap…the beginning of segments are characterized by having little overlapping speech, and [there is] sometimes a correlation between topic boundaries and sudden changes in speaker activity” (Galley & McKeown, 2003, p. 567).

Speaker agreement and disagreement is examined to aid in the advancement of automatic summarization efforts and the generation of increasingly competent machine translators that could report accurately on events during the course of a discussion that lead to a decision. As with the work of Galley & McKeown, a 2003 study by Hillard, Ostendorf, and Shriberg seeks to represent the essence of conversation by focusing on areas of tension and controversial decisions, and correctly representing action items. Focusing on prosodic features as categorized into negative, positive, and backchannel leads to a deeper understanding of social dynamics and cultural constructs present during conversation and represented in the summary.

Galley and McKeown research agreement and disagreement in conversational speech using Bayesian networks to model dependencies. As with the 2003 work of Hillard, Ostendorf, and Shriberg, the focus is on advancing the application of automatic summarization and to further pattern recognition during multi-participant discourse. As
opposed to focusing on the unique context in which a word was used—such as an area of tension or controversy, they utilized adjacency pair speech acts which “consist of two parts that are ordered, adjacent, and produced by different speakers, to identify the target of an agreeing or disagreeing utterance,” and given the second element of an adjacency pair, they can determine who is the speaker of the first element (Galley et al., 2004). In order to classify speakers as either the A or B speaker, and therefore designate which utterance was attributable to that speaker, they combined DA tagging with adjacency pair labeling.

Extending participant speech act identification into agreement and disagreement classification was achieved through the integration of features that were described as “good predictors of agreement: utterance length distinguishes agreement from disagreement, duration is also a good predictor of backchannels, since they tend to be quite short…a fair amount of silence and filled pauses is sometimes an indicator of disagreement” (Galley et al., 2004). This research was crucial in furthering conversational research in that it was the first to focus heavily on assigning DAs to certain speakers, therefore determining which are the most relevant in the scope of the meeting dialogue. Further work on the role utterances play in actionable item detection and meeting summarization can be seen in the 2006 work by Purver, Ehlen, and Niekrasz: *Detecting Action Items in Multi-party Meetings: Annotation and Initial Experiments.*

Addressing a lack of visual output for dialogue analysis, *ChAT: A Time Linked System for Conversational Analysis* (Gregory, et al., 2006) provides a graphically rich user interface that represents an annotated dialogue with named entities, participant
information, affect (agree, disagree, positive, negative), and topic segmentation after data ingest. The ChAT research is notable for its desire to advance conversational analysis with a visual representation that efficiently reduces required user time for conversational analysis and complete extraction of pertinent information. “The components can be run independent of the UI and in batch, resulting in an xml document containing the original transcripts and the metadata added by the processing components. This functionality allows the data to be manipulated by traditional text mining techniques, or to be viewed in any other visualization.” (Gregory, Love, Rose, & Schur, 2006, p. 56). Their approach is similar to other studies in that it aims to create relationships between speakers and topics, and present them in concise, clearly tagged formats.

There is a marked continuation of focus on the inherent nature of conversation and the characterization of speaker tendencies and speech acts—which can be used to directly alter the course of discussion in a meeting. Distracting others from what was an important piece of dialogue in order to gain traction from other speakers is a direct example of conversational manipulation. Mining of “speaker/author affiliations, positions within a social hierarchy…and the integration of even very limited information regarding inter-document relationship can significantly increase the accuracy of support/opposition classification” (Thomas, Pang, & Lee, 2006, p. 334). In a 2007 study by Somasundaran, Ruppenhofer, and Wiebe, adjacency pairs are examined as complementary features in examination of meetings and their associated dialogue structure. Their work extends the 2003 work of Wrede and Shriberg to “annotate the mental states of participants in meetings or interviews on the basis of multi-modal data, and they approach adjacency
pair information as detectors of agreement and disagreement amongst participants much as Galley et al. do in 2004” (Somasundaran et al., 2007, p. 32).

Extending the personal nature of conversational analysis to include interaction style, as a product of both personality recognition and sentiment analysis is reflected in *Extracting Social Meaning: Identifying Interactional Style in Spoken Conversation* by Jurafsky, Ranganath, and McFarland in 2009. Detection of flirtatiousness, awkwardness, and friendliness is certainly more complex in nature than agreeable versus disagreeable, much as noting characteristics of speakers with depression and charisma and the effects of those characteristics on conversational style is more complex than simply identifying extraversion and introversion. This study also aimed to advance the conversational abilities of natural dialogue agents.

More recently textual analysis and the use of natural language processing tools has turned attention towards sentiment analysis and topic detection in corpora containing short, informal discourse such as that found in blogs or on Twitter. Although Twitter events, or “tweets” can be considered to be singular events they are often a string of dialogue composed of several tweets directed at an individual and can therefore be classified as conversations ripe for mining and sentiment analysis. The use of SentiStrength was deemed successful in detecting overall sentiment of expressed terms, (Thelwall, Buckley, Palthoglou, & Cai, 2010) as it is designed to not only classify, but also measure the strength of positive or negative sentiments. Future research in this area will be heavily focused on large-scale classification of sentiment.

It is expected that this area of research will continue to refine speaker-to-speaker analysis, highlight individual personality characteristics, prosodic tendencies, and the
association of these features to topic segmentation and dialogue act predictions. Furthering this area of research will lead to increasingly refined techniques that can be applied to discourse analysis to the extent that a non-present party can view output and not only quickly gain insight into actionable items and decisions, but also to the nuanced areas of tension between individuals, or participant feelings towards topics and speakers, leading non-present parties to an area of understanding that would have previously only been accomplished by listening to audio recordings. With the refinement of discourse analysis and machine training the burden of analysis could ideally be transferred entirely to the machine, rather than being a shared responsibility between human and machine. Discourse analysis is still largely reliant on human interpretive capabilities such that analyzing interview transcripts or blog postings cannot yet be completely transferred to a machine. There are opportunities for more sophisticated analysis, such as predictive capabilities, which could have large social implications, particularly in the arena of politics and observations of larger socio-cultural trends.

Natural Language Processing, Scholarly Communication, and Content Analysis

Combining natural language processing tools with traditional data collection and analysis methods could strengthen the field of scholarly communications and, within the scope of this case study, increase awareness of sentiment towards open access publishing. Transcript analysis has traditionally involved lengthy coding sessions. While this may prove beneficial in more traditional social science research projects that are not limited by publication pressures or staff time, for the purpose of a librarian wanting to quickly determine the effectiveness of services offered this presents an alternate method for obtaining information. Qualitative interviews and discussions surrounding open access
publishing are plentiful, however, research combining sentiment analysis, natural
language processing tagging, summarization, or classification technologies and
ethnographic fieldwork are limited, reflecting a need to explore the integration of these
research areas.

Despite a wide-ranging discussion on the changing role of libraries and librarians
in respect to open access, there is still a limited exploration of the intersection of open
access, scholarly communications, and the utilization of automatic textual analysis as a
viable method for content analysis. White and Marsh, in discussing qualitative and
quantitative data analysis in LIS studies remark that:

Analysis is integrated into coding much more in qualitative content analysis than
in quantitative content analysis. The emphasis is always on answering the
research questions but considering as well any transformations that the initial
foreshadowing questions may have undergone during the coding or any new
questions or themes that emerge during the coding. Often the result of qualitative
analysis is a composite picture of the phenomenon being studied. The picture
carefully incorporates the context, including the population, the situation(s), and
the theoretical construct. The goal is to depict the “big picture” of a given subject,
displaying conceptual depth through thoughtful arrangement of a wealth of
detailed observations. (Marsh & White, 2006, p.39)

In the spirit of this definition of qualitative analysis as providing an image of
collected data it would seem appropriate to explore natural language processing tools as a
potentially efficient pathway to gleaning topical and thematic information from a corpus
as a preliminary step to human analysis and coding. This could result in reducing barriers
to utilizing interview methods as a rich source of information for researchers within the
academic library environment, and increase the ability of libraries to engage in this type
of research in order to assess the services they provide. Ethnographic interviewing can be
a lengthy and involved process, which presents yet another reason to utilize automatic
text processing. Uncovering unexpected patterns early in the research process may
provide sufficient evidence to warrant reformulation of the initial research questions or main study focus before investing large amounts of time and energy to the project.

The goal of this study is to test the viability of utilizing a textual analysis tool prior to engaging in detailed human coding and analysis. We propose a framework for utilizing textual analysis processing on interview transcripts. In order to test the framework, we utilized data from a previous study involving ethnographic interviewing techniques, which we will refer to as the Health Sciences Library (HSL) study.
Methods

In order to assess the viability of utilizing automatic text processing tools as a precursor to human analysis we tested a methodological framework of processes for investigating interview transcripts. The methods utilized in this study are loosely based on those examined in the cancer blogs research by Sujin Kim, and also integrate strategies offered through the TAPOR portal that encourage the use of multiple tools to uncover larger themes and areas of interest in a corpus.

Figure 1
Outline of Framework Utilized During the HSL Case Study Analysis

The goal of using TAPOR is not to answer research questions generated in preliminary analysis, but rather to assess the potential of the information presented in transcripts to address research questions without researchers having to engage in lengthy coding
sessions in order to reveal this information. The output from TAPOR analysis will be used to generate a summary of findings that could be presented to an administrator in order to determine if the transcripts should be subject to further analysis. In order to validate our findings and initial conclusions regarding the potential value of utilizing a simple natural language processing tool such as TAPOR we invited an outside reader who is an expert in the field of scholarly communications to provide input on the study and suggest ways that the executive summary could be improved or reformulated to better serve the interests of an academic institution.

The data used to test our framework were generated from interviews conducted as part of IRB study 10-0916 (Edwards, et al., 2010), or, as we have termed it, the Health Sciences Library (HSL) study. This study included 30-minute, face-to-face semi-structured interviews with researchers to discuss their publication practices. Participants were asked to recollect a recent project that led to the publication of an article and respond to queries about the evolution of the final product and decisions regarding final distribution. Examples of specific interview queries can be viewed in Appendix A. Transcriptions of audio-recorded interview sessions were completed by the interviewing researcher and stored on a password-protected site.

Potential interview participants were identified through purposive sampling measures based on their academic department and were all located at a major research university and identified as having recent publication activity as evidenced through database searches including Ulrich’s and ISI Web of Science, and through departmental and individual faculty/researcher websites available on the web. Participants responded to an email invitation to participate in a study examining author-fee subsidy programs and
their effects on authors’ publishing behaviors. The cohort of researchers and faculty who participated in interviews primarily represented the biomedical and health sciences fields. All data was de-identified in order to maintain anonymity.

In a 1999 study designed to assess users conception of the library, Sanna Talja remarked on the theoretical basis for the discourse analysis method:

the researcher abandons the assumption that there is only one truly accurate version of participants’ action and belief. Interview talk is, by nature, interpretation work concerning the topic in question. It is reflexive, theoretical, contextual, and textual, because the objects of talk (e.g., the library) are not abstract, ideal entities everyone sees in the same way. When talking about the library, participants do not solely produce a neutral description and express their opinion. They produce a version of the library… (Talja, 1999, p. 464)

Similarly, participants whose transcripts were utilized for this study produced a version of events, as best they were able to recall, leading to the publication of a journal article. As such, our framework assumes all analysis will be conducted in an ethnographic, iterative, and inductive way.

After reviewing the transcribed HSL interviews it was noted that as individual authors spoke about open access publishing they generally expressed a willingness to consider open access a viable option when assessing final publication venues. The nature of discussion prompts from the HSL study, as viewed in Appendix A, and conversations from the interviews led to the decision that within the confines of this case study open access should be a central focus of analysis. This conceptual overview led us into the process of developing our methodology for utilizing TAPOR. On a metacognitive level this began by developing a better understanding of the major concepts discussed in the interviews. For example, because some authors expressed a willingness to consider open access publishing after engaging in interview discussions one could hypothesize that collegial influence may factor into the decision-making process surrounding open access
publishing. A conceptual cluster, similar to what one would develop by drawing a simple mind map, was thus generated to include terms that could be considered descriptors of academic community such as peer, colleague, or collaborator.

Generating clusters ensures more complete coverage of a concept in transcript analysis. We would be misguided to conduct co-occurrence analysis on only one term, and expect that individuals discussing a concept would always use only that term to describe a concept. We are increasing the likelihood of complete concept coverage in co-occurrence analysis by selecting multiple terms and synonyms to represent natural language variations on concept expression by assuming that interview participants may use one of any number of terms to describe a concept in their responses.

Concept cluster creation is a multi-step process. In our case study we generated one major concept, open access, through human reading of the transcripts. Additional concept clusters were generated through an inductive process of comparing WordNet search results against frequency analysis output, and integrating prior knowledge of interview content. WordNet is a lexical database developed and maintained by Princeton University.

Although similar to a thesaurus, “in that it groups words together based on their meanings, …WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity. (wordnet.princeton.edu, 2011)

The investigator's prior knowledge of the interviews informed concept term generation, with some supporting terms discovered deductively, working from the concept term down and exploring terms suggested by WordNet. Concept clusters were deemed
complete after multiple sister term searches resulted in a high saturation return. The utilization of WordNet was added as a means of formalizing and verifying concept term coverage for the scope of our case study. A complete list of concept clusters that were utilized in framing this research can be seen in Table 3.

In our case study we generated concept clusters after a thorough reading of the HSL transcripts, and after utilizing TAPOR for word frequency analysis. We then utilized the keyword finder and co-occurrence tools. Figure 1 provides a model of the entire analysis process used for this case study. The order in which tools are used is not crucial to the success of utilizing TAPOR for preliminary transcript analysis. The order of steps presented in our model was determined to be the most efficient process based on several external factors, including the researcher’s general knowledge of the research area and familiarity with the original interview content. It should be noted that the process of generating concept clusters and engaging with TAPOR will vary slightly from study to study based on the researcher, the structural framework-particularly the institution the study is situated in, and other outlying factors not limited to, but including; time, funding, and data manipulation and technological comfort of the party who is responsible for analysis, that are unique to each study.

Generating research questions prior to utilizing TAPOR will allow researchers to determine which of the tools available via TAPOR would be most beneficial to their area of inquiry. After reading the HSL transcripts we generated the following research questions for our case study:

1. What are the main themes or areas of interest discussed during the interviews?
2. Of concepts identified through textual analysis, which have the strongest correlation with open access?

3. Is there a correlation between time to publication (ie: speed) and consideration of open access?

4. Is there a correlation between impact factor and open access publishing?

5. Is there a correlation between funding and open access publishing?

6. Is there a correlation between collegial attitudes towards open access publishing and consideration of open access?

7. Does discussion of the library occur frequently, infrequently, or not at all?

As suggested above, for our case study we generated research questions prior to the identification of major themes and concepts. Generating research questions prior to engaging with textual analysis tools will serve to narrow the focus of potential information pathways, and reduce human energies spent utilizing the toolkit. Entering into data analysis without at least some high-level questions is possible, but would not represent the best possible path to finding the information that is of most interest to the researcher. Developing a framework through research question generation will better facilitate the discovery of dialogue extracts that could provide the most revealing information about the dialogue as a whole, and could provide additional information to assist in the decision-making process regarding the potential value of deeper human analysis. In an effort to produce a framework that represents the most efficient methods possible by reducing duplication of effort, all of the original transcripts from the HSL study were merged into one text (.txt) file. Merging the transcripts into one file included
the removal of personal identifiers. In addition, all dialogue not attributed to interviewee responses was removed, resulting in a corpus of twenty interviewee responses (n=20).

Utilizing one text file eliminates the need to replicate textual analysis on each individual transcript. Combining transcripts was optimal for this case study as our interview participants represented a fairly cohesive group of individuals. In addition, because our cohort was small (n=20) this was the most efficient method for transcript management; however, if there were a larger selection of interview transcripts from a group of participants that could be more appropriately divided, such as by department or discipline, we would have likely separated the transcripts into groupings in order to make generalizations from the transcript data about groups of participants as a whole.

Three TAPOR tools were used for textual analysis in our case study; list words, keyword finder, and co-occurrence. By highlighting the statistical properties of text, we are able to see what terms or concepts are discussed, thereby illuminating the nature of discourse without reading, in entirety, all of the interview transcripts. The list words tool provides a list of all terms occurring in the corpus. Frequency analysis provides complete counts of all terms in the corpus, and was conducted both including and excluding stop words. Stop words are the most commonly used words in many human discussions and include words such as “and”, “or”, and “the”.

In textual analysis stop words are often considered noise, as they contribute little to the overall understanding of the nature of a discussion. Standardized stop word lists have been created for general use, and are often utilized in NLP analysis. For this case study the Glasgow stop word list was augmented by the inclusion of several terms that were diluting the frequency analysis on initial data runs. Augmenting a standardized stop
word list, such as Glasgow, is standard practice in order to remove “noise words to be ignored, however, as it is empiric and subjective, it is also biased and can be misleading. This is why stop word lists need to be understood and used with care” (Blanchard, 2007, p. 311). In our case study terms such as “like” and “yeah” are examples of the types of words that were added to the original Glasgow list. These terms contribute little to understanding the overall content of discourse, but since they are common in spontaneous speech, were occurring with high enough frequency that they could dilute later analysis conducted using the key word finder or co-occurrence tools. The results of frequency analysis can be found in Table 1.

Following frequency analysis and concept term generation with WordNet, the keyword tool was utilized to determine if there were any significant concepts that would be uncovered that were not revealed during frequency analysis. The keyword tool “tries to find the possible keywords or key phrases of a source text and recommend them to a user. The principle of this tool is based on the assumption that keywords or key phrases will occur more frequently in the text than other words or phrases” (TAPOR, 2011). In our case study we might expect to see such phrases as “open access journals” with high frequency. Results of keyword analysis can be found in Table 2.

Concept clusters informed the creation of a co-occurrence matrix. The co-occurrence matrix includes the key areas of interest to the case study as determined through concept cluster creation: time, money, speed, community, colleague, publish, and impact were considered primary terms for the major concepts in our case study, and were the main terms utilized during co-occurrence analysis. Additional terms used in the matrix were selected on the basis of their occurrence frequency in the text, and their
potential to address the seven research questions generated for this case study by virtue of their contextual significance. These terms can be viewed on the left side of the co-occurrence matrix, and are much less specific with respect to their significance in the case study.

Co-occurrence analysis can be conducted with varying context lengths, typically between 1 and 15, and can be conducted in various context types, including a word, sentence, line, or paragraph. Co-occurrence analysis could be considered an extension of basic key word analysis, and is conducted by selecting the primary term and an associated term for which the co-occurrence count is desired. TAPOR generates results based on the number of times the two terms occur within the user specified context. For example, because our case study focused on open access if we select “open” as the primary term and “access” as the second term, we can expect a high co-occurrence count. By reviewing co-occurring terms in a corpus we can gain a deeper understanding of relationships between concepts and terms, and potentially discover new concepts or relationships that were unexpected. TAPOR highlights the span of text in which terms co-occur, allowing users to quickly identify key phrases. In our case study we included these highlighted phrases as a part of our executive summary to the outside reader.

The goal of utilizing TAPOR was not to answer major research questions, but rather to identify topics discussed in interviews and determine if the information in the transcripts could begin to address research questions. Following data analysis we generated a one page executive report with a summary of findings. The report was created as an example of the type of report one might expect to give to a senior administrator in order to assist them in determining if preliminary findings warrant
continued exploration of data through human analysis and coding with tools such as SPSS or NVIVO based on the data’s potential to answer research questions. This report can be viewed in Appendix B.

The outside reader utilized in our case study is an expert in the field of scholarly communications and holds the J.D., the Master’s of Science in Library Science (M.S.L.S), the M.A. with a concentration in Law, Expression & New Technologies, and the B.A. in Constitutional History and Dramatic Art. The outside reader’s goal was to determine if the findings presented in the summary were rich enough to provide ample information to begin to address the research questions, and determine if a summary of this type would be useful in the scope of an academic library. In addition, the outside reader was asked if this report and method could be considered valuable to the extent that sufficient information was provided in the summary to allow an informed decision regarding future analysis of transcripts. He was asked to provide this information based on previous experiences with academic library studies, his knowledge of the working environment of research libraries, and his expert knowledge in the field.

For the scope of our case study the outside reader acted as a surrogate to a senior library administrator. In practice, if a group chose to use this tool it is likely that the administrator receiving the executive summary would be knowledgeable about the subject matter, as such we made the following assumptions when creating our summary:

1. The administrator receiving a report such as the executive summary included in Appendix B would have prior knowledge of the original study that generated transcripts.
2. The administrator would likely have asked for a summary of findings.
3. The administrator would be knowledgeable about the topics discussed during the interviews but would not have read any of the interview transcripts, as such the executive summary could be considered the first major report providing information about the transcripts.

The outside reader, Will Cross, received directions on his goals within the scope of our study via email. Directions were not prescriptive in nature, but rather provided a general frame of reference and the assumptions that could be made while conducting a review of the executive summary. Feedback was requested regarding the potential usefulness of this method as well as regarding general aesthetics of the report, including areas that could benefit from deeper explanation, or graphics that may provide additional insight into the nature of results. This feedback was requested as a measure of the readability of findings by an individual who may be unfamiliar with textual analysis tools and processing methods, and as a way to assess the overall potential value for utilizing this method in the scope of an academic library environment where an administrator may need to make a decision regarding further research in a short period of time.

Following input from the outside reader discussions and limitations of the case study were recorded and are reported in the discussions section along with what are perceived to be study limitations and benefits. This section also reports on potential extensions of this case study.
Results

The corpus contained 3,137 unique words other than those in the stop word list. There was a total word count of 66,448, reduced to 20,784 after removing stop words, resulting in an average word count per transcript of 3,322 words. The average transcript length, including interviewer prompts and interviewee response was 4,251 words, with a range of 1,350 to 9,840 words. Table 1 reports frequency counts for the top 10% of unique words not including those in the stop word list, and utilizing an inflectional stemmer.

Table 1

Top 10% of Unique Words, by Frequency of Occurrence.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Frequency</th>
<th>Rank</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>journal</td>
<td>713</td>
<td>21</td>
<td>different</td>
<td>97</td>
</tr>
<tr>
<td>2</td>
<td>think</td>
<td>595</td>
<td>22</td>
<td>read</td>
<td>93</td>
</tr>
<tr>
<td>3</td>
<td>publish</td>
<td>418</td>
<td>23</td>
<td>grant</td>
<td>92</td>
</tr>
<tr>
<td>4</td>
<td>access</td>
<td>321</td>
<td>24</td>
<td>sort</td>
<td>83</td>
</tr>
<tr>
<td>5</td>
<td>people</td>
<td>314</td>
<td>25</td>
<td>interest</td>
<td>83</td>
</tr>
<tr>
<td>6</td>
<td>paper</td>
<td>287</td>
<td>26</td>
<td>mean</td>
<td>82</td>
</tr>
<tr>
<td>7</td>
<td>open</td>
<td>281</td>
<td>27</td>
<td>little</td>
<td>81</td>
</tr>
<tr>
<td>8</td>
<td>kind</td>
<td>260</td>
<td>28</td>
<td>fund</td>
<td>81</td>
</tr>
<tr>
<td>9</td>
<td>work</td>
<td>169</td>
<td>29</td>
<td>student</td>
<td>80</td>
</tr>
<tr>
<td>10</td>
<td>time</td>
<td>158</td>
<td>30</td>
<td>plos</td>
<td>79</td>
</tr>
<tr>
<td>11</td>
<td>article</td>
<td>156</td>
<td>31</td>
<td>factor</td>
<td>78</td>
</tr>
<tr>
<td>12</td>
<td>review</td>
<td>125</td>
<td>32</td>
<td>particular</td>
<td>67</td>
</tr>
<tr>
<td>13</td>
<td>science</td>
<td>121</td>
<td>33</td>
<td>cost</td>
<td>66</td>
</tr>
<tr>
<td>14</td>
<td>year</td>
<td>116</td>
<td>34</td>
<td>high</td>
<td>66</td>
</tr>
<tr>
<td>15</td>
<td>research</td>
<td>112</td>
<td>35</td>
<td>start</td>
<td>66</td>
</tr>
<tr>
<td>16</td>
<td>publication</td>
<td>107</td>
<td>36</td>
<td>field</td>
<td>64</td>
</tr>
<tr>
<td>17</td>
<td>data</td>
<td>105</td>
<td>37</td>
<td>study</td>
<td>64</td>
</tr>
<tr>
<td>18</td>
<td>impact</td>
<td>104</td>
<td>38</td>
<td>figure</td>
<td>63</td>
</tr>
<tr>
<td>19</td>
<td>pay</td>
<td>104</td>
<td>39</td>
<td>library</td>
<td>61</td>
</tr>
<tr>
<td>20</td>
<td>money</td>
<td>99</td>
<td>40</td>
<td>editor</td>
<td>60</td>
</tr>
</tbody>
</table>
Table 2 reports results from the TAPOR keyword finder. We can compare findings in the keyword finder against frequency analysis data and see that 8 of the top 10 word pairs contained a frequently occurring term. The 2 word pairs not having at least one term in the most frequently occurring terms were the phrases “post doc”, and “editorial board”.

Table 2

*Key Words or Phrases as Suggested by TAPOR Keyword Finder*

<table>
<thead>
<tr>
<th>Word Pairs</th>
<th>Count</th>
<th>Word Triplets</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>open access</td>
<td>283</td>
<td>open access journal</td>
<td>30</td>
</tr>
<tr>
<td>Impact factor</td>
<td>42</td>
<td>open access publishing</td>
<td>9</td>
</tr>
<tr>
<td>Little bit</td>
<td>38</td>
<td>health services research</td>
<td>6</td>
</tr>
<tr>
<td>Access journals</td>
<td>30</td>
<td>non open access</td>
<td>6</td>
</tr>
<tr>
<td>things like</td>
<td>22</td>
<td>health sciences library</td>
<td>5</td>
</tr>
<tr>
<td>post doc</td>
<td>17</td>
<td>open access option</td>
<td>5</td>
</tr>
<tr>
<td>high impact</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>really good</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>editorial board</td>
<td>17</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Includes the 10 most frequently occurring word pairs, and the highest occurring word triplets, or phrases, across the corpus as a whole.

Concept clusters, which were used to generate a conceptual matrix for directed co-occurrence analysis can be viewed in Table 3. The terms most descriptive of the entire corpus and determined most important through contextual analysis are listed at the top of the chart, with linguistically similar terms underneath.
Table 3

*Concept Clusters that were Generated Using WordNet*

<table>
<thead>
<tr>
<th>Time</th>
<th>Money</th>
<th>Speed</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>time frame</td>
<td>fund</td>
<td>accelerate</td>
<td>factor</td>
</tr>
<tr>
<td>period</td>
<td>subsidize</td>
<td>swiftness</td>
<td>effect</td>
</tr>
<tr>
<td>duration</td>
<td>sponsor</td>
<td>quick</td>
<td>result</td>
</tr>
<tr>
<td>lifespan</td>
<td>provide</td>
<td>rapid</td>
<td>outcome</td>
</tr>
<tr>
<td>semester</td>
<td>support</td>
<td></td>
<td>by-product</td>
</tr>
<tr>
<td>year</td>
<td>price</td>
<td></td>
<td>change</td>
</tr>
<tr>
<td>interim</td>
<td>cost</td>
<td></td>
<td>influence</td>
</tr>
<tr>
<td>interval</td>
<td>value</td>
<td></td>
<td>consequence</td>
</tr>
<tr>
<td>meantime</td>
<td>charge</td>
<td></td>
<td>offspring</td>
</tr>
<tr>
<td>lead time</td>
<td>format</td>
<td></td>
<td>materialization</td>
</tr>
<tr>
<td>processing time</td>
<td></td>
<td></td>
<td>side effect</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Colleague</th>
<th>Publish</th>
<th>Community</th>
</tr>
</thead>
<tbody>
<tr>
<td>workfellow</td>
<td>release</td>
<td>department</td>
</tr>
<tr>
<td>co-worker</td>
<td>pre-print</td>
<td>center</td>
</tr>
<tr>
<td>fellow worker</td>
<td>produce</td>
<td></td>
</tr>
<tr>
<td>collaborator</td>
<td>generate</td>
<td></td>
</tr>
<tr>
<td>partner</td>
<td>print</td>
<td></td>
</tr>
<tr>
<td>affiliate</td>
<td>issue</td>
<td></td>
</tr>
<tr>
<td>adjunct</td>
<td>publication</td>
<td></td>
</tr>
<tr>
<td>associate</td>
<td>product</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Key terms were generated after review of frequency analysis output from TAPOR.

Co-occurrence analysis was conducted utilizing the conceptual matrix. Complete results from the co-occurrence analysis can be found in Table 4. Table 5 presents condensed findings representing only the concept terms deemed most significant based on range of frequency and occurrence.
Table 4

*Co-occurrence Matrix with Total Occurrence Counts from the Combined Text (.txt file)*

<table>
<thead>
<tr>
<th></th>
<th>Open</th>
<th>Access</th>
<th>Time</th>
<th>Impact</th>
<th>Library</th>
<th>Funding</th>
<th>Colleague</th>
</tr>
</thead>
<tbody>
<tr>
<td>article</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aware</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>colleague</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>costs</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fund</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>funding</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>grant</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>impact</td>
<td>6</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>interest</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>journal</td>
<td>51</td>
<td>11</td>
<td>30</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>know</td>
<td>29</td>
<td>12</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>library</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>money</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>pay</td>
<td>17</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>people</td>
<td>25</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>publication</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>publish</td>
<td>21</td>
<td>7</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>research</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>result</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>review</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>speed</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>student</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>support</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>time</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>work</td>
<td>3</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Total Count** 214 65 59 27 7 5
**Range** 0-51 0-12 0-30 0-8 0-2 0-1

*Note:* The matrix was generated using findings from TAPOR frequency analysis and concept cluster generation with WordNet.
These findings reflect some expected results, such as the high frequency return of “open access” and the frequency with which some terms occur in our corpus, such as “publish” and “journal”. From these data analysis runs we expected that several terms and key words would occur with high frequency, and that this would be reflected in each of the tools we utilized to analyze the corpus. This overlapping return confirms that the tool is presenting valid results, and also indicates that certain topics and themes are present in the corpus that may highlight valuable sentiments regarding open access publishing and author needs as they relate to the library environment.

From these findings we can begin to address the primary research question of the main study: does utilizing automatic text processing by way of a textual analysis portal prior to human analysis and coding result in findings that could lead a researcher to areas of significance in a medium sized corpus? Based on our findings, we believe that utilizing a textual analysis tool as a preliminary means to transcript analysis, such as TAPOR, combined with an executive report and a verbal summary of findings from the individual conducting transcript analysis could be valuable in reducing the amount of
time spent analyzing interview transcripts without the use of TAPOR, and could therefore be a valuable contribution to the research workflow utilized in preparing interview transcripts for analysis. This framework increases the usability of qualitative interviewing techniques within the scope of an academic library environment by reducing the potential loss of staff time on extended engagement with transcripts that may not warrant thorough qualitative analysis.

According to observations by the outside reader, the most helpful portion of the executive summary was the selected quotations, which “provide[d] some context that was useful…and g[ave] a nice snapshot of common phrases in the body of a text” (Cross, 2011). Furthermore, it was noted that the report “has value as a quick compilation of information,” however, the executive summary as provided in Appendix B could benefit from the addition of visual output, such as a simple word cloud, or other graphical representation of the analysis findings. This would serve as an additional gateway to understanding the statistical nature of the data as presented in Appendix B.

The need for additional visualizations was succinctly noted by the outside reader: “it would really be useful if [the report] could provide some good models for visualization/explanation or, even better, by offering analysis that goes beyond tallying to uncover patterns that a reader wouldn't naturally spot” (Cross, 2011). This sentiment was expanded to include the need for “striking visuals” that could be used in presentations to library committee members, or in other public presentations of findings. This context expands the environment that was considered when creating the methodology and presenting initial findings, as the report generated in our method was initially intended for internal use amongst academic library personnel only.
Discussion & Limitations

This study intended to test a methodological framework by exploring a specific case study whose research questions sought to explore broader themes that might emerge from interview transcripts as evidenced through frequency of terms, term co-occurrence, or key words identified in the corpus. Our case study focused on open access publishing and researcher decision-making processes related to academic publishing. As such we expected that certain terms would occur with high frequency, such as “open access”, however, we also wanted to determine if utilizing TAPOR would reveal any unexpected returns in the analysis, or provide helpful insight in a succinct manner. We were also interested in obtaining information from an outside reader regarding the potential functionality of this methodology and the associated executive report in the scope of an academic library environment.

Upon review of the results of TAPOR analysis it was determined that themes generally emerged from the corpus as expected. This can be illustrated through co-occurrence returns. For example, there was a high co-occurrence between “open access” and “publish”. The lack of some high co-occurrence returns yielded unexpected findings. Based on the researcher’s knowledge of the interview transcripts it was expected that there would be a high term co-occurrence between “funding” and “publication”, but results from co-occurrence analysis and frequency returns yielded no significant results, as compared to “open access” and “people”, which yielded a surprisingly high return (n=25) when compared with the overall highest term co-occurrence between “open
access” and “journal” (n=51). Although there is a weak tie between “open access” and “pay”, (n=17) and “open access” and “money” (n=9) these ties are, within the scope of our case study, conceptually different than “funding”, which may indicate a desire for money, rather than an action taken with money-such as money used to cover open access publication costs. It was also expected that collegial influence on decisions related to publication venue would be higher, as would be reflected in a high co-occurrence between some derivation of the terms “colleague” and “publish”; however, collegiality yielded the lowest overall co-occurrence count of the entire co-occurrence matrix (n=5).

A review of the findings and methodology would not be complete without a discussion of the omission of collocation analysis. The initial framework methodology included collocation analysis as a fourth step in the transcript analysis. During the creation of the framework the researcher found that results from the collocation analysis provided little insightful or expansive information. The value gained from conducting the collocation analysis was minimal compared to the amount of time that would be spent conducting comprehensive collocation analysis. The goal of this study was to establish and test the usefulness of a method for preliminary analysis of transcripts; as such the researcher recorded time-on-task data throughout analysis and focused efforts on creating a framework that would present the most efficient methods, including total time spent on analysis. It was determined that the elimination of collocation analysis did not detract from findings.

The utilization of the co-occurrence tool proved capable of providing sufficient information, such that when combined with findings from frequency analysis and conceptual clusters, they could direct a researcher to areas in the transcripts that would
begin to address the primary research questions. If an institution was utilizing this method and primary research questions were unavailable, the framework utilized in our case study would likely provide information regarding larger thematic elements in the transcripts. Research questions could then be generated from this information.

The limitations of this study are the result of the subjective nature of interview data and inherent in the nature of qualitative research. Despite best efforts to obtain high quality data analysis with the assistance of natural language processing tools the sentiments expressed in interviews are only subject to analysis to the extent of highlighting textual patterns that represent larger themes. The limited demographic variance among researchers and the small cohort size (n=20) could be seen as a limitation to findings, as a larger cohort would provide the opportunity to more fully test the functionality of automatic textual analysis. This study should merely act as a pilot to future research studies utilizing textual analysis and qualitative interview techniques, and not be viewed as a comprehensive review of the full capacity of this methodology.

Another limitation is the availability of a user-friendly natural language processing tool. TAPOR was selected for this study based on its high level of usability and graphical user interface. Requiring knowledge of Java could be an impediment to replicating research such as this in the future for those with less technical expertise, and not choosing to utilize TAPOR. In addition, not all institutions will have the time or means to access an outside reader with professional expertise in the field of open access and scholarly communications, or with expertise in the chosen area of inquiry.
Conclusions

This project supports academic librarians and institutional leaders in operationally defining needs and systematically addressing them with data driven research. Combining natural language processing tools with ethnographic fieldwork presents a new opportunity for further exploration of the results of combining these research methods. Minimizing time and effort invested in the preliminary investigation of interview transcripts by utilizing tools such as TAPOR could provide future researchers with an incentive to utilize qualitative interviewing techniques as they provide rich information regarding behaviors, processes, and methods related to any number of fields.

The findings in this study could be expanded by utilizing the methods we propose with a staff member who has little or no prior knowledge of interview transcripts used for preliminary analysis. Furthermore, analysis conducted utilizing a natural language processing tool other than TAPOR would contribute more broadly to the discussions of Sujin Kim as well as the work presented here. Contributions from researchers utilizing other NLP toolkits and methods will provide a more complete picture of the overall potential for utilizing simple natural language processing tools in academic research. These contributions will reveal the full potential of NLP tools for analyzing transcripts and other personal communications, such as blogs, that are prone to subjective discourse and complex conceptual relationships but are, nonetheless, valuable sources of data that are often untapped in the scope of an academic library environment.
Future studies could replicate this method utilizing research questions specific to their environment and desired findings. In the case of this data a library could, theoretically, conduct a preliminary investigation on transcripts in order to determine if they warrant more focused human attention and analysis. The individual conducting preliminary investigations need not have been affiliated with the initial interviews, but could determine if there are areas of interest in the data revealed during focused textual analysis runs with the use of an automatic text-processing tool. If there were no significant findings the study could be abandoned, or, if there were word clusters or concepts that appear with high frequency or in unexpected combinations, the study could be passed on to a researcher or librarian with more significant subject-based knowledge of the research questions at hand.

Future studies focused on expanding the discourse surrounding the applicability of natural language processing tools in qualitative analysis and research would contribute valuable information to those institutions wanting to explore this method of research, but having limited knowledge of natural language processing methodologies. A key element to the success of utilizing this method of preliminary analysis that was not considered prior to beginning the study is the importance of communication between the individual conducting data analysis and the individual or group seeking the information. The human interlocutor is paramount for translating textual analysis results into meaningful information for administrators hoping to assess the value of the interviews or discourse being analyzed. This conclusion is most accurately reflected in the comments of the outside reader for this study, Will Cross, as he noted:

Overall, I think the key to establishing the value of this tool is going to be providing several methods to leverage it that resonate with real users. Being able
to say "these are the top terms and you can use this information to reveal about the reports" will be very important. I get the sense that this could be really useful if a smart statistics-oriented person showed me how to harness it, but without that I struggled at times to figure out how to get the most value from the data.

By expanding the framework we have proposed here future studies will contribute to meaningful discussions of the application of NLP tools. Because the nature of this method is such that the NLP tools act as an intermediary form of analysis by neither gathering data, nor providing conclusive results, there is still room for methodical development of the proposed executive report. Developing a suggested framework for the development and presentation of an optimal executive report through the inclusion of visualizations as well as specific communication dialogue would be valuable in extending our framework. With little analysis conducted on optimal use of human translation capabilities other than the opinion of an outside reader, this area of inquiry leaves some gaps in the expected roles of the executive report and the individual delivering the report. From our study findings we believe that the human translation of analysis results is necessary to illustrate the full functionality of this method of preliminary analysis until the tools used in natural language processing analysis are able to better illustrate connections between terms in a corpus; however, a more stable framework for reporting results using this method of preliminary analysis is needed.

The nature of the framework presented in this study is such that it is situated in between complete transcript analysis using more powerful tools, and the complete lack of analysis due to a failure to engage with ethnographic interviewing methods because of the time required for analysis. Because of the unique position of this framework exploration of the full potential through future studies is encouraged, and aided by the generalizability of the methodology, which increases the likelihood of multi-institution
utilization and testing. The full potential of the methods explored in this study and the value in utilizing natural language processing tools for ethnographic data analysis will continue to be uncovered through contributions from future research and experimentation.
Resources


Moore, M., Edwards, P., Greenberg, J., & Swogger, S. (2010). *What are the effects that two author-fee subsidy programs have on researchers’ work practices and publishing behaviors?* Poster presentation at the Mid-Atlantic Chapter of the Medical Library Association Annual Meeting (A New Vision for Health Sciences Libraries), Chapel Hill, NC.


conversational speech. *Computational Linguistics*, 26(3): 339-73. DOI: 0.1162/089120100561737


Appendix A

Discussion prompts from interviews conducted as part of IRB study 10-0916.

Main questions/prompts:

1. “Please tell me about this article."
2. "Please tell me about the project that it reports about."

Sample follow-up questions:

1. "Why did you consider posting/publishing this article here?"
2. "Are there any other places that you would have rather published/distributed this article? How difficult was the decision? Why did you decide against these options or for the option you selected?"
3. “What helped as you thought about where to publish? What hindered as you thought about where to publish?”
4. "Were you looking for a particular venue to publish/distribute this work? Were you trying to publish/distribute to reach a particular audience?"
5. "How did you approach any issues related to author-fees for this article?"
6. "Were there any other alternative sources of funding for author-fees which you considered? If so, what were they?"
Appendix B

Service Evaluation Proposal: Open Access at the University Library

Submitted by: Amber N. Welch

This document contains a summary of findings from preliminary evaluation of interview transcripts utilizing the Textual Analysis Portal (TAPOR). These findings present statistics based on textual analysis and are indicative of the types of topics that were discussed during interviews conducted in Fall 2010 examining author attitudes towards open access publishing at a four-year research university. Findings are not conclusive, but merely indicate areas of interest within the transcripts that may be valuable for researchers interested in the Library's current service model. It is recommended that further analysis be conducted utilizing a more powerful statistical analysis tool, such as SPSS or NVIVO, if findings in this document are determined to be indicative of the types of information needed to reevaluate the current service model for open access initiatives.
2 Service Evaluation Proposal: Summary of Results from TAPOR Analysis

Research Questions:
1. What are the main themes or areas of interest discussed during the interviews?
2. Of concepts identified through textual analysis, which have the strongest association with open access?
3. Is there an association between time to publication (ie: speed) and consideration of open access?
4. Is there an association between impact factor and open access publishing?
5. Is there an association between funding and open access publishing?
6. Is there an association between collegial attitudes towards open access publishing and consideration of open access?
7. Does discussion of the library occur frequently, infrequently, or not at all?

Key Findings:
The corpus contained 3137 unique words other than those in the stop word list. There was a total word count of 66,448, reduced to 20,784 after removing stop words, resulting in an average word count per transcript of 3,322 words. There were 20 total transcripts analyzed.

<table>
<thead>
<tr>
<th>Top 25 Terms, by frequency:</th>
<th>Concept terms with highest co-occurring terms, by frequency:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Journal</strong></td>
<td><strong>Science</strong></td>
</tr>
<tr>
<td>Think</td>
<td>595</td>
</tr>
<tr>
<td>Publish</td>
<td>418</td>
</tr>
<tr>
<td>Access</td>
<td>321</td>
</tr>
<tr>
<td>People</td>
<td>314</td>
</tr>
<tr>
<td>Paper</td>
<td>287</td>
</tr>
<tr>
<td>Open</td>
<td>281</td>
</tr>
<tr>
<td>Kind</td>
<td>260</td>
</tr>
<tr>
<td>Work</td>
<td>169</td>
</tr>
<tr>
<td>Time</td>
<td>158</td>
</tr>
<tr>
<td>Article</td>
<td>156</td>
</tr>
<tr>
<td>Review</td>
<td>125</td>
</tr>
<tr>
<td>Interest</td>
<td></td>
</tr>
</tbody>
</table>

Top Word Pairs:

<table>
<thead>
<tr>
<th>Word Pairs</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open access</td>
<td>283</td>
</tr>
<tr>
<td>Impact factor</td>
<td>42</td>
</tr>
<tr>
<td>Little bit</td>
<td>38</td>
</tr>
<tr>
<td>Access journals</td>
<td>30</td>
</tr>
<tr>
<td>Things like</td>
<td>22</td>
</tr>
<tr>
<td>Post doc</td>
<td>17</td>
</tr>
<tr>
<td>High impact</td>
<td>17</td>
</tr>
<tr>
<td>Really good</td>
<td>17</td>
</tr>
<tr>
<td>Editorial board</td>
<td>17</td>
</tr>
</tbody>
</table>
Selected quotations from co-occurrence analysis:

...because I feel like most open access journal rely less on impact factor...

...you have to pay an additional fee to be open access and we've published under both models...

...in a journal called Genome Research which again has an open access option but you have to pay more for it...

...their papers are very similar and the experience with an open access journal it doesn't seem any different...

...every time you publish in a journal that doesn't have open access publishing you've got to sign a copyright waiver...

...I think I submitted it to another non-open access journal before and it got rejected....

...One of the things that's not great about open access, is that the authors pay...

...we can pay extra and make them open access and both times I've opted not to do that...

...its the impact factor, its the type of science, frequently people will start with higher impact journals and then as that doesn't go...

...the idea is to get it into a high impact journal as you can so as many people can read...

...in the department who are sufficiently aware or care about open access that they would spend their research money for it...

...but I value the open access enough to devote money to it....

...I paid a ton of money to get it immediately open access...

...people are becoming aware of the open access and serials crisis, but their behaviors not changing that...

...people are all more interested or knowing in support of open access, it's there, it's really just whether you've been exposed...

...generally people in the department are very supportive of open access publishing....

...And if we had had more time we might have selected a different journal....

You know, this is my first time through as a lead author on an open-access journal.

...to publish a paper it takes them a really long time to prepare figures for publication for example...