Application of machine learning to mapping primary causal factors in self reported safety narratives

S.D. Robinson *, W.J. Irwin, T.K. Kelly, X.O. Wu

Parks College of Engineering, Aviation and Technology, Saint Louis University, Saint Louis, MO 63103, USA

ARTICLE INFO

Article history:
Received 31 March 2014
Received in revised form 15 January 2015
Accepted 6 February 2015
Available online 23 February 2015

Keywords:
LSA
Adaptive taxonomy
Safety
Automatic indexing
Machine learning

ABSTRACT

A new method for analysis of text-based reports in accident coding is suggested. This approach utilizes latent semantic analysis to infer higher-order structures between documents and provide an unbiased metric to the narrative analysis process. Results from this study on a small sample of aviation safety narratives demonstrates an unsupervised categorization accuracy of 44% for primary-cause within the existing taxonomy. If provided with a large sample set, the indication is that a significant increase in accuracy is possible along with the possibility of recoding between data sets. Demonstrated is the ability of LSA to capture contextual proximity of a narrative.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Place yourself in the role of a safety manager, who is accountable for identifying and managing safety trends within your organization. A central element in assessing your organization's current safety climate and identifying safety trends involves making sense of voluntarily submitted safety reports. Several questions evolve for a safety manager attempting to make sense of a submitted report: “Is this incident of concern?” “Are other similar events happening within the organization?” “Does this report signify a trend?” “Is this an area of significant concern?” “What actions are appropriate to manage or mitigate this threat?” “What are the risks to the organization?” Central to answering these questions is the ability to efficiently identify reports of a similar nature within the organization and industry. Imagine having received the following narrative report.

“After arrival at gate, the Flight Attendant disarmed door 2R and then proceeded to door 2L where he began opening the door without disarming it. Realizing his mistake, he attempted to disarm the door, but the gate agent outside the door began trying to open the door resisting the flight attendants attempt to close and disarm the door. The slide pack fell onto the cabin floor but did not inflate. Company personnel were summoned and took control of the situation. As Captain, I was still in the cockpit finishing the last of my cleanup procedures and was unaware of the events as no one notified me. I became aware of the situation only when going to the door 2L area where I became aware of what was going on.” (ACN Report Number 839745, Primary Problem: Human Factors).

A safety professional, when confronted with a report narrative, engages in a process of sensemaking, parallel to that as described by Weick et al. (2005). Sensemaking ultimately answers the question of “what does an event mean?” (Weick et al., 2005, p. 410). Voluntary safety narrative reports play a role in the sense making process by providing a mechanism for collecting data that leads safety professionals to identify problem areas and discover meaningful trends. Coding taxonomies and text based searches provide safety managers with a tool for searching safety narrative databases for similar reports, useful in the sensemaking process. Machine learning techniques such as latent semantic analysis (LSA) provide an additional technique for safety managers engaged in sensemaking.

LSA aids the safety professional in understanding the meaning or significance of narrative reports, by relating them to other organizational events. This parallels the sensemaking steps of selection and retention described by Weick et al. (2005). LSA uses a different method from that provided by coding taxonomies to identify similar narrative reports. LSA has the potential to provide greater flexibility and to be more adaptive than traditional taxonomies. LSA matching provides an automated computer process for identifying similar report narratives that is less subject to
human coding limitations and more efficient in terms of human effort.

To provide an illustration, consider the aforementioned voluntarily submitted report narrative. The LSA process allows the safety practitioner to generate a list of report narratives both within and external to their safety reporting system. In this case, the submitted report was compared using LSA with a sample database generated from the Aviation Safety Reporting System (ASRS). Using a predetermined threshold (cosine <0.50), 37 report narratives of a similar nature were identified. In contrast to a taxonomic search, which relies on a coding structure, these reports were heuristically generated strictly based on textual similarity. To illustrate a limitation of the ASRS taxonomy, these reports were coded by ASRS experts with a range of primary problems, including Aircraft (17), Human Factors (11), Ambiguous (8), and MEL (1). The narrative reports that follow indicate the closest narratives to our exemplar, as contained in our ASRS sample dataset, from the primary problem categories of Aircraft, Human Factors, and Ambiguous.

The nearest report coded with a primary problem of Ambiguous is as follows:

Ready for pushback, Flight Attendant from door 2L came to cockpit to indicate that door 2L was not arming properly, that the wedge was not coming out. [I] went to door 2L to investigate. Looking at door, moved lever to disarm and looked again. Opened door handle and it was apparent that door was not disarmed, door with assist opened partially. Slide was attached at bottom and partly out of pack. I tried to close door but slide then deployed. No one was hurt; no one else was involved in manipulating the door. Maintenance was called and it was determined that the slide would be replaced and flight to operate with a delay. (ACN 1031966, LSA cosine 0.861).

The nearest report coded with a Primary Problem of Human Factors is as follows:

I was the ‘A’ Flight Attendant and was feeling nauseous and dizzy during descent. When I reached down to disarm the L1 door I must have disengaged the girt bar and instead of attaching it above to disarm, I rearmed it. I attempted to open L1 door along with the Agent, when I realized the door was still armed. We closed and I disarmed the door, but the slide pack had dropped into a position that prevented us from opening the door. Maintenance had to be called to remove the slide pack. We deplaned. I believe that I am experiencing symptoms of a sinus infection. The dizzy, nauseous sensations I was having contributed to a potentially deadly mistake. Even though I cross checked myself, had the red flag up, and made my announcement I will always be conscious of how the door feels and aware of the dragging sound the slide makes when the door has not been disarmed and you attempt to open it. That final awareness saved me from one of my worst fears. (ACN 983720, LSA cosine 0.836).

The nearest report coded with a Primary Problem of Aircraft is as follows:

This was a charter flight. The aircraft was parked in the north lot. Airstairs were brought to the aircraft door at 1L. It appeared the person trying to open the door was having difficulty opening the door. I cracked the door. He still appeared to be having difficulty, so I gave the door a push. The person on the other side was still having difficulty opening the door. I soon saw why. The side of the slide pan was caught on the side of the aircraft door. As the person on the outside continued to pull the door open the slide pan opened and the slide fell out, but did not deploy. A Mechanic arrived to detach the slide from the door. He said the door was armed. The arming mechanism was stuck between arm and disarm and we were unable to put it either the arm or the disarm position. I told him that I disarmed the door. After I disarmed it, I made the all call to disarm cross check and stand by for all calls. It is possible the next time a slide may in fact deploy. (ACN 969496, LSA cosine 0.832).

The similarity between the above reports should be self evident, despite differences in the primary problem coding. In each case, an inadvertent deployment of an exit slide was a possibility, as evidenced by the actual deployment in one case, and the reporter expressed concerns of a possible deployment in the other cases. In this case, use of the LSA process utilizing only the raw report narrative generated a list of 37 reports many of which are useful in the sense making process. These reports covered a range of primary problems and contributing factors. Of those reports, 16 reports dealt specifically with the improper arming or disarming of exit slides (including inadvertent deployments). 20 of the reports were related to improper door operation. The final report involved a damaged cockpit door.

In contrast, a similar search of the same database using multiple text strings was conducted. These search strings included “inadvertent slide deployment”, “door AND disarm AND flight attendant”, and “door slide OR exit slide OR inadvertent deployment OR improperly armed OR improperly disarmed”. In total, the text based searches captured 11 of the LSA generated report narratives, of which seven related to inadvertent slide deployments. The LSA process generated 13 reports beyond that of our simplified text search. Anecdotally this indicates that the LSA process provides safety managers with an additional tool beyond coding taxonomies and text searches for identifying similar report narratives within large databases. The reliance of LSA on raw report narratives avoids the time and effort required to code incoming reports, overcomes the limitations of existing taxonomies, and provides additional flexibility in generating safety reports across databases.

1.1. Latent semantic analysis

The methodology of latent semantic analysis (LSA), developed by Deerwester et al. (1990), has seen many applications since original publication. The process has subsequently been applied to diverse fields of inquiry, from educational theory (Landauer and Dumais, 1997) to automated document classification (Liu et al., 2004; Landauer et al., 1998; ZUKAS and Price, 2003). Thus far, textual analysis techniques such as LSA have not been widely applied within the field of safety management, despite the abundance of narratives alongside other quantitative data. In aviation, accident analysis narratives are commonly used to manually discriminate factors developed through traditional statistical methods applied to quantitative data.

LSA is a mathematical technique for inferring relations between words within bodies of text. Without the assumptions of other language processing methods, LSA extracts the occurrence of words in a text body and creates a term frequency document matrix. A singular value decomposition (SVD) is applied to the resulting matrix. The central matrix from the SVD is then truncated by the substitution of the lowest values with zero. This truncated, or reduced space, form of the matrix then provides the inferred relationships between terms used in similar contexts. In this reduced space, term associations are made that are not present in any single body of text. Thus latent relationships are revealed by this method.

This application of LSA has been successfully used to match similar texts, answer multi-choice based subject tests, and predict subjective ratings of texts. Studies of document classification problems have indicated that accuracies up to 93% (Huang, 2003) are possible when LSA is combined with support vector machines.
(SVM). Kakkonen et al. (2008) has demonstrated an accuracy of up to 90% for unsupervised LSA alone in predicting the grades assigned by human reviewers to individual text submissions.

1.2. Background

Historically, narrative reports have relied on coding taxonomies to facilitate classification of safety reports (see Bailey, 1989), to provide access to narrative databases (Wallace and Ross, 2004), and to allow quantitative assessment of safety reporting (Bailey, 1989). Taxonomies are useful in managing large volumes of data (Hawkins et al., 2003), providing order to complex systems (Nickerson et al., 2009), allowing for classification and quantification of safety reports, and for identifying patterns of similarity (Nickerson et al., 2009) or variation (Wallace and Ross, 2004). Taxonomies are also useful in describing and explaining safety related behaviors (Perry, 1989). At the theoretical level, the development of taxonomies provides a means for holistically studying particular objects of study (Bailey, 1989) while facilitating hypothesis testing of relationships among objects (Nickerson et al., 2009).

While taxonomies have proven useful in safety science, the use of taxonomies is not without limitation. The process of coding narrative reports into the taxonomy structure is cost and labor intensive (Hawkins et al., 2003), provides inconsistent and/or unreliable results (Wallace and Ross, 2004), and results in the loss of information as data is reduced into limited categories (Taib et al., 2011). Taxonomies function by forcing narrative reports into discrete groups with limited categories, which filter and reduce information available to facilitate comparison or classification (Taib et al., 2011). The structure of the taxonomy decides what is relevant by focusing on particular elements of the data that are predetermined, rather than summarizing or representing the data for meaning and implications (Wallace and Ross, 2004). By forcing data into limited, discrete groups there is the danger that significant information may be lost as narrow incident classifications (such as food borne illness) may be lost in the coding taxonomy, while larger groups (such as human factors or miscellaneous) become overpopulated (Wallace and Ross, 2004). Incomplete or over-generalized taxonomies result in a limited understanding of the incident and limit the recommendations that may be made (Taib et al., 2011). Hierarchical taxonomies may lead to a “stop rule” problem, where the end of the taxonomy may obscure or not fully describe the error (Hollnagel and Amalberti, 2001).

Taxonomies are further limited by their static nature and lack of flexibility. Once the taxonomy has been developed, it is time and labor intensive to recode previously coded reports to reflect the new coding system. Taxonomies need reevaluation as time, context, and understanding change (Wallace and Ross, 2004; Nickerson et al., 2009), many times having evolved in an ad hoc manner (Wallace and Ross, 2004). Since sensemaking is time and context dependent (Dervin, 1998), taxonomies are context specific (Wallace and Ross, 2004) with their usefulness being tied to the needs of the user (Nickerson et al., 2009; Hawkins et al., 2003; Turner, 1989), the nature of the problem being analyzed (Drabek, 1989), and situation or culture being studied (Wallace and Ross, 2004). Taxonomies are theory dependent, evolving over time as theory evolves (Turner, 1989). They may also be specific to the industry or system being classified (Wallace and Ross, 2004). Differences between competing taxonomies, using different coding taxonomies, limit the ability to manage or compare data across datasets. In some cases, taxonomies may fail to keep up with changes in the needs and knowledge of the organization.

The subjective nature of human coding further limits the usefulness of taxonomies. Typically a reliability rating of 70% is considered good for human coding systems (Wallace and Ross, 2004; Taib et al., 2012). The subjectivity of classification leads to unreliability in the absence of clear objective criteria for classification, overlapping characteristics (or applicability to multiple categories) (Smith, 2002), the polysemous nature (multiple meanings) of words, and the subjectivity of coders in deciding what is significant within the narrative (Wallace and Ross, 2004). Taxonomies as socially constructed objects which require a degree of consensus and compromise in establishing terms and meanings (Bailey, 1989; Quarantelli, 1988; Turner, 1989; Wallace and Ross, 2004). The nature of error (establishing correct or incorrect behavior) implies existing consensus about knowledge and theory, which may not exist (Hollnagel and Amalberti, 2001; Dekker, 2004).

The abstract nature of many aviation safety coding terms adds difficulty to the coding process. Dekker and Hollnagel (2004) addresses the dangers associated with mistaking human factors labels with the deeper insights they represent. When terms such as situational awareness, automation complacency, and loss of effective crew resource management are used, people “tacitly assume that others understand the concepts named by the labels in the same way” (Dekker and Hollnagel, 2004, p. 79). In similar fashion, when determining the primary-cause for an accident narrative, some discrepancy in interpretation may exist between categories such as “human factors” and “procedural errors”. In utilizing a different approach to managing safety related narratives, LSA offers potential benefits to aviation safety managers.

The LSA process is not, however, without limitation. First, it does not account for the knowledge and experience of the safety professional in analyzing reports for similarity (see Wallace and Ross, 2004). The LSA process also requires an exemplar report, vector angle, or group of LSA specific themes to begin searching for similar events. LSA is limited in inferring the intentions of actors, given that it relies on the terminology contained in reports rather than implied meanings. LSA while useful for comparison across databases and coding taxonomies, is bounded when crossing industries and cultures. Differences in terminology may impact the usefulness and reliability of LSA across these different datasets. A good example would be when crossing between medical and aviation reporting systems.

In summary, LSA is an alternative technique that has the potential to overcome many of the limitations inherent in coding taxonomies. Instead of relying on a front-end data coding process (pre-analysis), LSA uses a back-end process to manage, organize, and analyze narrative data. Back-end analysis of the narratives has several advantages over traditional coding by taxonomy. By relying on the analog narrative report, instead of a digitally reduced code, the majority of report information is retained for comparison and analysis. The observable features (actions and outcomes), along with contextual factors and conditions are retained for analysis (Taib et al., 2011; Rasmussen and Vicente, 1989). By avoiding the need to categorize reports into discrete, mutually exclusive categories, reports of a similar nature that previously might have been coded differently are linked for analysis. The use of LSA further avoids the subjectivity of the coding process, relying on the raw narrative rather than a subjective, interpretive classification into a coding taxonomy, thereby improving reliability and consistency. Through a process of theoretical sampling, exemplar reports that typify a current problem or event may be used to generate a group of similar narrative reports. This capacity expands the usefulness of LSA by allowing the comparison of narrative reports across databases that use different coding taxonomies.

2. Theory

Through the initial preparation process the corpus is converted to a rectangular document-term frequency matrix. Subsequently through single value decomposition, the matrix is decomposed
into three matrices, the product of which is the original matrix. Eq. (1) represents this process.\(^1\) Columns of the \( \mathbf{U} \) matrix are mutually orthogonal vectors of length equal to the dictionary length. Similarly, \( \mathbf{V}^T \) describes orthogonal vectors equal to the document number. The \( \sigma \) matrix is diagonal and scales the relationship between the other matrices to reconstruct the \( \mathbf{X} \) matrix. The columns of \( \mathbf{U} \) may be interpreted as topics within the subspace of the SVD since they indicate terms that commonly occur together. The first four of these topics is shown in Appendix B. Since \( \sigma \) scales the relationship, the best approximation to the original matrix in a reduced space (of rank \( \Gamma \)) is found by maintaining the \( \Gamma \) largest entries and setting all other entries to zero. This constitutes a least-squares approximation of the matrix to \( \Gamma \) dimensions. The total energy retained, or degree of fit, of the reduced space model is calculated as the sum of squares of the singular values found in the \( \sigma \) matrix (shown by Eq. (3)). The proportion of total energy retained by the model is then the ratio \( E_r/E \). Normal values fall between 80% and 95% for most LSA applications.

Within the LSA model documents are constructed as vectors with each word in the dictionary representing a dimension. Thus documents are compared for their similarity in vector form within the reduced space. Two methods of comparing the document vectors are commonly used in natural language programming. Through these methods document similarity is calculated.

The most frequently used method of determining document similarity is by use of the unit vector dot product. The unit vector dot product, or cosine, measures the projection between vectors in a multi-dimensional space. The calculation of cosine similarity between two documents vectors (x and y) is shown by Eq. (4). The cosine value is taken to represent the similarity between documents. This cosine value may fall between −1 and 1. Where the value is closer to unity, documents are very similar. Where the cosine value approaches zero, documents share no similarities. Alternatively the document vectors may also be compared by measurement of the absolute distance between them (euclidean distance). Unlike the cosine, the euclidean distance is not scaled and may take on any positive value. With this approach, documents are considered to be of greater similarity with smaller values.

These approaches differ significantly when provided with two very semantically similar documents of different length. The similarity of words used will produce a cosine value inferring similarity, while the increase in frequency of words within one document will cause the euclidean distance between documents to increase. Hence in this case of semantically similar documents, the value returned by the cosine method will infer similarity, while the euclidean distance returned value would infer that the documents are dissimilar.

\[
\mathbf{X} = \mathbf{U} \sigma \mathbf{V}^T
\]

\[
\mathbf{X}' = \mathbf{U} \sigma_\Gamma \mathbf{V}^T
\]

\[
E_r = \sum_{i=1}^{\Gamma} \sigma_{ii}^2
\]

\[
\text{Cosine} = \frac{x'y}{\|x\|\|y\|}
\]

**3. Method**

The purpose of this study is to evaluate the effectiveness of LSA techniques as a contextual and thematic evaluation tool for aviation safety report narratives. Specifically four research questions were presented:

- Does the use of LSA techniques reliably identify report narratives of similar themes or context?
- Can thematic similarity be used to predict causal similarity?
- What cosine values provide an appropriate threshold for identifying thematic or contextual similarities between report narratives?
- Does the use of LSA techniques provide an additional tool that may be reliably utilized in analyzing narratives?

A review of the literature failed to indicate any similar applications of LSA technology or specific methodologies for evaluating the LSA process as a tool for safety report analysis. In the absence of verifiable techniques for evaluating the application of LSA to safety reporting a three step evaluation process was chosen. First, in order to evaluate the strength of LSA techniques to reliably identify similarities between report narratives and to establish threshold values for report analysis, two narrative data sets were compared. One was derived from a query to the ASRS database (query corpus), the other consisted of narratives made up of random words generated from the aforementioned query corpus (random corpus). Secondly, to assess the viability of LSA techniques in automatic primary-cause coding of narratives, a comparison of LSA cosine values for nearest neighbor reports to ASRS primary-cause data was conducted. The percentage of nearest neighbor reports with matching primary-causes was determined. Thirdly, in order to further assess the usefulness of the LSA process in analyzing safety report narratives, a review was conducted of report narratives from the training corpus with cosine values above 0.80 and disagreeing primary-causes. The qualitative review of the reports identified was conducted by the researchers independently to ensure an unbiased assessment. The LSA process and each of these three methods will be addressed in more detail.

Data for the three test corpuses was taken from the ASRS repository and was limited to reports that were operated under Federal Aviation Regulation Part 121 and passenger carrying operations. This control had the effect of focusing the reporter type to that of professional pilots, the largest single demographic found within ASRS.

Data from January 2011 to January 2013, which consisted of 4497 reports after filtering for the given criteria, was utilized to create a primary training corpus for the LSA. A second sample was taken from the ASRS database to generate a query corpus, used as a means of cross-validation. This query corpus was filtered in the same manner as the training corpus from the date range January 2009 to December 2009 and consisted of 2987 reports. A third test corpus was generated from the query corpus by generating a sequence of report narratives consisting of randomly assigned words from the query corpus restructured into random word narratives of the same length as those in the query corpus. This random corpus was generated by random selection, without replacement, of words from a collection of all words found in the query corpus. The choice to select without replacement was made such that order, or algorithmic information content, amongst documents would be lost whilst overall content remained. The corpus was then delineated into documents of the same size as the query corpus suitable for LSA comparison to the training corpus. By generating a random corpus for LSA analysis a baseline measure of query corpus order and threshold cosine values could be developed. In the case of the training and query corpuses, primary-causes of each event were extracted from the ASRS database along with the narratives before processing.

In order to facilitate the LSA process, documents were taken through an automated cleaning process involving computer code

---

\(^1\) For an extensive discussion on singular value decomposition please see Henry and Hofrichter (2010).
character was then created from all documents in the training corpus. Stop words, such as ‘the’ and ‘a’, were removed from the dictionary if present in a list of 402 English words shown in Appendix C. Subsequently, words occurring less than 20 times were removed to ensure SVD stability for the small corpus size. This resulted in a dictionary of 2764 words. Each document was then converted into a document vector to construct a corpus matrix. The resulting corpus matrix was transformed by the term frequency-inverse document frequency (TFIDF) approach of Salton and Buckley (1988). Thus words that occur less frequently in the corpus were deemed to be more important and assigned a higher value. A single value decomposition (SVD) was then performed utilizing the gensim software developed by Rehurek et al. (2010). From the SVD, the upper 400 values were retained by removal of the lower values in the central matrix of the SVD. The choice of 400 semantic topics was selected by a heuristic process maximizing primary cause accuracy. In post hoc analysis, this decision was validated by a process which the authors believe to be new to the literature. This choice of topics to retain, resulted in a loss of 9.1% of the energy spectrum.

Utilizing LSA, post construct of the SVD, the queries were compared to the training corpus. Queries were compared with each of the documents in the training corpus by taking the vector dot product (cosine value) and placed in decreasing rank order. The primary-cause from the highest ranked document was returned as the probable primary-cause of the query document. This process resulted in the dichotomous outcome of cause match (where zero indicated that the documents did match) and the continuous predictor of cosine value.

The potential of the LSA cosine value as a measure of narrative context and themes was conducted by comparing the distributions of cosine values for the query and random corpuses considered against all reports and then to only nearest neighbors in the training corpus. From the literature review, no methodology was found to quantitatively determine the strength of the training corpus in capturing the contextual significance of the query. Therefore when providing a query to the training corpus, the concept of minimum value for significance or goodness of fit was absent. A rational approach to develop an internally consistent metric of significance was needed. The decision was made to compare the results of the nearest neighbor cosines for the random corpus against the query corpus. Where documents are similar, cosine values will be higher. Therefore in the case of a large training corpus a query is likely to return a higher cosine, since a similar document is more probable to exist. In the case of a corpus with well defined context, a higher cosine should be returned even with a smaller training corpus. In either case, the strength of the training corpus may be assessed graphically by viewing the separation between the random and query corpus or numerically by percentile scores. In the absence of more evidence, the cosine value of the 99th percentile of the random corpus nearest neighbor distribution was chosen as the minimum value for contextual significance. Where the cosine value associated with this percentile coincided with that of the query corpus nearest neighbor distribution was taken to be a measure of the strength for the training corpus.

The effectiveness of LSA primary-cause matching was evaluated by creation of a randomized primary-cause sample for comparison. For each primary-cause from the query corpus, comparison was made with a random, with replacement, selection of primary-cause from the training corpus. This resulted in a mean value of cause matching accuracy across the population.

In the assessment of the usefulness of LSA to the practitioner in analyzing narratives, a sample of high similarity reports with different primary-cause assignments was selected. A lower threshold of 0.80 was chosen for the cosine. This was considered a high value following a cursory review of reports and their nearest neighbors. In addition, this limit reduced the narrative pairs of interest to 64, which was considered by the researchers to be of manageable size. A selection of those reports which were considered of particular interest are presented in Appendix A.

4. Results

The central SVD (sigma) matrix reduction to 400 non-zero terms resulted in a reduction of the energy spectrum of 9.1%. The 400 values retained plotted by rank order are show by Fig. 1.

For validation of the information content value produced by the LSA, a comparison of each document within the random corpus to the training corpus was conducted. This returned a document proximity cosine value between each document in the random corpus and every document in the training corpus. The histogram of this data is shown in Fig. 2. This distribution returned a mean value of $\bar{X} = 0.1153$ (skew = 0.527, kurtosis = 0.1254, sd = 0.0662). The 99th percentile of the random corpus cosine values was found to be 0.292.

Fig. 1. Values of the diagonal $\sigma$ matrix in rank order.

Fig. 2. Histogram of document proximity cosine values for all documents within the training corpus when subjected to documents from the query corpus and the random corpus.
A comparison of each document within the query corpus to the training corpus was also conducted. This returned a document proximity cosine value between each document in the query corpus and every document in the training corpus. The histogram of this data is shown in Fig. 2. This distribution returned a mean value of $X = 0.0775$ (skew = 2.257, kurtosis = 9.0276, sd = 0.0701).

To determine the improvement of the nearest neighbor matching above chance, the primary-causes from the query corpus were matched with a randomly selected cause from the training corpus. Thus the distribution of the 18 cause categories (including a null entry) found in the ASRS database was represented proportionally. The resulting primary-cause accuracy match was determined to be $X = 0.227$. The pos that the primary-cause of a query will likely be the primary-cause of the nearest neighbor found within the training corpus was tested. This primary-cause matching technique returned an accuracy of $X = 0.449$. Thus, the matching precision of the LSA approach was determined to be considerably (22.2%) above that of chance.

Fig. 3 shows the distribution of nearest neighbor cosines from the query corpus and the random corpus when compared to the training corpus. The 99th percentile was found to occur at a cosine value of 0.516 for the random corpus. This corresponded to the same cosine value as the 26th percentile within the query corpus distribution.

5. Post hoc analysis

In the construction of this paper it became clear that two assertions had been made without substantial foundation, either through lack of rationale in the literature or directly by the authors. Firstly, the selection of topic number for the LSA was based on result accuracy. Hence, the choice of 400 topics was made after a number of computational iterations, where topic number was plotted against primary-cause matching accuracy. In the second instance, the qualitative review process initially used was determined to lack sufficient rigor to show that the cosine value provided a substantive measure of narrative similarity. These shortcomings of the primary analysis were addressed as follows.

5.1. Criterion for the selection of topics to retain

The number of semantic topics retained in the LSA process is traditionally determined by heuristic processes. Since many natural language processing problems attempt to recreate the results of a human undertaking, validated samples of problems and expected solutions exist. Therefore, processing techniques can be tailored to a sample problem set before being subject to an unseen challenge. When posed with the question of how many topics to retain using the LSA technique, the common answer is to measure the accuracy of the LSA output against a verified data set. Thus, no internally consistent process is essential to obtain an adequate result.

This outcome based approach fails to offer any logically grounded framework for the application of results outside the study being considered. When working with safety narratives, there are many different taxonomies by which to categorize a report. Thus, the LSA technique would likely result in different values of topic number for each different application even on the same data set.

In light of the inherent ambiguity in assignment of ‘cause’ from a self-reported narrative, the need for an alternate criterion became self evident. The key to development of a internal criterion came from the creation of a random corpus and the comparison of nearest neighbor cosine values. The summation of the cosine values of nearest neighbors describes the similarity of a given query corpus to the training corpus. Where the summation is divided by the number of queries, the outcome is an average of the nearest neighbor cosine. The numerical comparison of the query corpus average to the random corpus results in a measure of information above that of chance. This measure of information complexity may be maximized by manipulation of the number of topics retained. Where the information complexity is maximized, the topics retained may be considered ideal for the given training and query corpora.

This process of maximizing information complexity of the LSA was completed for the selected ASRS corpora. Fig. 4 demonstrates that the greatest degree of complexity occurs at approximately 400 topics retained. Thus, this new approach was reconciled with the initial outcome based selection of topic number.

5.2. Significance of the LSA nearest neighbors

Evaluation of several LSA nearest neighbor pairs intuitively indicated that reports of a similar nature were provided by LSA nearest neighbor pairs. In order to apply a more rigorous process to assessing the similarity of LSA nearest neighbor pairs, post hoc testing was done on a random sample of LSA nearest neighbor pairings. 26 nearest neighbor pairings were identified through
random selection to include LSA reports above and below the approximate threshold value of 0.50. Given that our interest was primarily reports above the threshold value, ten nearest neighbor pairs were selected at random from pairs above the threshold value. Three additional nearest neighbor pairs were randomly selected for coding from pairs below the threshold value. The selected collection of reports allows for the analysis of 287 non-nearest neighbor pairings in addition to the 13 nearest neighbor pairs selected.

To provide a systematic process for assessing the similarity of the nearest neighbor reports, each report was coded using an open coding process consistent with grounded theory (Glaser and Strauss, 1967). One of the researchers conducted the open coding of the reports, generating approximately 800 coding terms. These coding terms were reduced through a focused coding process (Glaser and Strauss, 1967; Charmaz, 2014) to develop a usable taxonomy consisting of 162 coding terms. The results of this coding process were tabulated, with the results being used to generate a 162 dimensional vector for each report. The vector was generated for each report by taking a count of the number of times each coding concept appeared in the report, plotted in each associated dimension. The resulting vectors from report pairs were compared and consolidated into a single cosine value using a similar process to that of the original LSA. To assess inter-rater reliability, a second coder utilized the developed taxonomy to code each report narrative, with a similar set of cosine values generated across each report pair.

The cosine values resulting from the qualitative coding process were compared with the LSA cosines for the same report pairings. A scatterplot of these results in contained in Figs. 5 and 6. The results (shown by Table 1) for researcher A indicate that there is a large correlation (consistent with Cohen, 1992) between LSA cosine and qualitative cosine values for all LSA nearest neighbor values. The reported Pearson's $r$ values are, $r = 0.615$ for all reports, $r = 0.664$ for reports above a cosine threshold value of 0.292, and $r = 0.897$ for reports above a cosine threshold value of 0.50. The results for researcher B are less conclusive, indicating a medium effect for the grouping of all reports ($r = 0.438$). The results for the 0.50 or 0.292 cosine thresholds were not statistically significant. The authors suspect that this lack of significance at the 0.50 and 0.292 cosine values is due to poor inter-rater reliability in the coding process or low statistical power due to the limited sample size. This suggests a need for further study involving a greater number of coders and a larger dataset.

In interpreting these results it is important to remain mindful of the normality of each distribution. Since this dataset was generated from a relatively small set of coded reports randomly selected based on nearest neighbor pairs, the cosine values between non-nearest neighbor pairs are generally low. The result is a positively skewed, leptokurtic distribution of cosine values for the LSA and qualitatively coded reports. Descriptive statistics for each set of cosine values is included in Table 2. The normality of each distribution was verified using the Shapiro–Wilk test (Table 2) which demonstrated that the 0.292 and 0.50 threshold distributions approached normal. The non-normality of the distributions for all reports is of limited concern given that the correlation values are generally robust to violations of the normality assumption and are likely to result in reduced power rather than inflated significance (Glass et al., 1972; Edgell and Noon, 1984; Norman, 2010).

It is worth noting that researcher A and researcher B have significantly differing backgrounds that may have contributed to differences in coding results. Researcher A is a former airline pilot with several thousand hours experience and a background in qualitative research methods. Researcher B is a trained air traffic controller with a background in quantitative research methods. Researcher B is a trained air traffic controller with a background in qualitative research methods. There is also the potential that language differences, unfamiliarity with the coding taxonomy, and unclear instructions regarding the coding process may have contributed to the differing results.

Overall, these results suggest that there is a correlation between report similarity values generated from a qualitative coding process with the cosine values generated by the LSA process. This suggests that LSA is a useful tool for identifying ASRS narrative reports possessing similar content, and that the LSA process at least in part replicates the results of a qualitative coding process when searching for similar reports. The differences in results between researcher A and researcher B are likely due to differences in coding and the smaller $n$ values at the 0.292 and 0.50 thresholds. It is suggested based on the results of this post hoc testing that further study is needed to evaluate the consistency of LSA cosine values when compared to a qualitative coding process. This study should utilize multiple qualitative coders, with a normalized distribution of LSA cosine values, and a greater number of reports above the associated cosine thresholds. This post hoc evaluation further suggests that LSA has the potential as a tool for evaluating the reliability of coders and taxonomies.
6. Discussion

The results presented here provide a proof of concept to the use of machine learning in safety programs and research. Of the research questions asked, all four were addressed successfully. This outcome was particularly surprising given the small training corpus used and demonstrates the power of the LSA approach. Appendix A provides empirical evidence and a substantial foundation for the discussion and post hoc should therefore not be overlooked.

Does the use of LSA techniques reliably identify report narratives of similar themes or context? An incident narrative is inherently a story written in reflection by the person who lived that experience. As with any story there is a beginning, middle, and end. Often there is also a conclusion with suggestions as to how the individual or organization is to avoid a similar issue. Therefore each narrative forms a complex series of events. Review of Appendix A clearly demonstrates the strong relationship between higher cosine value and closer thematic similarity.

Can thematic similarity be used to predict causal similarity? The moderate degree of primary-cause matching accuracy can be closely associated with the small training corpus, when compared with other LSA inquiries, used in this study. With LSA and other machine learning techniques, an increase in data provided for training results in improvements in accuracy. This is especially true for LSA matching alone, where queries are compared to the corpus used to train the LSA. Provided with the entire ASRS database or a major air carrier ASAP database the accuracy would likely increase substantially. Table 3 shows that the LSA matching approach to causal prediction does not improve with increasing cosine value. This result reflects the premise that incident cause and thematic similarity are different.

What cosine values provide an appropriate threshold for identifying thematic or contextual similarities between report narratives. The 99th percentile value of the nearest neighbor cosine values of the random corpus against a given training corpus may be considered as a reasonable point to which contextual similarity may be significant. This cosine value for significance as a subjective measure will have to be determined by the researcher based on their needs and from the corpora used in the inquiry. It is important to caution that in the case of a nearest neighbor, a small cosine value (close to zero or negative) returned from the training corpus indicates uniqueness of the given query. Such a query may be of higher importance to the researcher.

Does the use of LSA techniques provide an additional tool that may be reliably utilized in analyzing safety report narratives. Literature related to machine learning and safety taxonomies is notably absent in aviation and other high consequence industries. Where the LSA did not correctly match incident primary-cause, an anecdotal review of the narratives was conducted by the researchers. This survey provided evidence highlighting the inherent limitations of the training corpus used in the LSA and the ASRS taxonomy. Human coding is biased and prone to error even for highly experienced coders (Wallace and Ross, 2004). Documents may fit into several categories. The ASRS taxonomy was developed with a limited set of incidents that do not represent all categories that may significantly affect safety. One significant example of a series of incidents with a serious effect on safety but without a relevant category in the taxonomy was related to food. Air crews during normal operations may ingest food that, due to mislabeling or preparation, causes an allergic or toxic reaction. The possible effect on crew members is significant and has the potential to cause an accident, yet such events are coded with ‘Human Factors’ as the primary-cause within the ASRS taxonomy. Appendix B shows the topics developed by the LSA.

Table 1
Paired correlation (Pearson’s r) statistics for nearest neighbor cosine values for LSA when compared to researchers A and B for three cosine thresholds (all, 0.292, 0.50).

<table>
<thead>
<tr>
<th>Comparison</th>
<th>n</th>
<th>Corr.</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All cosine values</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSA – Researcher A</td>
<td>300</td>
<td>0.62</td>
<td>0.00</td>
</tr>
<tr>
<td>LSA – Researcher B</td>
<td>300</td>
<td>0.44</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Cosine values above 0.292</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSA – Researcher A</td>
<td>16</td>
<td>0.66</td>
<td>0.05</td>
</tr>
<tr>
<td>LSA – Researcher B</td>
<td>16</td>
<td>0.33</td>
<td>0.21</td>
</tr>
<tr>
<td><strong>Cosine values above 0.50</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSA – Researcher A</td>
<td>10</td>
<td>0.90</td>
<td>0.00</td>
</tr>
<tr>
<td>LSA – Researcher B</td>
<td>10</td>
<td>0.286</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Table 2
Descriptive statistics for nearest neighbor cosine values for LSA and researchers A and B for three cosine thresholds (all, 0.292, 0.50).

<table>
<thead>
<tr>
<th>n</th>
<th>X</th>
<th>sd</th>
<th>Var</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Shapiro–Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>val.</td>
<td>se</td>
<td>val.</td>
<td>se</td>
<td>val.</td>
<td>sig.</td>
</tr>
<tr>
<td><strong>All cosine values</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSA</td>
<td>300</td>
<td>0.063</td>
<td>0.124</td>
<td>0.015</td>
<td>3.34</td>
<td>0.14</td>
</tr>
<tr>
<td>A</td>
<td>300</td>
<td>0.139</td>
<td>0.133</td>
<td>0.018</td>
<td>1.39</td>
<td>0.14</td>
</tr>
<tr>
<td>B</td>
<td>300</td>
<td>0.174</td>
<td>0.142</td>
<td>0.021</td>
<td>0.76</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>Cosine values above 0.292</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSA</td>
<td>16</td>
<td>0.554</td>
<td>0.129</td>
<td>0.017</td>
<td>0.08</td>
<td>0.56</td>
</tr>
<tr>
<td>A</td>
<td>16</td>
<td>0.373</td>
<td>0.166</td>
<td>0.028</td>
<td>-0.21</td>
<td>0.56</td>
</tr>
<tr>
<td>B</td>
<td>16</td>
<td>0.345</td>
<td>0.152</td>
<td>0.023</td>
<td>-0.10</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>Cosine values above 0.50</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSA</td>
<td>10</td>
<td>0.632</td>
<td>0.086</td>
<td>0.007</td>
<td>0.81</td>
<td>0.69</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>0.451</td>
<td>0.128</td>
<td>0.016</td>
<td>0.46</td>
<td>0.69</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>0.373</td>
<td>0.142</td>
<td>0.020</td>
<td>-0.01</td>
<td>0.69</td>
</tr>
</tbody>
</table>

LSA represents the findings from latent semantic analysis, A those from researcher A, and B those from researcher B.
6.1. Possible applications

LSA query ranking allows for filtering and association of events in a similar manner to the functionality offered by a taxonomy. Since LSA is a machine-based approach human coding bias is absent. Additionally with commonly accessible computing power, rapid recoding and cross dataset comparison is possible. This process may be applied to any area where narratives occur alongside quantitative variables in large data sets, such as medicine, aviation, and social work.

7. Conclusion

The relative success of this approach to coding of safety incident narratives strongly suggests that further research in the application of machine learning to safety management would be highly productive. The LSA approach for event indexing and matching, as well as a qualitative assessment tool has been clearly demonstrated. This process highlights the inherent limitations of taxonomies and provides a new mechanism for meeting the needs of safety programs and researchers. This research highlights the potential of LSA as a tool to address the shortfalls of both qualitative and quantitative research methods in safety research. Provided access to a larger narrative database, contextual comparisons could be made across databases and disciplines.

Acknowledgement

The authors would like to thank the two unknown reviewers for their constructive comments to prior versions of this paper.

Appendix A. Narratives of interest

The following narratives were identified during anecdotal review of the data. Events that appeared to fit poorly into the taxonomy and those which had very close neighbors not of the same code are provided for review. Narrative numbers with the letter ‘Q’ following the index are taken from the query corpus. Those narratives taken from the training corpus are numbered without following characters.

Narratives 1388 (ACN 962540) and 1282 (ACN 959858). Respective primary-causes: ‘Human Factors’ and ‘Human Factors’.

... At approximately 2 h into flight we both had our crew meals. The First Officer had the Ravioli; and I had the Teriyaki Steak dinner. Within 15 min of eating the meal; I started experiencing the onset of an allergic reaction. My eyes began to itch; my nose started to run; and my hands began to itch. The rate at which these symptoms appeared was familiar to me as I am severely allergic to certain food items (peanuts; pine nuts; sesame seeds; pumpkin seeds; etc.). Within 30 min of eating the crew meal; my tongue had swollen to the point where I was not able to speak understandably; my eyes had swollen to the point where my vision was significantly restricted; my breathing was labored. ... I also retrieved my Epi-Pen from my suitcase. I instructed the jump-seater on how to administer the Epi-Pen ‘just in case.’ ... Post a memo alerting flight crews about the possibility of being served crew meals that may cause allergic reactions.

Cosine between this narrative and the following one was 0.535, which within the LSA document space represents a weak association, however this was the strongest available with the corpus and the documents were coded similarly.

I was the Captain and pilot not flying (PNF) for this leg. Prior to departure I had the very; very slightest onset of a sensation of vertigo. I was able to walk normally and there was no impairment to my ability to function as a pilot. After takeoff; the vertigo escalated at a truly amazing rate. After some twenty minutes; I handed off the PNF duties to the First Officer thinking that I was incurring some form of food poisoning. At a certain point; the vertigo was so extreme that when I went to stow my headset to my immediate left; I was ‘thrown’ hard to the left. My vision became jittery owing to the inner ear/vision conflict. I was soon using oxygen with my forehead braced against the glare shield. The sensation of vertigo was now extreme; like the worst ride in a vertigo trainer times five. Approximately an hour and a half from destination; I declared myself incapacitated. Utilizing the help of my number one Flight Attendant; I was able to shift from the left seat to the right jump seat. I then summoned a B767 First Officer who was in the cabin as a passenger and installed him in the left seat. I made a few constructive suggestions to the crew and let them proceed. A medical emergency was declared at some point. I urged the crew to proceed to destination as other than the vertigo; I had no other symptoms. On descent; I incurred rather violent vomiting. Utilizing my Captain’s authority; I admitted a nurse to the flight deck. She took my vitals and administered a cold compress to my neck; which helped alleviate the discomfort an appreciable amount. The nurse then left the cockpit. I urged the crew to coordinate a tow-in at the gate in the interests of safety. Although the tug was not waiting; little time was lost getting us to the gate and paramedics removed me from the flight deck. Several points come to mind looking back on this event. First; never having experienced vertigo; I had no idea a medical event like the one I experienced could occur. Had I even the slightest idea; we would not have departed. I might have declared myself incapacitated earlier than I did. I don’t believe anything bad became of it; but the crew would have had perhaps thirty more minutes to organize themselves. As it was; they did very well with the time (approximately 1 h 20 min) allotted to them. The role of the number one Flight Attendant cannot be understated. He was instrumental in getting me out of the left seat and into the jump seat; summoning medical help; providing a walk-around bottle; etc. Good CRM of the entire crew was paramount to the positive outcome of this event.

Narrative 4000 (ACN 1044952) and 3879 (ACN 1040480). Respective primary-causes: ‘Human Factors’ and ‘Aircraft’.

The First Officer lifted his backpack which was on the floor behind his seat. His bag strap caught onto the DC Standby Bus CB and there was a momentary power interruption in the DC standby bus. The subsequent result in the power interruption was the loss of the L ENG EEC normal mode. Per checklist and conference with Maintenance; both engines were placed in the EEC ALT MODE resulting in the loss of the autothrottles. Flight continued and landed with no further problems.

Cosine between this narrative and the following one was 0.593, which within the LSA document space represents a weak association, however this was the strongest available with the corpus and the documents were coded appropriately with a very different primary-cause.

The preflight Standby Power Test passed (no Bus ‘Off’ light) but generated a Standby Inverter Status message. The Status message could be erased; but it returned when the Standby Power Test was repeated. A check of the electrical EICAS Maintenance page showed no volts or frequencies on the Standby AC Bus during the test. In addition; components listed in the Flight Manual as being powered by the Standby AC Bus were unpowered during the test (e.g. the Captain’s flight instruments all had flags
because the left ADC was not being powered). Clearly we had a bad Standby Inverter. It was replaced and everything worked normally. Here’s the problem. Why did we not get a Bus ‘Off’ light or a Standby Bus ‘Off’ EICAS advisory during the test? The Flight Manual lists the components that are powered by the Standby AC Bus. One of these components is the Standby AC Bus ‘Off’ indication. This implies that the sensor to tell you the Standby AC Bus has no power is actually powered by that very same bus. If true; this means the warning will never work because it has no power. How ironic. The Flight Manual lists the same thing for the Standby DC Bus. If the plane is really wired this way; then looking for a bus off light or EICAS advisory during the Standby Power Test has no validity whatsoever. Our procedures regarding status messages during preflight allow us to press on if we can erase the status message. In this case; that would have led us departing with an inoperative Standby Inverter. Perhaps some status messages need a little more scrutiny; but unfortunately pilots are given very little information about the meanings of status messages. A table in the flight manual listing all status messages and their meanings would be quite helpful.

Narratives 3667 (ACN 1031966) and 1697 (ACN 969496). Respective primary-causes: ‘Ambiguous’ and ‘Aircraft’.

Ready for pushback; Flight Attendant from door 2L came to cockpit to indicate that door 2L was not arming properly; that the wedge was not coming out. I went to door 2L to investigate. Looking at door; moved lever to disarm and looked again. Opened door handle and it was apparent that door was not disarmed; door with assist opened partially. Slide was attached at bottom and partly out of pack. I tried to close door but slide then deployed. No one was hurt; no one else was involved in manipulating the door. Maintenance was called and it was determined that the slide would be replaced and flight to operate with a delay.

Cosine between this narrative and the following one was 0.862, which within the LSA document space represents a strong association, however the primary-causes were coded differently.

This was a charter flight. The aircraft was parked in the north lot. Airstairs were brought to the aircraft door at 1L. It appeared the person trying to open the door was having difficulty opening the door. I cracked the door. He still appeared to be having difficulty; so I gave the door a push. The person on the other side was still having difficulty opening the door. I soon saw why. The side of the slide pan was caught on the side of the aircraft door. As the person on the outside continued to pull the door open the slide pan opened and the slide fell out; but did not deploy. A Mechanic arrived to detach the slide from the door. He said the door was armed. The arming mechanism was stuck between arm and disarm and we were unable to put it in either the arm or the disarm position. I told him that I disarmed the door. After I disarmed it; I made the all call to disarm cross check and stand by for all call. As I made the all call I first made the announcement; ‘1L and 1R disarmed and cross checked.’ As I made the announcement I double checked each door. Then flight attendants at doors two and four confirmed their doors were disarmed and cross checked. This is my first charter flight and no one has made it clear who opens doors at what station. There should be some clarity here. While we were at this airport; those who met the flight waited to open the door. Then I was told (can’t remember if it was the Customer Service Representative or the GSC) that I had to open the door. The GSC and the Customer Service Coordinator were both in the galley when the incident occurred. Both said they saw me disarm the door. This past week while on a B757 after the passengers deplaned; a Mechanic came to door 2L with a screwdriver and he began pushing in the rubber tubing surrounding the door. Curious; I asked why he was doing that. He said he likes to double check. I asked if the slide pan could get caught on the door. He reluctantly said; ‘It could.’ My guess is this is an issue that has not occurred for the first time. I hope there will be a check on B757’s to make sure this does not happen again in the future. It is possible the next time a slide may in fact deploy.

Narratives 83Q (ACN 819565) and 2602 (ACN 991859). These documents were found by analysis of the cases that did not match following selection of cases with cosine values above 0.80. Respective primary-causes: ‘Ambiguous’ and ‘Airport’.

Closed to taxi via Taxiway C to Runway 22R. Turned left on Taxiway E instead of Taxiway ZA. Signage is not clear. The sign indicating an intersection turn to Runway 22R strongly suggests that the left turn should be at Taxiway E. Suggest a sign indicating that Taxiway E is the route to Runway 22L and marking the pavement. There were no conflicts and the somewhat longer taxi to Runway 22R was uneventful.

Cosine between this narrative and the following one was 0.892, which within the LSA document space represents a very strong association, however the primary-causes were coded differently.

The CHS airport taxiway signs at the Taxiway B and Taxiway C intersection are inadequate and/or missing. While moving southwest on Taxiway C toward Taxiway B; the only sign depicting the Taxiway B turn is placed past the Taxiway C/Taxiway B intersection. This placement past the intersection is nonstandard. The Taxiway B turn sign should be placed before the Taxiway C/Taxiway B intersection for standardization. While moving north on the ramp from the gates toward Taxiway B; there are no signs identifying Taxiway B. There should be a sign that identifies Taxiway B as the jet approaches Taxiway B from the ramp. I am not sure if Taxiway A is identified with a sign or not. We did not take that route out of the ramp area. While moving north on Taxiway B (unidentified taxiway) toward Taxiway C there is no Taxiway C turn sign (either direction). However; there is a good sign at the Taxiway B and Taxiway C intersection; which has an arrow pointing left toward Runway 3 and an arrow pointing right toward Runway 33. While moving northeast on Taxiway C toward Taxiway B; there are no signs identifying Taxiway B. There should be a Taxiway B turn sign along Taxiway C (northeast) located before the Taxiway B turn. Thanks for looking into the CHS taxi signs.

Narratives 901Q (ACN 836772) and 3812 (ACN 1037700). These documents were found by analysis of the cases that did not match following selection of cases with cosine values above 0.80. Respective primary-causes: ‘Human Factors’ and ‘Chart Or Publication’.

After the ground crew pushed us back from gate the aircraft was blocking the gate next to ours. An air carrier was going into that gate so we offered to move forward to accommodate them. Ramp control gave us taxi instructions to pull straight forward and taxi to Spot 3. After verifying where Spot 3 was; the Captain started to taxi. It appeared that the only way to get to Spot 3 was to taxi onto Kilo. The Captain and I figured that Spot 3 was on Kilo. However; when we got onto Kilo I called ground and told them we were on Kilo and that was when they told us that we were in violation of the FARS. The event occurred because the Captain and I were under the impression that ramp control wanted us to taxi down to the Spot 3 near the de-icing pad. Which is the de-icing pad Spot 3 not the Spot 3 where they put it in either the arm or the disarm position.
wanted us to be. I realized after the fact that on the DTW commercial chart north ramp page that the Spot 3 we should have been going to was right on Echo. I think Spots 3; 4; 5; 6; should be depicted on the DTW airport chart instead of just being on the DTW north ramp page. The lines on Echo need to be painted better. The taxi instructions need to be more specific.

Cosine between this narrative and the following one was 0.853, which within the LSA document space represents a very strong association, however the primary-causes were coded differently.

We were told by LAS Ramp Controller to taxi to Spot 7. When I looked at the Commercial chart for Concourse D; I located spot seven and began a taxi to what I thought was that spot. As I arrived at the spot it was in fact spot 8. If you look at the chart you will see spot 8; not next to spot seven (where in fact it is) but half way down the chart. When I tried to call ramp to inform them of my mistake I noticed the First Officer had already switched to Ground and we then received a runway change and amendment to our clearance. I continued the taxi as I felt this was the safest course of action.

Narratives 902Q (ACN 836773) and 338 (ACN 935283). These documents were found by analysis of the cases that did not match following selection of cases with cosine values above 0.80. Respective primary-causes: ‘Procedure’ and ‘Aircraft’.

‘We made a precautionary return to departure airport due to low oil quantity. Upon preflight inspection of the aircraft; we noticed the quantity was 8 quarts in the left engine and 9 quarts in the right. Captain called maintenance to see if we could get some oil added and they advised the quantity was sufficient to make it to destination. Shortly after takeoff; we noticed 2 quarts displayed in the left and 1 quart in the right. We leveled off according to procedure; reduced the power and advised the tower of a possible return. The left engine was now displaying 1 quart (matching the right) and after 30 s; we decided to return to the field. Our landing weight was under gross and we had normal indications for oil temperature and pressure. Upon return to the gate; the Captain called company and maintenance arrived; adding 9 quarts of oil to the left engine and 8 quarts to the right. The aircraft landed the previous evening with 6 quarts of oil in the left engine and 7 in the right. Had oil been added; this event would not have occurred.

Cosine between this narrative and the following one was 0.869, which within the LSA document space represents a very strong association, however the primary-causes were coded differently.

Our layover was extended due to the fact that the inbound aircraft had an in flight engine shutdown due to low oil quantity and low oil pressure. Mechanics were flown into work on the aircraft overnight. We were told the aircraft would be ready the following morning and we would fly the aircraft out. During the preflight we noted that both engines had 22 quarts of oil. We had normal takeoff and climb to FL350. During the initial cruise portion of the flight we noted that both engine oil quantities had stabilized at about 12 quarts. As the flight progressed we noticed the right engine oil quantity beginning to decrease. It slowly went from 12 to 10 to 9 to 8 to 6 quarts. At this point the Captain and I became concerned about the oil loss and decided the safest action would be to return to the departure airport. The communications with Dispatch and Maintenance Control were very difficult through commercial radio. We ended up relaying most of our information through a company flight that could use their ACARS. As we returned to the airport the oil quantity in the right engine continued to decrease to 4 quarts. We had all available checklists out and were prepared for an engine shutdown if necessary. We did not declare and emergency and discussed continuing to another airport for passenger convenience. As the oil quantity continued to decrease we decided our departure airport would be a better decision. The right engine oil pressure and temperature remained steady and we did not shut down an engine. We had an uneventful approach and landing into our departure airport. After landing and shutdown we noted the right engine had 13 quarts of oil remaining. In the course of about two and a half hours of flight we had lost over 9 quarts of oil. We had a Company engine specialist on our flight. He told us that this engine had a long history of problems with it. I don’t feel the proper maintenance was done on this engine. Also communication with Dispatch and Maintenance were very difficult; almost impossible via HF phone patch. We were then instructed to deadhead home on the next flight.

Narratives 1082Q (ACN 839745) and 3667 (ACN 1031966). These documents were found by analysis of the cases that did not match following selection of cases with cosine values above 0.80. Respective primary-causes: ‘Human Factors’ and ‘Ambiguous’.

After arrival at gate; the Flight Attendant disarmed door 2R and then proceeded to door 2L where he began opening the door without disarming it. Realizing his mistake; he attempted to disarm the door; but the gate agent outside the door began trying to open the door resisting the flight attendants attempt to close and disarm the door. The slide pack fell onto the cabin floor but did not inflate. Company personnel were summoned and took control of the situation. As Captain; I was still in the cockpit finishing the last of my cleanup procedures and was unaware of the events as no one notified me. I became aware of the situation only when going to the door 2L area where I became aware of what was going on.

Cosine between this narrative and the following one was 0.862, which within the LSA document space represents a very strong association, however the primary-causes were coded differently.

Ready for pushback; Flight Attendant from door 2L came to cockpit to indicate that door 2L was not arming properly; that the wedge was not coming out. [I] went to door 2L to investigate. Looking at door; moved lever to disarm and looked again. Opened door handle and it was apparent that door was not disarmed; door with assist opened partially. Slide was attached at bottom and partly out of pack. I tried to close door but slide then deployed. No one was hurt; no one else was involved in manipulating the door. Maintenance was called and it was determined that the slide would be replaced and flight to operate with a delay.

Appendix B. Topics generated by the analysis

Table B.4 shows the terms and weights for the first four (of 400) topics developed by the LSA. Topic 1 contains many common terms, but is strongly associated with takeoff and landing events in the context of mechanical issues. This is the most generic topic within the document corpus. Topic 2 unlike topic 1 demonstrates negative weightings for terms within the topic. Terms with negative weightings are those which are penalized when present in a document. Thus the negatively weighted terms ‘engine’ and ‘gear’ suggest that topic 2 is not associated with mechanical issues, rather with approach and clearance problems. Topic 3 highlights issues with configuration during approach and landing. Topic 4 represents higher altitude events since terms that represent low
Table B.4
LSA topics produced in the reduced semantic space for the ASRS narratives database sample.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Weight</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.142</td>
<td>engine</td>
</tr>
<tr>
<td>2</td>
<td>0.126</td>
<td>runway</td>
</tr>
<tr>
<td>3</td>
<td>0.116</td>
<td>captain</td>
</tr>
<tr>
<td>4</td>
<td>0.115</td>
<td>ft</td>
</tr>
</tbody>
</table>

Altitude is negatively weighted. Positively weighted are terms associated with mechanical issues such as engine and pack failures.

**Appendix C. Stop words used to filter the dictionary**

- 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'again', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', 'should', 'now'.

**References**


