7-9-2009

The wrong Wright stuff : mapping human error in aviation

Stacey Hendrickson

Stacey Hendrickson

Candidate

Psychology

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Timothy E Goldsmith, Chairperson

Harold Delaney

Rick Johnson

Aileen Oliver

THE WRONG WRIGHT STUFF:

MAPPING HUMAN ERROR IN AVIATION

BY

STACEY M. L. HENDRICKSON

B.S., Psychology & Biology, University of New Mexico, 1998
M.B.A., Operations Management & Marketing,
University of New Mexico, 2003
M.S., Psychology, University of New Mexico, 2004

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

Psychology

The University of New Mexico
Albuquerque, New Mexico

May, 2009

DEDICATION

To my Grandma Audie: she taught me to tie my shoes, and so much more…

ACKNOWLEDGMENTS

There are many people who have contributed to the success of this project. There are, however, a few people who deserve special recognition for their contributions.

Dr. Tim Goldsmith, my advisor and dissertation chair, has been a great mentor and teacher. His guidance and kindness have helped me to succeed in school and in life.

Dr. Harold Delaney helped me through all stages of my college education. His guidance and teaching continue to help me grow as a researcher and statistician.

The other members of my committee, Dr. Peder Johnson and Dr. Caren Wenner, contributed thoughtful comments that made this a much stronger project.

Robert Larranaga responded at a moment's notice and helped this document reach finality.

Wendy Johnson helped me to push through writer's block on this project and many others. Our houses were decorated and our wardrobes filled!

My family always lifted me up when I have felt down and kept me grounded when my head was too far in the clouds.

Finally, to my husband, Gerald, I love you deeply. We have suffered through our dissertations together, and I know there is no one else I would rather have by my side. You have been the spark that has kept me going and the taskmaster when I was ready to give in. Thank you for all you are and for all you do. Because of you, "I am rooted in the me that is on this adventure!"

THE WRONG WRIGHT STUFF:
MAPPING HUMAN ERROR IN AVIATION

BY

STACEY M. L. HENDRICKSON

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

Psychology

The University of New Mexico
Albuquerque, New Mexico

May, 2009

THE WRONG WRIGHT STUFF:
MAPPING HUMAN ERROR IN AVIATION

by

Stacey M. L. Hendrickson

B.S., Psychology & Biology, University of New Mexico, 1998
M.B.A., Operations Management & Marketing, University of New Mexico, 2003
M.S., Psychology, University of New Mexico, 2004
Ph.D., Psychology, University of New Mexico, 2009

ABSTRACT

The Aviation Safety Reporting System (ASRS) was instituted to aid the Federal Aviation Administration in tracking trends in aviation incidents so that, ultimately, safety measures and training could be implemented to decrease the occurrence of accidents and incidents within the industry. The current system relies on hand coding of reports to recognize current trends and alert the proper parties. Although the filing party may enter some codified data describing the surrounding scenario (e.g., time of day, weather), there is no opportunity to specify a category if the problem is human error. Considering the prevalence of human error within these incidents (around 55% based on a report by Boeing, 2006), a greater understanding of the driving factors is needed.

The current study was an investigation of the human error components of airline incident reports. Text analysis tools were applied to ASRS incident narrative reports to

determine a classification based on human performance for commercial and general aviation. The results from the current study demonstrate that an empirically based approach can be used to uncover latent categories within the "Flight Crew Human Performance" classified reports. The combined approach of latent semantic analysis, *k*-means clustering, and keyword analysis were used successfully in developing a nine element classification of commercial aviation reports and twelve element classification of general aviation reports.

The taxonomies suggested by the current study for both commercial and general aviation reveal categories beyond just human error elements. The classification scheme suggested for the commercial aviation reports most closely resembled the ACCERS taxonomy developed by Krokos and Baker (2005; see also Baker & Krokos, 2007), which was constructed to help in categorizing all incident reports. The classification suggested for general aviation reports did not closely resemble any existing classification scheme. Although the suggested taxonomy shared categories such as situational awareness and communication with classifications such as crew resource management (CRM) or single pilot resource management (SRM), the current classification also holds non-human elements such as weather and context.

The taxonomies for both commercial and general aviation revealed a category for context, and the difficulty of flying into certain airports was apparent. These findings can be implemented to improve training programs by assisting in the creation of contextually based training scenarios. Furthermore, based on findings for general aviation in particular, pilots could benefit from increased training in situational awareness and monitoring of notices and airspace.

Table of Contents

List of Figures

List of Tables

Introduction

The Wright brothers succeeded in completing the first human flight in 1903. In the century since that eventful flight, air travel obviously has grown in importance. By the year 2004, U.S. air carriers were logging more than 8 billion miles of flight (NTSB, 2007). However, not all of these flights have come to a successful end. As air flight has grown in availability and popularity, the need to control and understand flight accidents has also grown.

As aviation systems have become more reliable and capable, they have become more complex. This complexity challenges the human in the system in his or her ability to interact with and control the system and operate it error free. Within the world of flight, one mistake or slip can cost money and, more importantly, lives.

Given the endemic nature of human error, there are two options to coping with it. The first alternative is to design it out of the system completely. The second option is to design the system to a level so that the occurrence and impact of the errors are minimized.

Although efforts to automate the human out of the system are underway, the day when there will be no human involvement within these systems is in the distant future. For instance, in the industry of aviation, although closer now to eliminating the need for human involvement within air traffic control, it is still highly unlikely the pilot will be completely removed from the cockpit. Even the use of auto-pilot requires much input from the human pilot.

For now, the inevitability of the human element within the system must be accepted and methods devised for dealing with its inclusion. The current study explored

the issue of human error within aviation. Concurrent with this exploration, methods for detecting human error, classifying it, and decreasing the occurrence through better training are discussed.

Human Error in Aviation

From its humble beginnings on the field in Kitty Hawk, North Carolina, flight continues to become safer with each passing year. However, there continue to be serious, and often fatal, accidents reported on the news. These accidents are rarely due to mechanical failure, but rather a failing on the part of the human operator. Boeing (2006) reported that between the years 1996 and 2005, 55% of all hull-loss accidents[1] were due to the fault of the flight crew. This compares to a total of only 17% due to the aircraft.

This estimate has come as a shock to some who believed the increase in the amount of automaticity within the cockpit would decrease the amount of human error. However, with the increase in automaticity has come a change in the types of errors flight crews make with a greater number of errors involving the use, or misuse, of the plane's automated systems (Kern, 2001). The errors that occur through the misuse, misunderstanding, or lack of familiarity of new equipment, account for only a portion of the mistakes made by pilots and crews within the cockpit.

Errors by humans can be defined in three ways: slips, lapses, and mistakes (Norman, 1981; Reason, 1990). Slips, lapses, and mistakes occur at different stages of the conception and execution of plans. Slips and lapses refer to those errors that come

---

[1] Boeing (2006) defines hull-loss accidents as any accident in which substantial damage that is beyond economic repair results to the aircraft. These accidents include those in which the aircraft if missing or when the aircraft is seriously damaged or inaccessible.

about from faulty execution of some action. They occur due to execution failures. Reason further distinguished between slips and lapses stating, “Whereas slips are potentially observable as externalized actions-not-as-planned (slips of the tongue, slips of the pen, slips of action), the term lapse is generally reserved for more covert error forms, largely involving failures of memory, that do not necessarily manifest themselves in actual behavior and may only be apparent to the person who experiences them” (pg. 9). Mistakes, on the other hand, occur through the misapplication of some plan. In a mistake, even though the execution of the plan may be perfect, the result from the execution is not what was originally intended. Slips and lapses are typically less complex and easier to detect. For this reason, mistakes often constitute a much greater danger and may go unnoticed for a longer period of time.

Within complex systems such as aviation, mistakes often do not occur through the faulty actions of one person. In most situations it is the accumulation of many smaller slips, lapses, and mistakes that finally results in the bigger error. This phenomenon, known as the Swiss cheese model, was first proposed by Reason in 1990. Figure 1 demonstrates how the smaller errors can build up to let a bigger error occur. This occurrence argues for the potentially severe consequences of allowing small mistakes, lapses, and slips to go uncorrected.

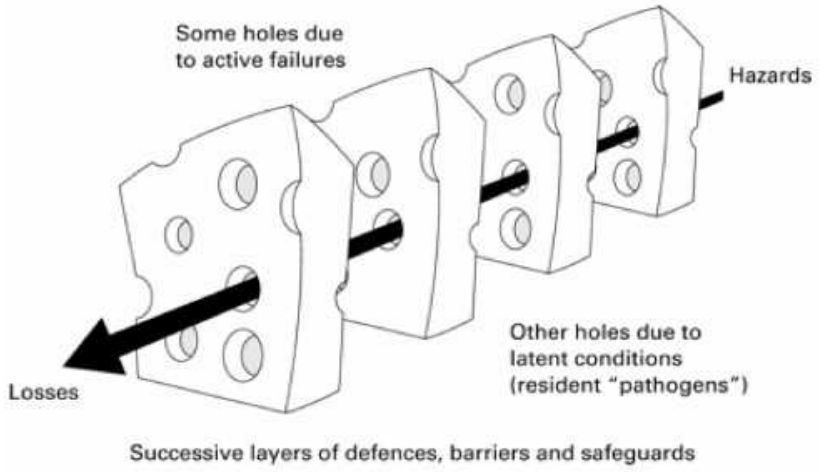

Figure 1. "Swiss cheese" model proposed by Reason (1990; depiction from 2001) to account for compounding errors.

A model of threat and error reduction proposed by Helmreich, Wilhelm, Klinect, and Merritt (2001) attempted to assuage these compounding slips and mistakes to eventually eliminate, or at least identify, the problem. Their proposed error management was an approach to limit the occurrence and impact of human error through better design of the system. It uses all available data to better understand the drivers of error and prevent it through a combination of options such as training, policy, and procedures (Helmreich, 1998). Helmreich et al.'s model classified the errors made by flight crews into five categories: intentional non-compliance, procedural, communication, proficiency, and operational decision.

The smaller mistakes that compound to form a much larger accident are by themselves not truly dangerous. However, in their combined form, they can be fatal. The investigation into some of the most deadly aviation accidents is time consuming and often fruitless. By the time the accident has occurred, the precipitating factors are often

so far buried it is impossible to understand the full extent of what happened. It is hoped that by looking more closely at incident reports instead of accidents, it will be easier to determine the precipitating factors and not be buried by all the elements of the resulting disaster.

Aviation Accidents and Incidents

ASRS incident reports. The Federal Aviation Administration (FAA) recognized the significance of this issue and began an effort in 1975 to monitor minor aviation mishaps. Due to this recognition, a collaborative effort was undertaken between the FAA and the National Aeronautics and Space Administration (NASA) to collect incident reports within the airline industry. The repository for these reports is maintained by NASA and is called the Aviation Safety Reporting System (ASRS). The primary purpose of ASRS is to identify problem areas and deficiencies within the aviation industry and respond with appropriate remedies. The ultimate goal is to reduce the number and severity of aviation accidents (ASRS Program Overview, n.d.). See Appendix A for an example ASRS report.

All the personnel involved in aviation operations (e.g., pilots, air traffic controllers, flight attendants, mechanics, ground crew) are encouraged to submit reports to ASRS for an unsafe incident in which they were involved or witnessed. The submission of the report is voluntary and confidential. The FAA will not take punitive action against the filer or punish the unintentional violation of statutes and regulations

reported through ASRS.[2] This rule allows airline personnel to feel more comfortable in filing a report without fear of retribution. In this way the FAA is able to keep a more accurate count of accidents and incidents (ASRS Program Overview, n.d.).

By June 2006, more than 700,000 reports had been submitted (About ASRS Data, n.d.). In addition to commercial aviation personnel, general aviation pilots also file reports. Aviation safety experts analyze all the reports by having two analysts read through each report and identify safety themes. Once aviation hazards are identified, they are flagged and the appropriate FAA office is alerted so that they can respond promptly. The analysts also classify the reports based on the underlying cause(s) for the incident. The classification of the incident along with any notes the analyst adds to the submitted report are then incorporated into the ASRS report (ASRS Program Overview, n.d.).

One of the key pieces of information included in the original report is a narrative describing the event. This narrative is rich in information about the incident and allows the person to describe the scenario, the events leading up to the incident, his or her reaction, and his or her recommendation on how to avoid a similar occurrence in the future (ASRS Program Overview, n.d.). These narratives are generally 100-200 words in length, but some may be substantially longer or shorter. It is primarily through this narrative description that the analyst classifies the incident using expert judgment.

There is no standard or systematic method for classifying the ASRS reports. Upon entering a report within the ASRS system, the participant is asked to record some

---

[2] There are exceptions to this provision in cases where a deliberate or particularly egregious violation of procedure has occurred.

preliminary information to help describe the incident and surrounding context. The information requested covers items such as: time, location, weather, role of pilot flying, and position of reporter. A copy of the reporting form is contained in Appendix B ("ASRS General Reporting Form", 1994).

Beyond this classification by environment, the only other classification of the reports is done by the analysts at the Aviation Safety Information Analysis and Sharing (ASIAS) safety office of the FAA. A representative from this office (James Fee, personal communication) explained when researchers make queries about a certain type of report (e.g., landing trouble or weather problems), the analysts often use keyword searches to pull relevant reports from the full database. The analyst will then read through the reports retrieved to ensure they address the researcher's original request. This system of pulling reports is inefficient and likely presents an incomplete picture due to omission of relevant articles not found with the keyword search.

Crew resource management in commercial aviation. An initiative to control the occurrence of flight accidents brought about by human action was begun with the incorporation of crew resource management (CRM) within training programs in the late 1970's (Kern, 2001). A large part of working effectively, and error free, within a complex system is understanding how to manage one's own situation in a team environment. Skills in both cognitive and social areas are identified and trained in CRM programs. These programs focus on the human element within complex systems and attempt to better prepare the person to cope in stressful situations.

CRM skills are typically categorized as cognitive or social (Flin & Martin, 2001). Although the specific labels differ across research setting and airlines, the concepts are

fairly consistent and include categories such as: decision-making, situational awareness, workload management, leadership (or "followership"), communication, and teamwork. Drawing from the CRM taxonomy, Lauber (1993) sought to classify the errors seen in aviation within a CRM framework. His system classifies errors into seven categories: preoccupation with minor mechanical problems, leadership, delegation of tasks, setting priorities, monitoring, effectively using available data, and effective communication.

Training in CRM has met with success in reducing human errors (Diehl, 2001). However, training programs lack standardization across industries and, within aviation, across carriers (Salas, Wilson, Burke, Wightman, & Howse, 2006). Most airlines have developed their own classification schemes making it difficult to compare ASRS reports across carriers.

Crew resource management in general aviation. In contrast to commercial aviation, general aviation (GA) pilots must deal with a different set of issues. In fact, CRM concepts and proper training might prove even more helpful in the GA setting as it is 20 times more hazardous than commercial aviation (Kern, 2001). However, the CRM taxonomy must be restructured because, unlike commercial aviation where pilots are regularly part of a flight crew of two or more members, within GA pilots often fly alone. Therefore, many of the CRM concepts important in commercial aviation (e.g., teamwork, leadership) are irrelevant in this setting.

Given the single pilot environment common within GA, CRM concepts within this setting are often termed single pilot resource management (SRM). SRM is the use of all resources available to the pilot (on-board and off the aircraft) during and before the flight in order to achieve a safe flight (Kern, 2001; Summers, Ayers, Connolly, &

Robertson, 2007). These resources include all hardware, software, and liveware[3] options. "For example, a non-pilot can help scan for traffic and arrange charts, Flight Watch can keep the pilot updated on changing weather, Flight Following provides radar services, and full use of the autopilot (if installed) may free the pilot to perform other cockpit duties" (Glista, 2004, p. 8). Table 1 lists the SRM elements described by Summers et al. (2007).

Table 1. *Definition of single pilot resource management (SRM) elements.*

SRM Element	Definition
Aeronautical Decision Making (ADM)	Consistently making timely, appropriate and informed decisions regarding the current task.
Risk Management (RM)	Having knowledge of the purpose of all available resources and using them appropriately.
Task Management (TM)	Similar to workload management, it is the appropriate prioritizing of the tasks at hand.
Automation Management (AM)	Having knowledge of and appropriately programming and using the modes of cockpit automation.
Controlled Flight into Terrain (CFIT) Awareness	Understanding and applying techniques to avoid CFIT encounters (especially during instrument rated flights).
Situational Awareness (SA)	Having awareness of and responding appropriately to all factors of the flight (e.g., traffic, weather, fuel state, aircraft mechanical condition, and pilot fatigue level).

[3] Liveware is defined as the other people (e.g., ATC, ground crew, passengers) available to help aid the pilot in operating the systems of the aircraft.

Restructuring the training programs for GA pilots is important to combat the increased risk associated with general aviation. One suggested change is to replace the current training programs that focus on training "stick and rudder" skills in isolation with a scenario-based training program similar to that used in commercial aviation (Wright, 2004). Glista (2004) suggested the current training overlooks the major causes of GA fatal accidents, which are a lack of situational awareness, risk assessment or management, and poor aeronautical decision-making. Increasing awareness of these skills as well as standardizing their conceptualization in training programs should help to decrease their negative impact in GA flights.

Classifying ASRS reports. Evidence suggests training in CRM (or SRM) programs assists the pilot in reducing the mistakes that are made overall (Diehl, 2001). However, a clear definition of the specific mistakes is often missing. A richer context for CRM training could be accomplished with a better understanding of the mistakes being made by the flight crew or single pilot.

Failings in CRM skills are often cited in NTSB investigative reports following accidents (Kayten, 1993). To aid in the avoidance of accidents, these failures in CRM skills should be investigated during incidents as well. Although human error in general may be found to be a contributing factor in an incident, a detailed account or listing of deficient CRM skills is lacking. A classification of the human factors elements present in the ASRS reports is needed for a better understanding of the initiating factors. Furthermore, to aid in the better understanding of CRM skills and eventually to obtain a standardized classification of these skills, a link should be established between the human

factors categorization of the ASRS reports and the future development of CRM programs.

Beaubien and Baker (2002), in a review of various aviation incident reporting systems currently being used, cited a weakness of ASRS as being that most of the information collected is in text format and reason codes do not exist for coding the reports. Similarly, the reporting system used by the United Kingdom (Confidential Human Factors Incident Reporting Programme [CHIRP]) and that used by Australia (Confidential Aviation Incident Reporting [CAIR]) seem to suffer the same problem and lack a standard and validated taxonomy for human error events. Including reason codes for classifying the reports would help analysts sort through the thousands of reports and organize them into meaningful categories for use in training or research. The current system of having the analyst cull through so many reports is tedious and inefficient. A few taxonomies and classification systems have been developed that may aid in the sorting of these incident reports.

Of particular interest to the current research project are classification systems that focus on the human error within accidents and incidents. Shappell and Wiegmann (1997, 1998, 1999) developed the Human Factors Classification System (HFACS) in an attempt to describe the holes in Reason's (1990) "Swiss cheese" model allowing airline accidents to occur. HFACS describes four areas of concern in which failures may combine to cause an accident: organizational influences, unsafe supervision, preconditions for unsafe acts, and unsafe acts. Figure 2 displays their conceptualization of these four areas represented in Reason's (1990) "Swiss cheese" model.

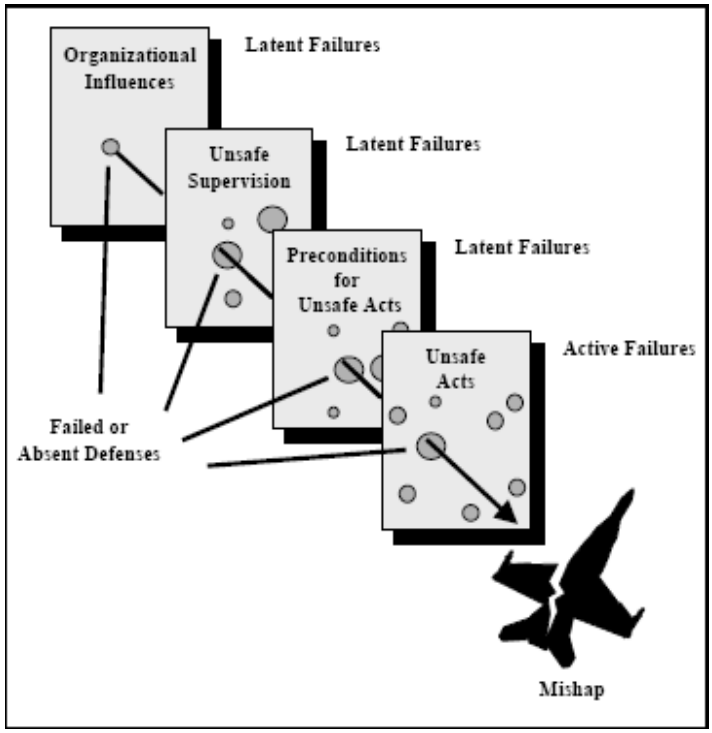

Figure 2. Shappell and Wiegmann's (2000) adaptation of Reason's (1990) "Swiss cheese" model.

More recently, Krokos and Baker (2005; see also Baker & Krokos, 2007) described a system developed to help classify reports received through the Aviation Safety Action Program (ASAP[4]). The classification system, titled Aviation Causal Contributors for Error Reporting Systems (ACCERS), was developed to classify the reports based on the underlying pilot error, but was to be used in the classification of all reported incidents. Krokos and Baker developed an initial categorization consisting of nine categories by reviewing any existing taxonomies and then enlisting the aid of subject matter experts to determine final category labels. This list was collapsed to seven items

[4] ASAP is similar in nature to ASRS, but is airline specific.

based on the expert knowledge of three senior pilots and a review by senior-level pilots (2007). Table 2 shows these two final levels of categories.

Table 2. *Final Classification Stages in the Development of ACCERS.*

Initial 9 Category Solution	**Revised 7 Category Solution**
Procedural Issues or Deviations	Policies or Procedures
Error Made by Other People Pilot Error Weight and Balance Issues	Human Error
CRM or Physiological Factors	Human Factors
Organizational Factors	Organizational Factors
Equipment Limitations or Failures	Hardware
Weather	Weather or Environment
	Airspace or Air Traffic Control
Unexplained Events	

The current conceptualization of ACCERS or the HFACS classification system asks the reporting employee to classify the report based on these proposed categories at the time of filing. Therefore, to be used effectively the classification systems should be integrated into the filing system from the inception. At this time, no such classification has been initiated within ASRS reporting. Therefore, it remains to be seen if these classification systems can be used to classify older reports already entered in the system.

The lack of a standard classification scheme available for sorting the numerous incident reports submitted to ASRS has led to difficulties in analyzing the incidents.

There is great depth of information available within these reports, and yet it has remained largely untapped. Queries made by airlines to NASA requesting reports representative of a trend of interest are usually answered with an assortment of reports produced through a simple keyword search (James Fee, personal communication). This method of retrieval not only fails to capture all the relevant reports to a topic, it is also unreliable in its results and requires a large investment of the analyst's time to peruse the documents to ensure they meet the needs of the inquirer.

Furthermore, the airlines are at a loss as how to fully evaluate and analyze the reports they are presented with. Although much emphasis on incidents and accidents is placed on what happened, a more important question is why it happened. The text narrative within the incident reports may contain the answer to this question. This information is invaluable to strengthening training programs and making flight even more reliable than it already is. To aid analysts and airlines in the analysis of these narratives, a classification scheme for the human error components of these incidents is needed. A major goal of the current project is to derive such a classification system.

Solving the problem of appropriately classifying past reports (and continuing forward for future reports) is twofold. First, an appropriate classification system must be determined. Second, an efficient method for culling through the many reports already collected must be presented and then these reports need to be classified appropriately.

Text analysis is an attractive option for studying the airline incident reports because of the wealth of information contained in the narratives. The narrative data contained in the ASRS incident reports is rich with contextual cues and information regarding the part played by human error in incidents and near accidents. Through a

careful consideration and examination of these documents, a clearer picture can be formed regarding how important a role human error plays. This information could be used in constructing training exercises and define emerging trends of reckless or unnoticed behavior that need to be corrected. One method proposed for tackling this problem is the text analysis methodology of latent semantic analysis.

Latent Semantic Analysis

Definition of latent semantic analysis. Latent semantic analysis (LSA) was initially proposed in the area of information retrieval. It is often referenced in this capacity as latent semantic indexing (LSI; see for example Kolda & O'Leary, 1998; Letsche & Berry, 1997). As a tool for information retrieval, LSA has compared very favorably against more traditional methods of vector retrieval approaches such as that proposed by Salton and McGill (1983 discussed in Dumais, 2003).

In vector retrieval approaches, the unique terms present in a collection of documents represent the axes or dimensions in a multidimensional space. The documents are represented as vectors within this space. Document retrieval is accomplished by calculating the cosine similarity between a probe document and a test document. Because each document is represented by a vector of numbers that code purely for the existence of words, a serious failing of this method is the exclusion of documents that may be semantically related but do not include words contained in the probe document. The restriction created by the term dimensions being orthogonal to each other keeps synonyms orthogonal and independent of each other. Therefore, the search for one word (e.g., doctor) will not retrieve documents including only a synonym (e.g., physician).

The success of LSA in information retrieval, and other applications that will be discussed shortly, is largely due to its ability to deal with variability in word usage and discern polysemous and synonymous words. A polysemous word is one that has two or more separate definitions. *Play* is a polysemous word as you may watch children *play* at recess, or you may go to the theater to watch a *play*. Synonymous words are different words that share a similar meaning, such as *doctor* and *physician*.

LSA is able to discriminate between polysemous words by observing the co-occurrence of terms around them. For instance, LSA will recognize the word play occurring in two separate instances. The first will co-occur frequently with words such as *children, recess,* and *fun*. The second instance of *play* would co-occur with *theater, actor,* and *actress*. LSA gains an advantage in its ability to discriminate these uses of the term and adjusts similarities between documents accordingly.

Even though synonyms do not occur often together in a single document, they will co-occur with many of the same terms. For instance, although *doctor* and *physician* may not occur together, other terms such as, *hospital, nurse,* and *sick* will co-occur in the same documents. Due to this ability to discern synonyms, LSA is able to return high similarities between documents that do not share the key term, but have synonyms instead.

Similar to the vector retrieval method discussed earlier, LSA also represents the sample of documents and terms in a multidimensional space. However, LSA uses a dimension reduction technique that necessitates the number of dimensions to be less than the number of terms (or documents) available. This smaller dimensional space forces relationships to exist between terms. In this space, "LSA simultaneously models the

relationships among documents based on their constituent words, and the relationships between words based on their usage in similar documents" (Dumais, 2003, pg. 493). Information retrieval done on such a space returns similar documents that contain synonyms and is able to discriminate between documents containing polysemes.

Applications of latent semantic analysis. Because of the stated abilities of LSA, researchers have found it to have many applications outside of information retrieval. Landauer and colleagues stress LSA's capability to emulate language acquisition. Landauer and Dumais (1997) showed LSA to be quite accomplished at learning English and performing on the *Test of English as a Foreign Language* (TOEFL) when fed a large body of text. In this study, LSA was trained on 30,473 articles from the *Groliers Academic American Encyclopedia* and then tested on 80 synonym questions from the TOEFL. They found LSA compared favorably with a large sample of students from non-English speaking countries taking the TOEFL as an admission requirement to U.S. schools.

Landauer, Laham, and Foltz (2003) also posited LSA as a replacement for human graders in assessing essay exams. Essays were gathered from a wide range of educational abilities (fourth grade through medical school students) and a wide variety of topics (e.g., neural conduction, Freudian concepts, history of the Panama Canal). Landauer et al. found that LSA correlated as highly with human raters as the raters correlated with each other. The use of LSA in such a domain offers educators an automated method for grading essays.

LSA has also been used in research on emerging trend detection (ETD; Kontostathis, Holzman, & Pottenger, 2004). Kontostathis et al. used five collections of

documents (including four years of INSPEC scientific abstracts and a collection of OOSE [object-oriented software engineering] articles) that had previously been evaluated to determine truth sets. Truth sets are lists of emerging and non-emerging trends within a collection of documents and are created to serve as comparative bases for testing ETD methods. For these sets of documents investigated, Kontostathis et al. demonstrated that LSA facilitated the detection of around 92% of the emerging trends. The application of a dimensionality reduction technique such as singular value decomposition (SVD) allowed related terms to be identified and clustered appropriately. These clusters were then used to reveal emerging trends based on the inclusion of terms previously identified as emerging or non-emerging indicators as well as the replication of these constructs across time periods. A method for accurate and efficient emerging trend detection in which new and important themes and topics can be seen is important for all businesses that must monitor a particular field or topic area.

Finally, and perhaps most relevant to the present study, LSA has been applied to clustering documents; however, this application has been infrequent. Lerman (1999) demonstrated that applying hierarchical clustering to the reduced term space produced through the application of LSA was effective in correctly clustering documents. The results of clustering 1,000 documents (representing five evenly sized groups of TREC [Text REtrieval Conference] topics) following the application of LSA was contrasted to clustering documents displayed in a full term space (i.e., before the application of LSA). The clustering of the set of 1,000 documents following the application of LSA outperformed clustering the documents in the term space for all values of dimensionality except for the largest value tested of 500. Precision levels, which are defined as the ratio

of the number of correctly assigned documents to the size of the cluster, were used to compare the performance of the two methods. The precision levels ranged from 85% to 100% following the application of LSA with dimensionality values of 3 through 100. These precision values were compared to the precision values of 81% and 87% obtained for the term space clustering. When the dimensionality level was increased to 500 in LSA, the performance was slightly worse than clustering in the term space (level of precision of 77% and 84.5% compared to 81% and 87%, respectively).

Elsas (2005) also investigated the usefulness of applying LSA to clustering documents. He compared the performance of LSA to another dimensionality reduction technique, independent component analysis (ICA), in clustering a dataset composed of 11 groups of 1,000 mutually exclusive documents.[5] Although he hypothesized that ICA would outperform LSA given it is "specifically identifying dimensions that exhibit a more 'clusterable' characteristic" (pg. 24), there was no appreciable difference between the two methods at the lower dimensionalities evaluated (e.g., 10 dimensions).

Mechanics of latent semantic analysis. LSA is a complex mathematical construct that purports to garner semantic information from text. It uses SVD to discover meaning, recognize synonyms, discern homonyms, and relate higher order semantic relations within the text. All of this information is gleaned from examining the co-occurrence of terms within documents.

---

[5] The set of documents Elsas used were composed of World Wide Web (www) pages collected by Sinka and Corne (2002). Sinka and Corne collected the WWW pages in an attempt to generate a standard text collection for use in document clustering research. In creating this set of documents, the authors relied on the Open Directory Project (http://www.dmoz.org) and Yahoo! Categories (http://www.yahoo.com) for categories that had been created by human judgment.

LSA operates on text alone to gather its representations of the meanings of words. It accomplishes its goal of extracting structure from a set of words in a series of steps. As a first step, LSA converts a corpus of text into a term by document, *m* by *n*, matrix where the *m* rows represent unique terms within the text and the *n* columns represent the documents. Individual cells within the matrix are the frequency of occurrence of a term within the document. Documents may refer to an entire paper or smaller sections of a complete piece such as a paragraph or a single sentence (Landauer & Dumais, 1997).

In the next step, the term by document matrix is submitted to a preprocessing step prior to the calculation of SVD. The preprocessing step expresses the importance of each term (i.e., importance in ability to discriminate between documents) by applying a weighting function to each cell. A weighting scheme typically includes reference to both the term's local weight and its global weight. The local weight of a term addresses the frequency of the term within a document, whereas the global weight expresses the frequency of occurrence across all the documents.

Dumais (1991) explored various combinations of local and global weightings to discern which performed best in document retrieval. She explained that common forms of local weighting are: term frequency (how often the term occurs within the document), binary (zero if the term does not occur and one for any occurrence greater than zero), and log (of the term frequency plus one). Global weighting measures explored are shown in Table 3 and included: Normal, GfIdf (global frequency inverse document frequency), Idf (inverse document frequency), and Entropy. Within the formulas, the variables are defined as: tf_{ij} is the frequency of term *i* in document *j*, gf_i is the frequency of term *i* over

all the documents, df_i is the number of documents in which term *i* appears, *n* is the total number of documents, and $p_{ij} = \frac{tf_{ij}}{gf_i}$.

Table 3. *Global weighting schemes evaluated by Dumais (1991).*

Global Weighting Scheme	**Formula**
Normal	$\sqrt{\frac{1}{\sum_j tf_{ij}^2}}$
GfIdf	$\frac{gf_i}{df_i}$
Idf	$\log_2\left(\frac{n}{df_i}\right)+1$
Entropy	$1-\sum_j \frac{p_{ij}\log(p_{ij})}{\log(n)}$

The log-entropy weighting scheme, a combination of the local weight (log of [term frequency + 1]) and the general weight (Entropy) multiplied together, is one of the most common and was found by Dumais (1991) to be the most effective in information retrieval applications.

The purpose of the preprocessing step is to weight in importance those words that are the most discriminating between documents. Words that occur too often (referred to as 'stop' terms including, for instance, *it*, *the*, *and*, *is*) do not help in discriminating meaning between documents. Because these words occur with relatively the same frequency within all documents, looking merely at their occurrence tells nothing regarding the difference between the documents because all the documents look alike.

Similarly, those words occurring only in one or two documents are too limiting and also offer no assistance in discriminating between documents. Therefore, the weighting scheme minimizes the impact of the terms occurring too frequently or not frequently enough, and increases in importance those terms that reveal discrimination between some documents and commonalities in others. Following this preprocessing step, LSA reduces the dimensionality of the matrix using SVD.

SVD is a process similar to principal component analysis (PCA) and is used to reduce the dimensionality of a multidimensional space. A key distinction between PCA and SVD is that PCA analyzes objects and components separately, whereas SVD analyses both together. SVD is a matrix manipulation allowing for the reduction of dimensions and the transformation of a nonsymmetrical matrix into a symmetrical one. By reducing the dimensionality of a matrix, SVD purports to eliminate noise contained in the original matrix and capture the most important associations between the words and documents (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990).

It is easiest to understand how SVD works by first considering a square matrix. Any square matrix **M** can be broken down into three components: $\mathbf{M} = \mathbf{A} * \mathbf{E} * \mathbf{A}^{\mathbf{T}}$ where **E** is a diagonal matrix of the eigenvalues and **A** and $\mathbf{A}^{\mathbf{T}}$ represent the eigenvectors of the matrix **M**. SVD reduces the dimensionality of the space by eliminating a portion of the eigenvalues and associated eigenvectors in the matrices. Before elimination, the eigenvalues are arranged in order from highest to lowest. A certain percentage of the lowest eigenvalues and eigenvectors are eliminated, keeping only *k* dimensions. Thus, when the resultant $\mathbf{E}_{\mathbf{k}}$, $\mathbf{A}_{\mathbf{k}}$, and $\mathbf{A}^{\mathbf{T}}_{\mathbf{k}}$ matrices are multiplied together, an approximation of the original matrix **M** is obtained (Kintsch, 1998).

This idea can be extended to non-square matrices such as those encountered in LSA. In this instance, the matrix **X** can be represented as: $\mathbf{X} = \mathbf{T}\,\mathbf{\Sigma}\,\mathbf{D}^{\mathbf{T}}$ where **T** is a *t* x *r* orthogonal matrix, **D** is an *r* x *d* orthogonal matrix, and **Σ** is an *r* x *r* matrix. In this case, *t* represents the number of terms, *d* represents the number document, and *r* is the rank of the original matrix. The matrix **Σ** is a diagonal matrix in which the diagonal values are the singular values, similar to the eigenvalues of the **E** matrix mentioned earlier. Only *k* of the original *r* dimensions are retained in the reduction by SVD. The product of $\mathbf{T_k}$ $\mathbf{\Sigma_\kappa}\,\mathbf{D^T_k}$, after the reduction of lowest singular values, is the singular value decomposition of the matrix **X** (Landauer, Foltz, & Laham, 1998; Leon, 1998). Figure 3 presents a schematic of the reduction accomplished through SVD.

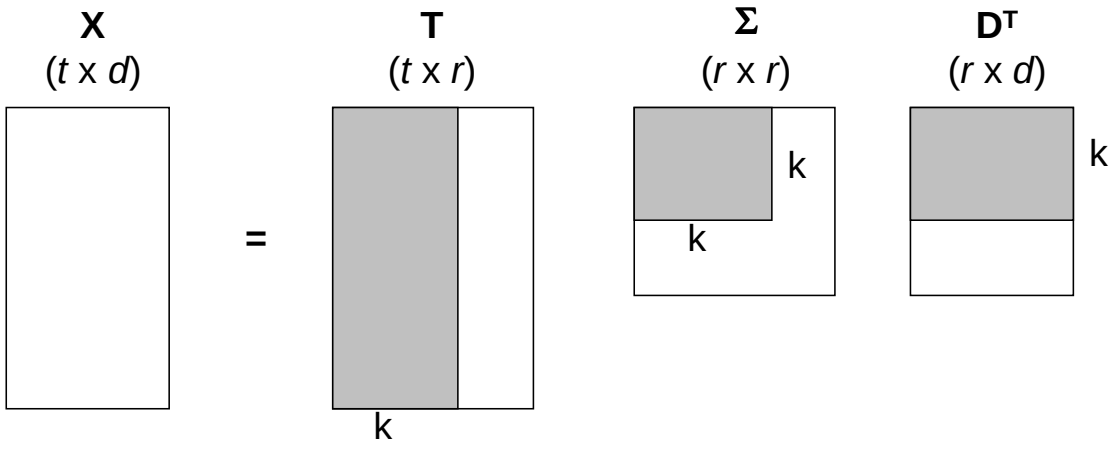

Figure 3. Diagram of SVD.

The number of retained singular values represents the dimensionality of the resultant matrix. Therefore, the words and documents can be represented as vectors in a multidimensional space. Landauer and colleagues often argue for the importance of maintaining 300 dimensions (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). However, the exact number of dimensions necessary for adequately representing

the meaning within a set of documents has been argued. For instance, Dumais (1991) stated that if the document sets are relatively homogenous, 100 dimensions are adequate for capturing meaning.

The number of dimensions is contingent on the dataset being examined. The number of terms available dictates, to some extent, the number of dimensions useful. Dumais (2003) sums the problem up nicely, "With too few dimensions, LSA performance is poor because there is not enough representational richness. With too many dimensions, performance decreases because LSA models the noise in the data thus reducing generalization accuracy" (pg. 497).

The reduced matrix produced through the application of SVD minimizes the "noise" or extraneous information from the original matrix and unveils the semantic structure of terms and documents. The components of this matrix represent the terms and documents in a multidimensional space in which similar items are near each other. The similarity between terms or between documents may be determined by measuring the angle between the vectors created in the multidimensional space (Martin & Berry, 2007). A common measure of similarity is that of cosine, but other measures such as Euclidean distance may also be calculated.

Clustering Documents

The output from LSA may be used to classify or cluster documents. The combination of SVD and clustering techniques offers a potentially powerful method to analyze and make sense of extremely large datasets. One difficulty in analyzing large datasets is their complexity and inclusion of often unnecessary detail that clouds interpretation. The application of SVD represents the knowledge in a more compact way.

This compaction may help to eliminate noise in the data and capture underlying regularities or structure in the data that might be obscured in its full form (Skillicorn, 2007). It is easier then to find defined clusters within this reduced form.

There are a variety of clustering algorithms available. The algorithms may be either hierarchical or partitional in the manner in which they carve up the dataset. Hierarchical methods give a complete hierarchical diagram of how objects are similar. Partitional methods, on the other hand, demonstrate how objects cluster at a single level (Skillicorn, 2007).

A widely-used partitional clustering method is *k*-means clustering. In its basic form, the *k*-means algorithm progresses by first randomly sorting the documents into *k* number of clusters around randomly chosen centroids (Johnson & Wichern, 2002, pg. 694). The mean distance from the group centroid for all documents within the cluster is calculated. The next step in the algorithm determines the proper members of the *k* clusters by proceeding through the documents and calculating all pairwise distances. A document is assigned to the cluster in which the distance between it and the other members of the cluster is a minimum. The centroid of that cluster is then recalculated including the newest member.

These steps are repeated until no further assignments can be made. This clustering algorithm has a drawback in that the formation of clusters is dependent upon the initial clustering and may be rather arbitrary (Willett, 1988). Therefore, it is advised to do repeated applications of the algorithm using different initial clusters (Kachigan, 1986; Kauffman, L & Rousseeuw, P. J., 2005; Skillicorn, D. 2007).

A second method of clustering, hierarchical clustering, can proceed in either an agglomerative or divisive manner. The agglomerative method begins with each object representing its own cluster and then progressively joining clusters until all objects are represented in one large group. The divisive method progresses in the opposite direction starting with all objects joined in one large group then dividing the cluster until eventually all objects are single member clusters (Hair & Black, 2004).

Clusters may be formed within hierarchical clustering a number of ways. Some of the most common methods are: single linkage, complete linkage, average linkage, and Ward's method. The distinguishing factor between these methods is the manner in which distance between an object and a cluster is calculated.

Single linkage technique is based on minimum distance and defines membership within a group on the nearest neighbor concept. Proceeding based on the nearest neighbor technique, two objects that are most similar, and not already within the same cluster, are joined within a cluster. Membership within a cluster is based on distance to only one other member within the cluster – the member to which it is closest. This clustering method may result in long chains of data because an object will be added to a cluster based only on another single member of the cluster. The resulting cluster may have little internal cohesion where the first and last object (or two ends of the cluster) may be very dissimilar to each other (Willett, 1988).

In contrast to the single linkage method, the complete linkage method bases membership within clusters on the farthest neighbor distance. This method tends to produce compact clusters, minimizing the distance between any two members of the clusters (Kaufman & Rousseeuw, 2005). An object is joined to a cluster when the

distance between the object and the farthest member to it in the cluster is minimum as compared to the farthest members of the other clusters. In other words, an object is assigned to the group in which the distance between it and the most dissimilar object to it in a cluster is less compared to any other cluster. Although this definition of membership avoids the chaining effect common with single linkage, it tends to be overly restrictive and creates many small tight clusters (Willett, 1988).

A third method, group average, helps avoid both problems encountered by single and complete linkage. In this method, the distance between an object and a cluster is based on a composite measure of the cluster. The composite measure is the average distance from the object to every other object within the cluster. The object is joined to the cluster for which it has the smallest average distance to all other members within the cluster.

Ward's method was developed to minimize the amount of information lost in cluster merging. Put another way, "the objective of Ward's method is to find at each stage those two groups whose fusion gives the minimum increase in the total within-group error sum of squares" (Gan, Ma, & Wu, 2007, p. 135). A drawback to Ward's method is it is only explicitly defined if the Euclidean distance is used to measure the similarity between objects (Willett, 1988).

The cophenetic correlation can be calculated to determine if the hierarchical clustering tree is a good representation of the data. This correlation is a Pearson product-moment correlation comparing the resulting hierarchical clustering tree to the similarity data matrix containing the distance or similarity measurements between the objects. Lack of agreement between the two representations results in a correlation value near

zero. Similarly, a correlation near 1.0 demonstrates concordance between the two representations (Romesburg, 2004).

Unlike *k*-means clustering, hierarchical clustering does not require parameters such as the number of clusters to be specified. This is especially helpful in data-mining applications in which the value for the number of clusters may not be readily apparent. Although the number of clusters is not determined *a priori*, the number of clusters to retain once hierarchical clustering has been applied still must be decided. Knowing at what point to stop the analysis is an issue of weighing the solution's structure against the clusters' homogeneity. The simplest structure is one large cluster, and the most homogenous is all individual clusters (Hair & Black, 2004).

Within hierarchical clustering, deciding where to stop the clustering can be referred to as *pruning the tree* because hierarchical clustering is often presented in a dendrogram that plots the formation of the clusters and resembles a tree. For instance, in agglomerative clustering, the tree progresses from the point in which all objects are represented as individual clusters, and steps along joining the smaller clusters until the final stage in which all objects are joined in a single cluster. Pruning of the tree cuts off the lower branches of the hierarchical tree, discarding those early steps in which the objects were individual clusters.

It may be that simply by viewing the dendrogram, a natural pruning point can be determined. For instance, Figure 4 depicts an example dendrogram that shows the hierarchical clustering of eight objects. The horizontal axis represents the objects to be clustered, and the vertical axis represents the distance between the objects or clusters. For the purposes of this example, the vertical axis displays distances (e.g., Euclidean

distance), but it could also display similarity between the objects (e.g., cosine). In Figure 4, the first step joins objects A and B, which are at a distance of 1.5 units from each other. Linkages continue to be made based on the distances between the objects and clusters until the final step (Step 4) in which the final two remaining clusters are joined to form a single cluster.

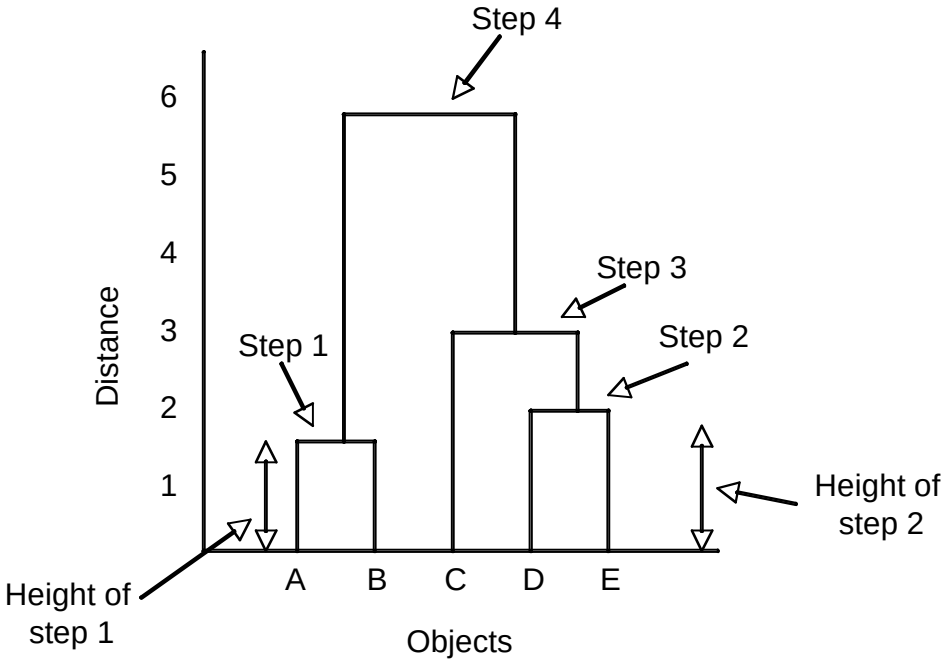

Figure 4. Dendrogram Graph.

The dendrogram in Figure 4 appears to have a clear pruning point at the distance of three units. The next step joining clusters (Step 4) is at a visually significantly greater height joining clusters at a height of 5.5 units. Visually, there is a natural break at the third step. If pruned at the height of three units (i.e., just after Step 3), two clusters are formed, the first cluster contains objects A and B and the second cluster contains objects

C, D, and E. However, hierarchical trees are often not this clear. Therefore, different pruning points may be set by the researcher based on the height of the links.

Some software programs, such as *MATLAB* (Mathworks, 2007), allow the calculation of the inconsistency coefficient, defined as the difference between the height of the current link compared to the average height of all the links below it. If the link being evaluated is not necessary, the inconsistency coefficient will be large, signifying the link is inconsistent with the other links formed. Conversely, a low inconsistency coefficient means the link is consistent with the other links. A value for the inconsistency coefficient may be specified by the researcher as a cut-off point for pruning the tree. If a large distinction between inconsistency coefficients can not be determined such that a clear pruning point is evident, the researcher will need to examine the clusters resulting from different pruning points and make a qualitative judgment as to the best representation of the data.

For both hierarchical clustering as well as partitional methods, it may be necessary for the researcher to judge the goodness-of-cluster across various clustering outputs. One method for judging which clustering output is superior is to measure the within to between cluster variability. A within to between (WB) ratio may be calculated to compare the within cluster cohesion to between cluster similarity. Specifically, the WB ratio is calculated as the average distance of all points within the same cluster divided by the average distance of all points across clusters. One would expect this number to be less than 1.0 based on the expectation that the average distance between points within a cluster should be less than the average distance across clusters. If the

value is equal to or greater than 1.0, the points are not well represented within the current clustering scheme.

Another method for evaluating goodness-of-cluster is through silhouette plots and the average silhouette value. The average silhouette value is calculated by comparing the placement of each object within its cluster to how well it would fit in the nearest cluster (Rousseeuw, 1987). Specifically, the average distance is computed for each object to every other object within its same cluster, and these distances are averaged. Next, for each object, the average distance is computed from it to objects within other clusters. The minimum distance is decided such that a nearest neighbor cluster is determined for each object (i.e., a next best cluster choice is found for each object). The difference between the average distance of the object with other objects within its cluster and the average distance of the object to other objects in its nearest neighboring cluster is divided by the maximum of these two averages. The closer the resulting number is to 1.0, the better the object is classified. The closer the resulting number is to -1.0, the better the object would be classified in its neighbor cluster. All of these values can be averaged for a clustering scheme to determine overall how well it clusters all the items.

Labeling Clusters

Once clusters have been decided upon, either through the use of a hierarchical or a partitional method, the next task is to label the groups. The labeling is largely a qualitative exercise and may be quite subjective. A human rater may develop labels by reading through the documents contained within a cluster and inferring the commonality between the documents. This method of labeling is limited by the ability of the person to read through the entire collection of documents; therefore, the label may be biased by the

selection of documents the reader was able to attend to. Furthermore, the label may be influenced by the biases of the reader.

One method for limiting the subjectivity in the interpretation of the clusters is to adopt an empirical approach to the labeling. For instance, Manning, Raghavan, and Schutze (2008) described the method of differential cluster labeling which compares the distribution of terms across clusters to develop appropriate labels. A similar method is to examine keywords within a cluster and allow these to drive the labeling process. Although still somewhat subjective in terms of the interpretation of the keywords, the analysis of determining keywords to aid in the labeling of the clusters helps to avoid some of the subjectivism plaguing much of qualitative research.

Keywords can be found by comparing the usage of a word within a cluster to its usage in the whole corpus of documents. Words are defined as being *key* if they distinguish the document from the other documents within the corpus because the term appeared with a different frequency. The *keyness* of a term may be computed by calculating a chi-square statistic comparing the frequency of the term in the document to the frequency of the term in the corpus of documents. A significant chi-square signifies a key term. A review of these keywords may be used to determine labels for the clusters.

WordSmith, a program developed by Scott (2008), may be used in developing word lists and distinguishing keywords within sets of documents. *WordSmith* determines keywords by comparing the frequency of the use of the term within a document collection of interest to the term's use in some larger corpus of documents, as defined by the user. A chi-square statistic is computed to determine if the term is used significantly more (or less) within the documents of interest.

These keywords can then be used to evaluate the meaning of the set of documents and its similarity to the larger corpus chosen. For instance, Scott (1997) described the benefits of using keywords to help reveal socially important concepts and stereotypes when applied to the analysis of news stories from a select time. Berber-Sardinha (1999) reviewed the KeyWord tool within the *WordSmith* program and explained its usefulness of finding keywords in distinguishing documents or distinguishing documents from a larger corpus.

Purpose of this Project

The current classification schemes of airline accidents and incidents have been developed through a top-down, rational approach. For instance, ACCERS (Krokos & Baker, 2005) was developed through reviewing the literature and interviewing pilots. Through these interviews and reviews, the authors determined relevant categories under which ASAP reports could be filed. In contrast, the current study embraces an empirical method for determining a classification scheme for aviation incidents.

The novel approach embraced by the current study implemented a computer automated classification of the ASRS human error documents. There were two main advantages over a human-centered approach. First, by using a computer to aid in the textual analysis of the documents within the ASRS database, the narratives are processed reliably (Krippendorff, 2004; Popping, 2000). Human raters cannot help but bring biases into reading and rating of documents. These biases may interfere with the interpretation of the document and may bias the classification of it. Another advantage to the automatizing of the classification process is the ability of the computer to process large amounts of data. Without excessive amounts of time available, a human rater is often

limited to sampling documents from a larger collection, whereas a computer is able to process the entire collection in minimal time.

An additional feature of the empirical approach of the current study is that it offers a finer grain definition of human error than that proposed in earlier classifications (e.g., ACCERS, HFACS). This goal is accomplished through the use of a bottom-up, statistical approach focusing closer attention on the data already gathered within ASRS. The motivation behind this plan was that through the use of text analysis in this manner, and being closer to the data, a finer distinction can be made between the existing types of human error associated with errors in flying.

Implementation of the empirical approach to the classification of the ASRS human error narratives within the current study was accomplished in a series of steps. First, the application of SVD was evaluated to determine the impact of dimension reduction on creating discernible clusters. Next, hierarchical and partitional clustering methods were examined to ascertain the best clustering scheme. Finally, term analysis was performed to aid in the labeling of the resulting documents. The combination of these analytical techniques was evaluated as a whole to determine the benefits of computer-automated classification.

Hypotheses

This project searched for human error types within aviation incidents and explored LSA's ability to assist in this discovery and discern patterns within aviation incident narratives. The first hypothesis was that LSA, and specifically the application of SVD, would be better, both in efficiency and results, in categorizing the narratives into

meaningful clusters compared to a non-LSA alternative. The non-LSA alternative tested was an option that offers simplicity in the analysis by deleting the calculation of SVD.

Categorizing the ASRS human error incident reports into meaningful categories requires the use of a clustering algorithm. Given the exploratory nature of this project, and the lack of foreknowledge regarding the number of clusters that would emerge, hierarchical clustering was hypothesized to be better suited to the clustering of the narratives compared to the partitional method of *k*-means clustering, which requires some foreknowledge of the number of clusters expected.

Finally, the third hypothesis concerned the clusters produced. The sample of narratives used in this study consisted of reports previously identified as problems with "Flight Crew Human Performance". Therefore, it was hypothesized the resulting taxonomies would closely resemble the principles defined in CRM or SRM classifications. Theses taxonomies focus either on human skills (e.g., classic CRM taxonomy such as explained by Flin & Martin, 2001) or on human error (e.g., Lauber, 1993). These classifications are in contrast to current aviation incident or accident classifications that focus on the entire scope of causes (e.g., ACCERS by Baker & Krokos, 2007; Krokos & Baker, 2005) and offer only the classes of human error and human factors to represent the bulk of the issues with human performance.

Method

Selection of Text

There were approximately 130,000 ASRS narratives available for analysis. The FAA collected these narratives during the years, 1988-2006. When analyzing and classifying a report, analysts were asked to assess the primary problem responsible for the incident. Although most choices are technical in nature, there is an option to file the report under "Flight Crew Human Performance". Approximately 60,000 of these narratives have been identified in this category and represent both commercial and general aviation communities. This subset of narratives was analyzed for the current study.

There were a total of 36,506 documents attributed to commercial aviation (CA) filed under "Flight Crew Human Performance". This set of reports was divided equally into six samples composed of 6,084-6,085 narrative reports. Table 4 shows the number of terms included within each sample of CA documents. Terms were only retained within the sample for analysis if they occurred in at least 0.2% (n = 12) of the documents[6]

[6] The decision to set the threshold to 0.2% was determined so that enough terms would be eliminated from the matrix to facilitate the computation of the analyses. For instance, the full commercial aviation set 1 contained 17,195 terms. The computing requirements (e.g., computer processing memory) are too great for the calculation of many of the analysis steps computed in this study (e.g., calculating the cosine similarity between documents in the non-LSA solution). Setting the threshold to 1% (or requiring the term to occur in approximately 60 documents) removed too many terms (15,745) leaving a scant 1,450 terms to be analyzed. Therefore, the threshold was stepped down to 0.2% to retain a more reasonable number of terms. Furthermore, some strategy for the removal of words not used in more than one document was necessary to remove nonsense de-identifiers used within the narratives to mask the identifying information such as pilot name.

and were composed of at least three letters. All terms were retained that met these criteria.

Table 4. *Number of documents and terms within each CA sample.*

Sample Set	Number of Documents	Number of Terms Total	Number of Terms Retained
CA Set 1	6,084	17,630	3,775
CA Set 2	6,084	17,665	3,842
CA Set 3	6,084	17,367	3,766
CA Set 4	6,084	17,431	3,798
CA Set 5	6,085	17,625	3,779
CA Set 6	6,085	17,519	3,795

There were 23,599 general aviation (GA) ASRS reports classified as "Flight Crew Human Performance" problems. This set of reports was divided into four samples of 5,899-5,900 narratives. Table 5 shows the number of terms included within each of the GA samples. Similar to the analyses for the documents within the CA sample, terms were only retained for analysis if they occurred in at least 0.2% (n = 12) of the documents and were composed of at least three letters.

Table 5. *Number of documents and terms within each GA sample.*

Sample Set	Number of Documents	Number of Terms Total	Number of Terms Retained
GA Set 1	5,900	19,673	4,267
GA Set 2	5,900	19,513	4,225
GA Set 3	5,900	19,858	4,268
GA Set 4	5,899	19,484	4,236

Hardware and Software

All analyses were run on a multi-core (2 x 2.66 GHz dual-core Intel Xeon processor) Mac Pro using 16 GB of 667 MHz FB-DIMM RAM with 2.5 TB of storage. This system is 64-bit native and runs Mac OS X 10.5.4. The 16 GB of RAM used in this set-up represented the maximum amount of memory the 32-bit *MATLAB* software used in this exercise could address.

The documents were stored within a MySQL database. The collection of CA and GA documents were randomly sorted within MySQL. After sorting, the subsets of approximately 6,000 documents were pulled from the database and each of the sets was saved in a text file. The term by document matrix was created through the use of the Text to Matrix Generator (TMG; a *MATLAB* toolbox created by Zeimekis and Gallopoulus, 2007), which took as input a text file that contained the set of 6,000 documents. All matrix decomposition and clustering calculations were done with the software *MATLAB* and the statistics toolbox (Mathworks, 2007). The *MATLAB* syntax used in this study is included in Appendix C.

Text Processing

In the completion of LSA, the following steps were done. First, the log-entropy weighting scheme was applied to the term by document matrix following the removal of terms not occurring in at least 0.2% of the documents. SVD was then performed on this weighted matrix, keeping 150 dimensions. Although Landauer and others have found 300 dimensions to be optimum in a number of studies (e.g., Landauer & Dumais, 1997; Magliano & Millis, 2003), other research has shown the benefit of a smaller number of dimensions (e.g., Dumais, 1991; Elsas, 2005). Ultimately, the best way to determine the optimum number of clusters is to compare the output of LSA to the evaluation of domain experts or to some other external validation criterion (Magliano & Millis, 2003; Quesada 2007). Given the lack of an existing classification criterion for comparison, the current study used 150 dimensions representing a compromise among the number of dimensions suggested in the literature.

It is often the case that words will be stemmed prior to being analyzed through LSA. Stemming is the process of removing suffixes to reduce terms to their base form so that multiple versions of a term are represented only once in the term by document matrix. This practice was not done in the current study as the terms within the ASRS database have already undergone some standardization. For instance, all forms of the term aircraft or airplane map onto arcft (i.e., aircraft).

Due to the sparsity of the matix, the *MATLAB* SVD command “svds” was used to perform singular value decomposition. The application of LSA on the weighted term by document matrix produced orthonormal term by rank and rank by document matrices, as well as the diagonal matrix of singular values. The non-LSA solution was calculated the

same way, except when SVD was performed, as many dimensions were retained as there were terms. Clustering could not be done directly on the weighted term by document matrix due to its sparsity. The calculation of SVD, even though all dimensions were retained, reduced the sparsity of the matrix, enabling the clustering algorithms to be run.

The clustering algorithms were performed using the cosine similarity between the documents. The cosine similarity can be found by taking the dot product of the document space formed by multiplying the matrix of right singular vectors by the diagonal matrix of singular values[7]. The rows within the $\mathbf{D\Sigma}$ matrix represent the coordinates of the documents within the multidimensional space (Deerwester et al., 1990).

Hierarchical and *k*-means clustering were performed within *MATLAB* using the commands "pdist", "linkage", "cluster", and "kmeans". Clusters of various sizes were evaluated. Once the clusters were produced, the goodness-of-cluster was assessed by calculating the WB ratio and silhouette plots. The WB ratio was calculated in *MATLAB* separate from the clustering algorithms. The clustering scheme with the best goodness-of-cluster statistics was selected for continued analysis and labeling. For more information on the use of each of the commands see Appendix C.

Determining the best representative and descriptive labels for the selected clustering scheme involved the use of *WordSmith* (Scott, 2008). The top 20 keywords, as determined based on a chi-square analysis, were used to determine the most appropriate label.

---

[7] The matrices of the left and right singular vectors are represented within Figure 3 by **T** and **D**. The diagonal matrix of singular values is **Σ**.

Results and Discussion

Commercial Aviation Documents

LSA. The LSA solutions using SVD to reduce the dimension of the original term by document matrix to 150 dimensions and the non-LSA solution using SVD to retain the same number of dimensions as terms were computed for each of the six CA samples. The LSA solution was clearly more efficient in terms of computational time required, taking approximately half the time to perform the calculations required (SVD and clustering).

Hierarchical clustering. Hierarchical clustering was calculated through the use of the linkage functions of average, single, and complete using the LSA solution from the CA sample set 1. The linkage functions were calculating using the cosine similarity proximity matrix.[8] The cophenetic correlation was calculated between each of the hierarchical clustering trees produced through each of these linkage methods to the original cosine similarity matrix to determine which method produced the clustering tree that most closely resembled the proximity matrix. The resulting cophenetic correlations for each of the linkage methods are displayed in Table 6. The average linkage method performed best, therefore, it was used in all following hierarchical clustering procedures.

[8] Clustering algorithms within *MATLAB* technically work on distance measures, so the cosine similarity is represented as 1 – cosine in all computations.

Table 6. *Cophenetic correlation between hierarchical clustering produced with different linkage methods and the cosine similarity matrix.*

Linkage Method	**Cophenetic Correlation**
Average	0.4686
Single	0.0364
Complete	0.2543

It is often the case that the hierarchical tree itself can be evaluated for a clear place for pruning the tree. However, given the size of the tree produced with such a large dataset, the resulting tree was not helpful in determining the correct pruning spot. Therefore, the inconsistency coefficients were examined to determine where to prune the hierarchical cluster tree. The inconsistency coefficients ranged from 0 through 1.1547. There was no clear jump in values dictating where a clear cutoff would be. However, the value 0.7071 was the most frequent value, suggesting many links were made at this point. Upon further examination of the number of clusters formed when this inconsistency coefficient was set as the limit to bound the hierarchical clustering, 3,932 clusters were created. When the limit was set at the next highest inconsistency coefficient value (0.7074), 2,480 clusters were produced.

Since the evaluation of inconsistency coefficients gave no clear answer as to where the hierarchical cluster tree should be pruned, a fixed number of clusters was evaluated. As a starting point, the number of clusters was set to nine, which is the same number of categories represented during one trial of the ACCERS taxonomy. This clustering resulted in a large number of the documents contained in two groups, one of

size 5,019 and one of size 1,014. Most other clusters had either small or single group membership.

Therefore, neither setting an inconsistency coefficient cutoff nor setting a maximum number of clusters produced a tractable number of thematic clusters. The results for hierarchical clustering were similar for the other sample sets of commercial or general aviation, therefore, hierarchical clustering was not pursued further.

The problem encountered with the use of hierarchical clustering, namely the production of one large cluster and the fractioning off of smaller groups, may be attributed to the curse of dimensionality. The problem when clustering high dimensional data is the lack of meaning or distinction in the similarity of the objects. Specifically, as the number of dimensions increases, the similarity of objects becomes meaningless as they become equidistant from each other (Parsons, Haque, & Liu, 2004). Therefore, as the clustering algorithm attempts to form clusters by grouping objects most similar, all objects will appear equally similar to one another. Furthermore, many of the dimensions seen in high dimensional data may be irrelevant, thereby masking the true relationships between the objects (Parson, Haque, & Liu, 2004). A clustering algorithm attempting to cluster on these irrelevant features produces an inordinate number of clusters that are not thematically coherent.

K-means clustering. Clustering by the *k*-means method was applied to each of the CA samples with *k* values specified initially as four through nine to mimic the existing CRM and aviation accident taxonomies. To improve the performance of the algorithm, the initial cluster centroid position was chosen by replicating the clustering 100 times, each with a new set of initial cluster centroid points. The solution that produced the

lowest within cluster sums of point-to-centroid distance was chosen as the starting centroid for further analysis.

To compare goodness-of-fit across these clustering results, the WB ratio representing the ratio of the average distance between members within a cluster to the average distance of members across clusters was calculated for each clustering scheme. The average ratio was compared across clustering schemes. The results comparing the LSA and non-LSA solutions are shown in Figure 5 and indicate that the LSA solution provided better clustering outputs than the non-LSA solution. In other words, there was greater differentiation between the clusters produced with the LSA solution than with the non-LSA solution. The range of the WB ratios for the LSA solution was 0.90 (for $k = 4$) to 0.85 (for $k = 9$). The range for the non-LSA solution was 0.98 (for $k = 4$) to 0.97 (for $k = 9$). Recall that a WB ratio of 1.0 indicates no distinction between clusters. The lowest WB ratio for the non-LSA solution was 0.97, very near to 1.0, indicating a lack of discrimination between the clusters. The LSA solution, on the other hand, was approximately a tenth of a proportion lower, pointing to the ability of the LSA solution to better discriminate between clusters.

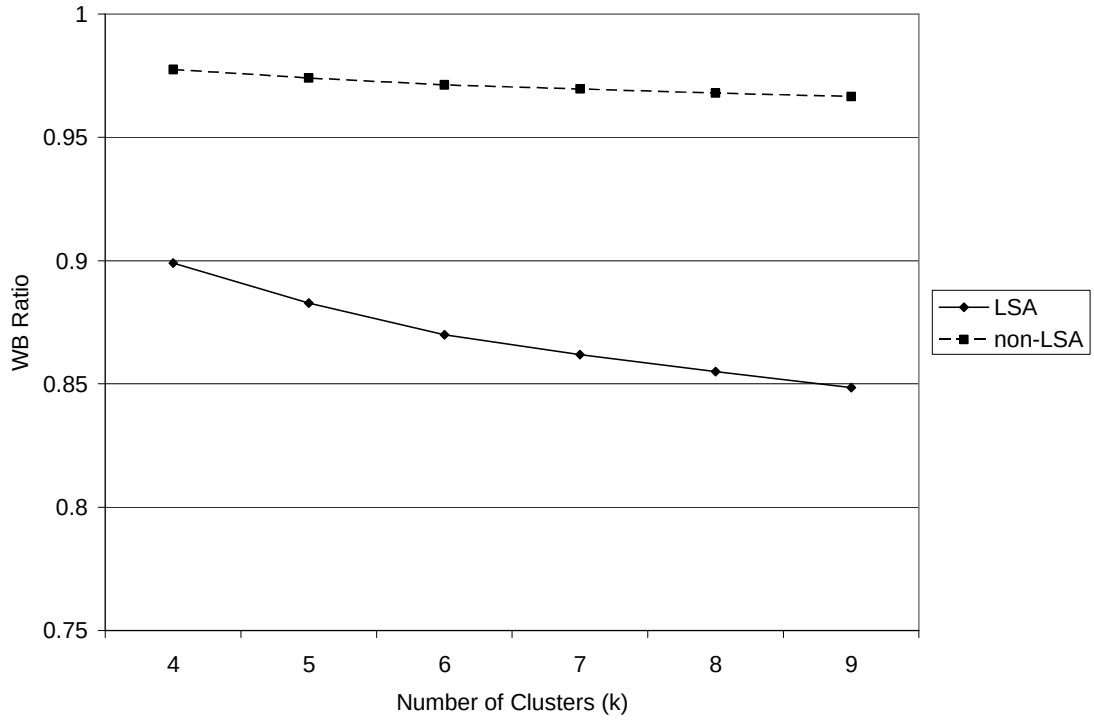

Figure 5. Ratio of within to between variability across clusters from CA documents after application of LSA vs. non-LSA solutions.

Also apparent in Figure 5 is the trend for the WB ratio to continue to improve as the number of clusters increased. Due to this trend of decreasing WB ratios, it was determined that higher values for *k* should be tested. However, given the superior performance of the LSA solution, the additional values for *k* were only evaluated following the application of LSA. Various values for *k* were tried and for each of these the goodness-of-fit was evaluated. The calculation of the WB ratio at higher values of *k* became analytically complex and programmatically unreasonable, so a secondary measure, the average silhouette value, was used to evaluate the goodness of clustering.

Initial *k* values were tested for each of the CA sets ranging from 4 up to 100. The average silhouette value for each of these *k* values is presented in Table 7.

Table 7. *Average silhouette values for various values of k on CA incidents.*

	Average Silhouette Value					
***k* Value**	**CA1**	**CA2**	**CA3**	**CA4**	**CA5**	**CA6**
4	0.0715	0.0716	0.0692	0.0714	0.0700	0.0709
5	0.0747	0.0767	0.0728	0.0749	0.0729	0.0759
6	0.0830	0.0852	0.0808	0.0832	0.0816	0.0844
7	0.0857	0.0866	0.0827	0.0856	0.0834	0.0864
8	0.0832	0.0868	0.0829	0.0879	0.0877	0.0882
9	0.0850	0.0887	0.0867	0.0915	0.0875	0.0868
15	0.0846	0.0832	0.0861	0.0903	0.0983	0.0840
30	0.0999	0.1020	0.1028	0.1026	0.0992	0.1033
40	0.1030	**0.1096**	**0.1069**	0.1055	0.1013	0.1027
45	0.1051	0.1087	0.1014	0.1067	**0.1059**	**0.1042**
50	0.1061	0.1045	0.1028	**0.1098**	0.0997	0.1008
60	**0.1071**	0.1016	0.1039	0.1036	0.1009	0.1002
75	0.1037	0.0954	0.1035	0.1040	0.0971	0.0957
100	0.0997	0.0961	0.1010	0.1077	0.0959	0.0923

Note. The largest average silhouette value for each CA set is in boldface.

The best average silhouette values for each of the CA sets were at *k* equal to 40 through 60. Therefore, every value of *k* within the range of 35 – 60 was tested and evaluated. Overall average silhouette values pointed to two choices: *k* of 53 or 54. Both

of these clustering schemes had an average silhouette value of 0.1072. However, the average silhouette values for *k* equal to 54 was greater than the average silhouette value for *k* equal to 53 on four of the six CA sets. Therefore, it was decided to settle on 54 clusters to be evaluated for labeling.

Labeling of clusters. The 54 clusters from each of the CA document sets were next evaluated for keywords so that appropriate labels could be assigned. The evaluation of labels for each of the clusters proceeded through a mix of quantitative and qualitative analyses. For each of the six sets of CA documents, separate text documents were created holding the documents for each of these individual clusters. In other words, 54 text documents were created for each of the six samples equating to 324 text documents in all.

All the clusters within a single sample of CA reports were analyzed via *WordSmith* to determine keywords. A first step was to create a word list for each of the clusters and a complete wordlist for the CA sample set. The word list tool within *WordSmith* was used to create a complete list of words for each cluster and a word list for each sample of text. After creating the wordlists, the keyword tool within *WordSmith* compared the frequency of words occurring within each cluster to the frequency of occurrence within the larger sample. Chi-square statistics were computed for each word. The top 20 words, as ranked with the chi square statistic, were evaluated to determine labels for each of the clusters. The keywords for each of the clusters are shown in Appendix D.

Although 54 clusters emerged after the k-means clustering, the keywords evident in some of the clusters fit well into a single cluster. For instance, clusters 10 and 39 from

the first CA sample both contained keywords attributable to weight and fuel calculation issues (e.g., *lbs* [pounds], *fuel, wt* [weight], *bal* [balance]). Similarly, the 46th and 50th clusters from this same sample set had keywords about weather issues (e.g., *visibility, wx* [weather], *fog, ice, storm, tstms* [thunderstorms], *winds*). Based on the qualitative analysis of these keywords, labels were formed for each of the clusters. Due to the overlap in some of the keywords among clusters, only 31 groups (including a miscellaneous group) were labeled. A sample of about five documents from each of the 54 clusters was read to ensure the collapsing into 31 groups and the applied classification labels were appropriate.

Across the sample sets, many of the same keywords for clusters were evident indicating the clustering was stable across different subsets of documents. For instance, the category "Fatigue" showed remarkable similarity across the sample sets. To illustrate, Table 8 displays the keywords for the clusters labeled "Fatigue" from each of the sample sets. The common occurrence of the terms *rest, fatigue, sleep, tired* make it relatively easy to label this category "Fatigue". However, even such a clear case such as this one evidences some of the subjectivity in labeling. For instance, this category might also be labeled "Scheduling". However, for the purposes of the current study in uncovering the human error within these documents, the label of the category is biased toward showcasing the human element (i.e., in this instance the physiological effect of fatigue).

Table 8. *Keywords for each of the clusters from the CA sample set representing the category "Fatigue".*

CA Set 1 Cluster 54	**CA set 2 Cluster 52**	**CA set 3 Cluster 4**	**CA Set 4 Cluster 14**	**CA Set 5 Cluster 19**	**CA Set 6 Cluster 11**
Hours	Hours	Hours	Hours	Hours	Hours
Duty	Duty	Rest	Duty	Duty	Rest
Day	Trip	Day	Day	Day	Duty
Rest	Day	Duty	Rest	Rest	Day
Trip	Rest	Scheduled	Trip	Hour	Trip
Scheduled	Hour	Sleep	Scheduling	Sleep	Sleep
Hour	Scheduling	Hour	Hour	Fatigue	Crew
Fatigue	Scheduled	Fatigue	Crew	Night	Scheduling
Scheduling	Sleep	Trip	Days	Trip	Fatigue
Crew	Fatigue	Days	Scheduled	Scheduled	Hour
Sleep	Period	Night	Fatigue	Hotel	Scheduled
Days	Time	Scheduling	Schedule	Days	Night
Minutes	Legal	Crew	Legal	Tired	Legal
Period	Flight	Reduced	Sleep	Scheduling	Hotel
Block	Leg	Legal	Period	Leg	Days
Time	Tired	Minutes	Time	Crew	Schedule
Hotel	Days	Tired	Company	Schedule	Time
Night	Night	Period	Night	Trips	Block
Tired	Legs	Schedule	Flight	Pilot not flying	Period
Legs	Schedule	Hotel	Assignment	AM	Reduced

Upon reading and further consideration of the clusters, the categories could be collapsed to create a total of nine categories. Not all of the original 31 clusters fit into these final nine categories. For instance, the documents that focused on the need for the analyst to call back the reporter for further information of the report contained a mix of incidents. Therefore, for further classification of these reports, greater detail is needed. The original and collapsed categories as well as descriptions of the categories are presented in Table 9. Appendix E presents the division of the document subsets into each of the categories.

As a comparison to the earlier presented WB ratio for the nine-category solution obtained through the application of *k*-means clustering, a WB ratio was computed for the first CA set. The average WB ratio for this derived nine-category solution was 0.91. This ratio is higher than the originally obtained average WB ratio for the *k*-means derived nine-category solution, which was 0.85.

The collapsing of the clusters into first the 31-category and then the 9-category solutions was somewhat subjective and was unable to be accomplished through a quantitative comparison of the keywords within the document sets. A comparison of the top 20 keywords represented within the document sets that were included within one category were compared to obtain some measure of equivalence between these document sets. A ratio of the number of repeats (i.e., a count of the instances in which a keyword is repeated across at least two of the included document sets) to the total number of keywords within the included document sets was calculated (see Appendix E for a full table of these values). For the 31-category solution, the range for this ratio was 0 through 0.56. The range for the 9-category solution was 0 through 0.48. One reason for the low

ratio values is because this measure of intersection of keywords fails to capture similarities such as *brake* and *brakes*. One method for improving this measure may be to stem the terms prior to applying the keyword analysis. Furthermore, the calculation of intersecting terms does not recognize synonyms.

Table 9. *Labeling and description of CA clusters.*

Original 31 Clusters	Collapsed Clusters	Cluster Meaning
Wind Weather Ice	Weather	Flying in inclement weather, including the appropriate use of equipment and skills.
Air Collision / TCASII Restricted Airspace Flight Plan Navigation	Situational Awareness	Being aware of were you should be and were you shouldn't be. Knowing your current location.
Altitude Speed Landing Gear Engine Issues Autopilot	Attention / Monitoring	Paying attention to instruments and equipment and completed checklist items in preparing the instruments/equipment appropriately.
Weight	Weight	Correctly calculating weight and balance
FAA Inspection Maintenance Inspection	Inspection	Being prepared for and responding to inspections
Cabin & Passenger Issues	Interpersonal	Dealing appropriately with passenger issues
ATC Communication / Radio Issues	Communication	Communication with other crew members and with ATC.
Fatigue	Physiological	Physiological effects
Taxi Runway Issues Parking / Pushback Take-off	Context (Runway & Take-off)	Context effects - especially during take-off and runway issues (e.g., knowing where the hold stop is on runway)
Landing Visual Approach Descent / Approach Holding	Context (Landing)	Context effects – especially during landing (e.g., avoiding traffic during approach and setting correct heading)
Location Issues	Context	Context effects at specific airports
Reporter Callback Helicopter Issues Miscellaneous	Miscellaneous	A mixture of narratives that need to be explored deeper for appropriate classification.

The consolidation of categories from the original 54 to the final nine groups is similar to that produced by Krokos and Baker (2005; see also Baker & Krokos, 2007) in the development of ACCERS. In the initial phases of the development of ACCERS,

Krokos and Baker explained that 300 causal contributors were listed as consistently appearing in the ASRS and ASAP reports read. Through the use of subject matter experts and pilots, the list was pared down to first 94, then 9, and finally 7 categories.

In the original conceptualization of the taxonomy for this project, the categories were meant to apply only to human error. However, upon inspection of the clusters, some could be found to be attributable to non-human elements (e.g., weather, phase of flight such as landing). Therefore, the constituents of the taxonomy of the current project come closer to resembling ACCERS, which was originally created as a total classification scheme, than it does to a CRM classification that includes only cognitive and social elements.

A comparison of the results from the current study to any of the CRM or error based classifications, for instance the classic CRM taxonomy or Lauber's (1993) error based classification, is inconclusive as elements from both CRM taxonomies are evident in the current study's classification. Within Lauber's seven categories, the categories of "Monitoring" and "Effective Communication" were both represented in the current study's solution. For the classic CRM taxonomy, the categories of "Situational Awareness" and "Communication" were both present in the current study's resulting categories. However, in general, neither of the CRM classifications fit very well as they excluded non-human elements.

The taxonomy that resulted from the current study did not match perfectly to that proposed by ACCERS. Some of the discrepancy between the current taxonomy and ACCERS may be explained with the awareness that ACCERS was based on the full set of incident reports whereas the currently proposed system was based only on those

reports classified as being due to “Flight Crew Human Performance”. However, the qualitative analysis of the currently proposed scheme revealed all of the labels proposed with ACCERS would also have been appropriate at some level. Table 10 provides a closer comparison of the two classifications to help clarify this last point.

Table 10. *Comparison of CA taxonomy to ACCERS.*

CA Taxonomy	Comparison to ACCERS
Weather	This class maps fairly cleanly onto the group in ACCERS identified as: *Weather or Environment*.
Situational Awareness	The factor 'situational awareness' is included within the broader heading of *Human Factors* within ACCERS to refer to a lack of awareness of what is going on around yourself as well as a lack of effort in discovering important situational variables. However, in the current taxonomy, this category also includes factors included within ACCERS under the heading *Human Error* such as a lack of awareness of flightspace or the misprogramming of controls.
Attention / Monitoring	This class matches many of the elements included in *Human Error* from ACCERS including the improper use of autopilot controls and lack of attention in regards to altitude and altitude settings. However, some elements are also included within the group *Hardware*, which includes the problem of malfunctioning equipment.
Weight	The documents included within this class most closely matched *Human Error* within ACCERS. These incidents were commonly due to a miscalculation of the weight, balance, or fuel for the flight.
Inspection	The narratives within this group could fit into a couple of the ACCERS categories. First, *Policies or Procedures* explains those incidents within this group that are due to confusing or conflicting inspection policies and practices. Next, *Hardware* describes those events that were reported due to a piece of equipment failing repeated times. Finally, *Organizational Factors* covers those issues that arise due to inadequate overview or monitoring by the ground management.
Interpersonal	Interpersonal issues including miscommunication between team members, teamwork among the crew, and difficulty in dealing with passengers fits most closely under the heading of *Human Factors* in ACCERS.
Communications	Most of the issues brought up in this heading are related to those factors included in ACCERS' *Airspace or Air Traffic Control*.
Physiological	Primarily this heading refers to fatigue, which is included in ACCERS' *Human Factors*. However, it may also include too little time between flights, which is an element of *Organizational Factors*.
Context	Some of the issues that arise in this category are covered by *Human Factors* in ACCERS such as performing work during times of high task load and saturation. Other issues are related to those covered by *Airspace or ATC* in which communications with ATC may either be incorrect or ill-timed due to high frequency or lack of monitoring and difficulty with the airport may be due to poor markings or signs.

General Aviation Documents

LSA. The LSA solution using SVD to reduce the dimension of the original term by document matrix to 150 dimensions and the non-LSA solution using SVD to retain the same number of dimensions as terms were computed for each of the four GA samples. Similar to the calculations on the CA sets, both machine and real time calculations took considerably longer for the non-LSA solutions than that required for the LSA solutions.

K-means clustering. Clustering by the *k*-means method was applied to each of the GA samples with *k* values initially specified as four through nine. As was done in the k-means clustering of the CA reports, for these initial clusters, the initial cluster centroid position was chosen by replicating the clustering 100 times, each with a new set of initial cluster centroid points. The solution that produced the lowest within cluster sums of point-to-centroid distance was chosen as the solution for that *k* value tested.

The WB ratio comparing within cluster similarity to across cluster similarity was calculated for each of the resulting classifications. As was seen for the CA reports, the classification seemed to show improvement with each increase in the number of clusters. Figure 6 shows the improvement in performance as *k* was increased for both the LSA and the non-LSA solutions. It is also clear by comparing these graphs that the LSA solution presented better clusters than the non-LSA solution. The range of the within to between variability ratios for the LSA solution was 0.852 – 0.909 compared to the range for the non-LSA solution which was 0.970 – 0.981. The results for general aviation are very similar to that seen for commercial aviation. Once again the non-LSA solution does not show much distinction between the clusters (i.e., the WB ratio is very near to 1.0). The

LSA solution was approximately 10% lower and was able to distinguish between clusters.

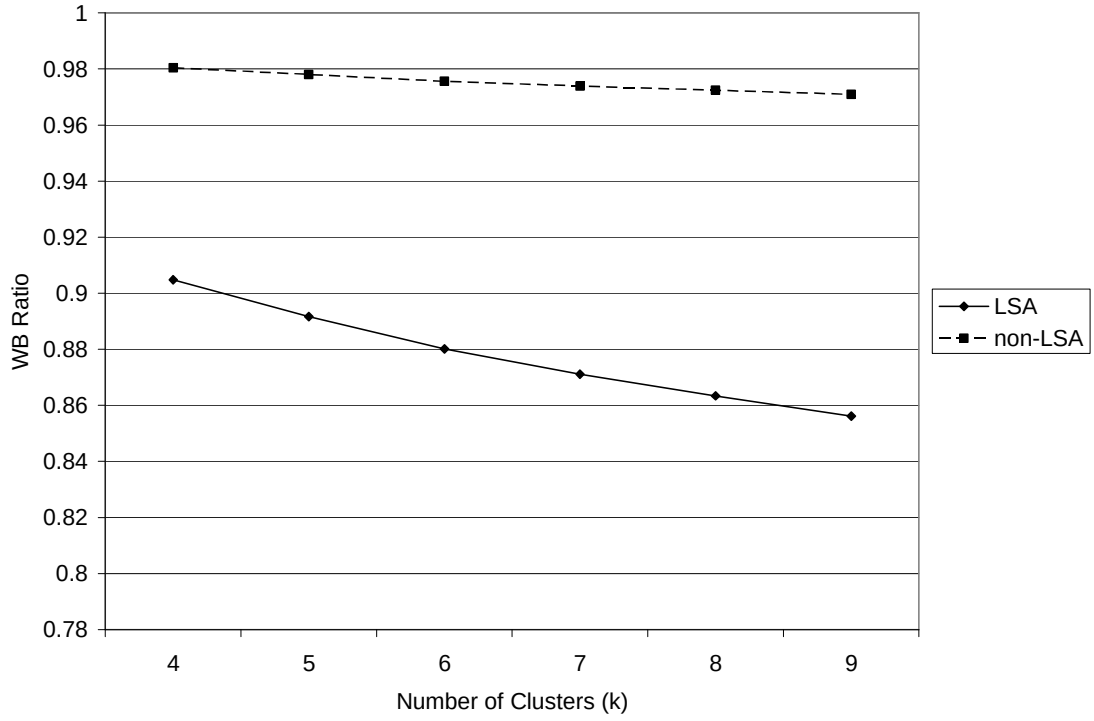

Figure 6. Ratio of within to between variability across clusters from GA documents after application of LSA vs. non-LSA solutions.

Since it appeared that the goodness-of-clustering would improve beyond nine clusters, larger values of *k* were evaluated. An initial sampling of *k* values was tested for values from 4 through 100 to determine what range to focus on. As was done in the evaluation of *k* values for the CA narrative sets, these clusters were evaluated for goodness-of-fit based on the average silhouette value. The average silhouette value for each of these *k* values is presented in Table 11.

Table 11. *Average silhouette values for various values of k on GA incidents.*

	Average Silhouette Values			
***k* Value**	**CA1**	**CA2**	**CA3**	**CA4**
4	0.0652	0.0632	0.0614	0.0626
5	0.0661	0.0647	0.0696	0.0712
6	0.0736	0.0716	0.0778	0.0793
7	0.0781	0.0776	0.0823	0.0833
8	0.0817	0.0808	0.0858	0.0850
9	0.0834	0.0783	0.0874	0.0831
15	0.0939	0.0866	0.0946	0.0970
30	**0.1027**	0.0963	0.0991	**0.1033**
40	0.0977	0.0967	**0.1025**	0.1012
45	0.1026	**0.0986**	0.0995	0.1009
50	0.1019	0.0957	0.0909	0.1008
60	0.0927	0.0935	0.0956	0.0982
75	0.0946	0.0979	0.0914	0.0942
100	0.0889	0.0898	0.0840	0.0899

Note. The largest average silhouette value for each GA set is in boldface.

The best average silhouette values for each of the GA sets were at *k* equal to 30 through 45. Therefore, every value of *k* within the range of 25-45 was tested and evaluated. The highest average silhouette value within this range was 0.1028 at the *k* value of 35. Therefore, 35 clusters were carried forward for labeling.

Labeling of clusters. The KeyWord and WordList tools within *WordSmith* were used to determine the top 20 keywords from each of the 35 clusters. Separate word lists were built for each cluster and compared to the word lists derived from for the GA documents for that set. From the keywords formed, cluster labels were determined for the clusters. The top 20 keywords for each of the clusters are presented in Appendix F. Some clusters were relatively similar in their content and so were placed under one heading. For instance, Clusters 10 and 20 from the first GA sample both had keywords generally about the taxi phase of flight (e.g., *taxi, txwy* [taxiway], *cross, gnd* [ground], *active*). Based on the keywords of the clusters, and due to the overlap within some of the clusters, a total of 33 categories were created. The four sample sets of GA documents showed exceptional similarity in most of the categories. For instance, Table 12 displays the keywords for the four sample sets for the category "Weather". The common occurrence of terms such as *clouds*, *wx* (weather), *VFR* (visual flight rules), and *visibility* imply the categories dealt with issues involving low visibility caused by inclement weather.

Table 12. *Keywords for each of the clusters from the GA sample set representing the category "Weather".*

GA Set 1 Cluster 54	GA set 2 Cluster 52	GA set 3 Cluster 4	GA Set 4 Cluster 14
Clouds	Clouds	Clouds	Clouds
Visual flight rules	Weather	Weather	Visual flight rules
Weather	Visual flight rules	Visual flight rules	Weather
Visibility	Visibility	Cloud	Visibility
Conditions	Conditions	Conditions	Conditions
Cloud	Cloud	Layer	Cloud
Layer	Scattered	Visibility	Scattered
Ceiling	Layer	Ceiling	Layer
Overcast	Ceiling	Scattered	Instrument flight rules
Scattered	Overcast	Broken	Ceiling
Broken	Instrument flight rules	Instrument meteorological conditions	Broken
Instrument flight rules	Broken	Fog	I
Mile	Mile	Instrument flight rules	Feet
Instrument meteorological conditions	Fog	Icing	Fog
Fog	Rain	Ceilings	Instrument meteorological conditions
I	Instrument meteorological conditions	I	Ceilings
Rain	Below	Overcast	Overcast
Forecast	Hole	Ice	Mile
Hole	Encountered	Feet	Forecast
Feet	Ceilings	Top	Hole

A sample of approximately five documents from each cluster was read to determine if the cluster label was appropriate. Upon further consideration of the categories, it was realized they could be collapsed further into a final twelve category solutions. Twelve headings resulted from this consolidation. The original cluster headings, the collapsed headings, and descriptions of the categories are presented in Table 13. Appendix G presents the division of the document subsets into each of the categories.

Similar to that discussed in the collapsing of document sets done in the creation of the commercial aviation taxonomy, this collapsing was also partially subjective. A comparison of the top 20 keywords represented within the document sets that were included within one category were compared to obtain some measure of equivalence between these document sets. A ratio of the number of repeats to the total number of keywords within the included document sets was calculated (see Appendix G for a full table of these values). For the 33-category solution, the range for this ratio was 0.05 through 0.55. The range for the 12-category solution was 0 through 0.50. Again, as described for the commercial aviation reports, one possible reason for the low measures resulting from the comparison of the documents sets is the treatment of essentially identical words as dissimilar (e.g., *cloud* and *clouds*).

The resulting classification differs from the SRM taxonomy in that there are non-human elements such as phase of flight, weather, and mechanical issues within the categories of the current study. The SRM taxonomy, similar to the classic CRM taxonomy examined for commercial aviation, focuses on the cognitive aspects of being in the cockpit (e.g., task management, aeronautical decision making). Only the category of

"Situational Awareness" was common between the current study's GA classification and the SRM taxonomy. In fact, the current study's classification was more similar to the classic CRM taxonomy discussed in the results for commercial aviation in which the additional category of "Communication" was similar.

Table 13. *Labeling and description of GA clusters.*

Original 33 Clusters	Collapsed Clusters	Cluster Meaning
Weather Wind Ice	Weather	Flying in inclement weather, including the appropriate use of equipment and skills.
Fuel / Weight	Calculation / Weight	Calculating fuel consumption and weight correctly.
Altitude Autopilot Control Instrument Flight ILS Approach	Use of Instruments	Using and monitoring instruments correctly and being able to fly through instruments.
Break Issues Landing Gear Propeller Issues	Mechanical Issues	Exercising proper care, use and monitoring of equipment.
Student / Instructor	Teaching	The relationship between the student and instructor including communication and appropriate instruction.
NOTAMs / TFRs	Monitoring	Monitoring advisory reports and staying up to date on current closures and temporary restricted spaces.
Communication / Radio	Communication	Communication between people on board and between the pilot and the control tower.
Restricted Airspace Navigation TCA's Air Collision	Situational Awareness	Having awareness of where you are, including staying on the appropriate route and out of restricted airspace.
Ramp / Parking Taxi Take-off Departure	Context (Runway / Take-off)	Context effects – especially during take-off and runway issues (e.g., not crossing an active runway).
Night Flying Arrival / Scheduling Landing	Context (Landing)	Context effects – especially during landing (e.g., knowing where the airport is and what runway to use).
Helicopter Aerobatic Parachuting Hot Air Balloons Gliders	Types of Aircraft	The flying of different types of aircraft and being familiar with the rules and regulations of each.
Team	Interpersonal	Relationship between people on board.
FAA Inspection	Inspection	Being prepared for and passing an inspection.
Reporter Callback	N/A	Not able to be classified without further detail.

Summary and Conclusion

Text Analysis Methods

The results from the current study demonstrated that the combined use of LSA, *k*-means clustering, and keyword analysis can be used to develop a taxonomy for classification of "Flight Crew Human Performance" ASRS documents. A set of incident reports representing human error within the commercial and general aviation were selected for study and were classified through the use of the combined analytical techniques.

The representation of the documents in the reduced dimensional space following the application of LSA resulted in more distinct clusters compared to a representation of the documents in a non-reduced term space. This finding supported the first hypothesis and lends credence to the idea that SVD is the key component that aids clustering. The benefit gained by the use of SVD is due to the dimensionality reduction representing the term by document matrix in a more compact form.

By reducing the dimensionality of the term by document matrix, terms are no longer forced to be orthogonal to each other and synonyms may be found. Therefore, documents not containing the same term may still exhibit similarity to one another as long as they contain synonymous terms. This similarity facilitates the clustering of like documents – even those documents that have no terms in common.

Following the application of LSA, the documents could be clustered through the use of *k*-means clustering. Although the initial results following the clustering of the documents through the use of the *k*-means algorithm appeared to result in a large number of categories (in comparison with existing aviation reporting systems), an examination of

the clusters showed some similarity between clusters and resulted in a taxonomy comparable to an existing aviation scheme. The finding supporting the use of *k*-means clustering was in contrast to the second hypothesis stating a preference for hierarchical clustering. Although hierarchical clustering was initially preferred given the exploratory nature of this study, this clustering resulted in a large number of very small clusters that were unable to be thematically coded.

Finally, it was through the analysis of the keywords within each of the clusters that appropriate labels could be attached to the clusters and to the final groupings. Although the final conceptualization of the categories was qualitative in nature, the primary step of analyzing the clusters for keywords removed much of the subjectivity inherent in most labeling exercises.

The empirically based approach pursued in the current study represented a new method for classifying the ASRS narratives. The classifications and taxonomies that are in use today within the airline industry were developed through expert judgment of relevant themes or through a rational, top-down approach involving the human coding of narrative reports. These systems are often biased by the expectations of the human experts and may be an incomplete picture of all the contributing factors to airline incidents.

The empirical approach embraced by the current study allowed the narrative data (narratives put into words by the reporting agent – pilot, ATC, or other) to drive the classification scheme. By automating the process, all narratives were represented and included within the analysis. This all inclusiveness is in contrast to a human rater who is often limited by time constraints and must, therefore, rely on a sampling of the narratives.

Furthermore, by limiting the influencing bias of a human rater to interpreting only the keywords, latent categories dealing with human performance within the cockpit were revealed. Through this unique combination of analytical methods, a taxonomy for commercial and general aviation emerged that showed a heavy influence of context as well as human elements such as skills in communication and situational awareness.

Human Error Taxonomy

The classification of the CA documents in the current study most closely resembled the ACCERS taxonomy developed by Krokos and Baker (2005; see also Baker & Krokos, 2007). The most frequent match between the elements represented in the current proposed taxonomy and ACCERS classification were *Human Factors* and *Human Error*. This finding makes sense considering the current proposed classification is built on only those narratives identified as primarily human error.

One advantage of the current classification over ACCERS is its ability to automatically classify the ASRS reports by their narratives. This classification can be done through the use of LSA by means of information filtering described by Dumais (2007). New incident reports entered into the system are compared to the existing corpus of reports. The existing corpus may be organized based on the currently proposed categorization or any existing taxonomy (e.g., ACCERS). The newly entered report is added to the category to which it is most similar (based on a similarity measure such as cosine). In contrast, the application of ACCERS to filed reports requires human interaction in the correct classification of the report.

The current proposed taxonomy may also be used to help distinguish between the two categories *Human Error* and *Human Factors* within ACCERS. Baker and Krokos

(2007) reported that pilots had difficulty distinguishing between these categories. As they further clarified, “In addition, researchers and practitioners alike have traditionally had difficulty distinguishing between the outcome of performance (i.e., human error) and performance itself (i.e., human factors)” (pg. 197). The current analysis elucidated those factors that drive the incidents within these classes.

In regards to general aviation, the closest existing taxonomy is that offered by Summers et al. (2007). Their taxonomy was originally conceptualized for training purposes primarily to help aid flight management and decision-making skills. The primary difference between the resulting taxonomy from the current study and the single pilot resource management (SRM) classification is the inclusion of non-human elements such as mechanical issues within the current scheme. Although NASA analysts previously identified the narratives included in the current analysis as primarily being due to “Flight Crew Human Performance”, some of the narratives were based on mechanical issues. It is likely that the NASA analysts included these incidents because the mechanical issue impacted the human performance or decision-making in some way.

Another key deviation between the current taxonomy and the classic SRM classification is the impact of the interaction with other people. SRM is built on the idea of the pilot being alone in the cockpit and in the decision making process. However, what was commonly seen in the narratives analyzed in the current study was an interaction between the pilot and another person on board. Although the formal partnership between captain and first officer may not exist in GA, there are often times when a second pilot or other person (such as a passenger or family member) may be on the flight. This inclusion of other people may serve as a help or hindrance to the pilot

flying. The second person may aid the pilot in troubleshooting and problem solving. Conversely, the other person may distract the pilot. The exclusion of communication and interpersonal relationships from SRM means it does not fit the current analysis of GA perfectly.

Cross Validation by a Subject Matter Expert

Review by subject matter experts (SMEs) was found to be a crucial component in the development of the ACCERS taxonomy (Krokos & Baker, 2005). The SMEs helped in validating the final labels assigned to the groups as well as prune the classification system from nine to seven headings. Therefore, as an extension of the current study, the keywords for the first set of documents for the CA and GA collections were reviewed by a SME. The SME tasked with this project was a certified flight instructor/instrument airplane transport pilot.

The SME reviewed the keywords for the first CA and GA set of document clusters to assign category labels. The labels proposed by the SME were overall consistent with the labels originally proposed for the 31-category solution for CA and 33-category solution for GA, but showed some discrepancies. Specifically, the two labeling schemes matched on 59% of the 54 CA clusters and on 63% of the 35 GA clusters. Agreement was most closely seen for those sets of documents related to weather, ground incidents (e.g., runway incursions, ground orientations), course deviations and navigational errors, and issues during landing or approach.

In the reduction of the 31-category CA solution and the 33-category GA solution to the 9- and 12-category solutions, respectively, the SME recommended the separation of the context category into at least two components composed of take-off and landing

phases. The tasks during these phases are sufficiently different to urge this separation. Furthermore, the SME was able to offer labels more relevant to the human participant in the event. For instance, documents that were originally classified within this study as pertaining to restricted airspace and some listed as communication problems could be reclassified as procedural errors. The SME also recommended re-labeling the group currently classified as Inspection as CFR Violations. He also corrected the misclassification of Interpersonal reports as Mechanical issues, therefore, recommending that the taxonomy for CA include a Mechanical category similar to that already in use in the GA taxonomy.

Impact of Findings

Dekker (2006) advised that human error could be viewed from either a human-centered perspective in which the person is the cause of the mishap or accident or from a system perspective in which the error is a symptom of something deeper. The assessment of the ASRS narratives helped to shed light on what the deeper problems may be. For instance, commercial pilots flying when they are sick might be a symptom of an organizational culture that places more importance on completing a flight than on safety. GA pilots might be similarly reluctant to cancel flights. Wright (2004) reported that the primary cause of fatal general aviation accidents is pilots intentionally flying instrument rated flights they are not cleared for. They take intentional risks for the purpose of completing a flight.

What can be done to improve flight safety and reduce the potential for error or even eliminate error-producing situations? Furthermore, what needs to be done to resolve the deeper issues that allow human errors to surface? Dekker (2006) provided

instruction on what has been done in the past and been shown to be unsuccessful. First, writing more procedures does not help to solve the problem. Procedures are often written to correct the problem at hand. Organizations can quickly become over-proceduralized creating a situation in which a gridlock occurs if the rules and procedures are followed to the letter.

Adding more technology is also an inappropriate solution. Increasing the technology and components in the system only increases the complexity, which only changes the errors that occur or relocates them (Dekker, 2006). The operator is left trying to understand how to interpret and respond to the new system. Finally, removing or reprimanding the operator who committed the error is not the answer. This solution does nothing to address the deeper problem underlying the accident. In fact, it may make it more difficult to discover the real problem if people start hiding mistakes in an attempt to escape punishment.

The key to understanding mistakes (one type of human error) is to understand why individuals make the decisions they do. At the time, the decision seemed the correct course of action to the person contemplating it. The investigation of the ASRS narratives is a good place to start in exploring the situation surrounding the decision to act and the context driving the decision. The categories revealed with the current study helped to identify situations that lead to the majority of mistakes. For instance, many mistakes are made during landing and takeoff, which are contexts with high taskloads. Many other mistakes are made during bad weather, which represents unfamiliar settings. Training programs that focus on these contexts will better prepare the pilot for responding in unfamiliar settings or making decisions in high workload arenas. Furthermore, the

categories can be used to drive more publications such as *Callback*[9], which brings awareness to tricky or dangerous situations.

One key method for training with CA is through line oriented flight training (LOFT). LOFT training encompasses a full flight and excels in helping pilots develop CRM skills (Federal Aviation Administration, 2004). Both the CA and the GA taxonomies resulting form the current study suggested a "Context" category. The existence of this category implies pilots are having difficulty implementing skills at specific phases of flight, or, alternatively, that these are the most inherently complex phases and have the most potential for error. Therefore, contextually based training in which skills are trained within a context or phase of flight might be beneficial.

The currently developed taxonomy for factors influencing human error can be used to help develop scenarios for critical areas or situations that result in repeated mistakes or violations. For instance, a recurrent theme in some of the context-based reports was trouble with specific airports such as Los Angeles (LAX), San Francisco (SFO) and Washington, D.C. (DCA). Each of these airports presents unique challenges a pilot should be prepared for. For example, the airspace and runways at LAX are complex and congested requiring increased awareness and training on the runway system. SFO is also congested and often difficult for pilots to maneuver through due to the arrangement of the runways and nearby bridges. Finally, flying around Washington, D.C. can be particularly taxing for both CA and GA pilots due to the large amount of restricted airspace in the area.

---

[9] *Callback* is a publication distributed by NASA as part of the ASRS program.

Other improvements can be made in the training of GA pilots. As cited earlier, Kern (2001) urged improved training for GA pilots as general aviation is 20 times more hazardous than commercial aviation. One reason for the higher incidence of accidents that befall GA pilots is the tendency of these pilots to switch aircraft types, often without proper training. One of the primary dangers in switching aircraft is the propensity of the pilot to take the procedures appropriate for the familiar aircraft and inappropriately apply them to the new aircraft.

Furthermore, GA pilots will often fly instrument rated flights for which they are *not* instrument rated (Kern, 2001; Wright, 2004). As GA makes the transition to glass cockpits, problems will arise for pilots in transitioning between aircraft. "In the past, GA aircraft cockpit displays, avionics and navigation equipment all looked the same no matter who manufactured the unit… Advanced technology systems and displays, on the other hand, look different and the way the pilot uses them may differ… Today's regulations do not require a pilot to be formally tested or even have an instructor endorsement when transitioning from one of these airplanes to another" (Glista, 2004, p. 6).

Scenario-based training is endorsed by Summers et al. (2007) to help build the decision-making skills that are often deficient in GA pilots. It is difficult to train CRM or SRM skills independent of the situation as the pilot often lacks the critical insight in how to apply the skill in the situation. Therefore, it is useful to integrate the CRM/SRM skills into scenario-based training similar to LOFT training used in CA.

Challenges and Future Efforts

Within the current study, the application of LSA, and more specifically the application of SVD, was used to reduce the term by document matrix to 150 dimensions. This number of dimensions represented a compromise of the number of dimensions commonly found successful in LSA research. However, an examination of other dimensionalities should be explored to ensure proper representation of the data. Furthermore, Elsas (2005) found that much lower values of dimensions (e.g., ten dimensions) were preferable in the application of LSA to clustering versus information retrieval. Therefore, the study using *k*-means clustering and keyword analysis could be expanded to test further values of dimensions ranging from 10 through 300 to determine the best representation of the data.

The initial review of the keywords representative of the document sets for the commercial and general aviation reports by the subject matter expert should be followed by a full vetting of the labels assigned to the categories. The initial review provided valuable feedback on the consistency of the labels but was not extensive enough to validate the final number of categories or to come to a final consensus on the appropriate label. Finally, the categories have not been tested for use in the field.

The current classification scheme was developed based on incident reports previously classified by NASA analysts as due to “Flight Crew Human Performance”. Although this selection of narratives was helpful in gaining a better understanding of what drives human error in the aviation industry, to be more useful to the NASA analysts as well as the aviation researchers, a more complete classification based on the entire ASRS set should be pursued. It is thought that this expanded analysis would benefit from

the current work in better clarifying the minutiae of human error and also allow for the classification of incidents that are not attributed to the human element.

Appendix A. Sample ASRS Report

Time / Day

Date : 200604
Local Time Of Day : 1801 To 2400
Day : Mon

Place

Locale Reference Airport : DFW Airport
State Reference : TX
Altitude MSL (Mean Sea Level) Single Value : 17000

Environment

Flight Conditions : VMC (visual meteorological conditions)
Light : Dusk

Aircraft : 1

Controlling Facilities TRACON (terminal radar approach control facility) : D10.TRACON
Operator Common Carrier : Air Carrier
Make Model Name : B767-300 and 300 ER
Operating Under FAR (federal aviation regulation) Part : Part 121
Flight Phase Climbout : Initial
Flight Phase Climbout : Intermediate Altitude
Flight Phase Climbout : Takeoff
Route In Use Departure SID (standard instrument departure) : DARTZ
Flight Plan : IFR (instrument flight rules)

Component : 1

Aircraft Component : FMS/FMC (flight management system/ flight management computer)

Person : 1

Affiliation Company : Air Carrier
Function Flight Crew : Captain
Function Oversight : PIC (pilot in command)
ASRS Report : 694974

Person : 2

Affiliation Company : Air Carrier
Function Flight Crew : First Officer
ASRS Report : 694969

Person : 3

Affiliation Government : FAA
Function Controller : Departure

Events

Anomaly Aircraft Equipment Problem : Critical
Anomaly Other Spatial Deviation
Anomaly Other Anomaly Other
Independent Detector Other Flight CrewB
Independent Detector Other Flight CrewA
Resolutory Action Other

Assessments

Problem Areas : Aircraft
Problem Areas : Flight Crew Human Performance
Problem Areas : FAA
Problem Areas : Chart Or Publication
Primary Problem : Ambiguous

Situations

Narrative

During preparation for departure from DFW, I fell victim to a classic case of pattern interruption. There were numerous distractions in the cockpit when I pulled up the clearance. I failed to notice the amendment to use another standard instrument departure (SID). Unfortunately, this new SID has the same initial waypoints as the original SID. Checking in with the ground control and giving him our runway and initial waypoint did nothing to help us catch our error. Fortunately, before we departed from the ground track that is common to both SIDS, we were given a direct routing to a waypoint down the road. It was at that point that we realized our mistake. Supplemental information from ACN 694969: the verification process is useless when more than one area navigation departure uses the same first fix. The area navigation departure verification should include runway, assigned departure, and first fix.

Synopsis

B767-300 flight crew failed to program a change in their area navigation standard instrument departure procedure at DFW. Queue frequency runway/waypoint check fails to warn them because both standard instrument departures utilize the same initial waypoint, TREXX.

Appendix B. ASRS Reporting Form

B

DO NOT REPORT AIRCRAFT ACCIDENTS AND CRIMINAL ACTIVITIES ON THIS FORM.
ACCIDENTS AND CRIMINAL ACTIVITIES ARE NOT INCLUDED IN THE ASRS PROGRAM AND SHOULD NOT BE SUBMITTED TO NASA.
ALL IDENTITIES CONTAINED IN THIS REPORT WILL BE REMOVED TO ASSURE COMPLETE REPORTER ANONYMITY.

(SPACE BELOW RESERVED FOR ASRS DATE/TIME STAMP)

IDENTIFICATION STRIP: Please fill in all blanks to ensure return of strip.
NO RECORD WILL BE KEPT OF YOUR IDENTITY. This section will be returned to you.

TELEPHONE NUMBERS where we may reach you for further details of this occurrence:

HOME Area _____ No. _____ Hours _____
WORK Area _____ No. _____ Hours _____

NAME _____
ADDRESS/PO BOX _____
_____
CITY _____ **STATE** _____ **ZIP** _____

TYPE OF EVENT/SITUATION
_____
_____
DATE OF OCCURRENCE _____ (MM/DD/YYYY)
LOCAL TIME (24 hr. clock) _____ (HH:MM)

PLEASE FILL IN APPROPRIATE SPACES AND CHECK ALL ITEMS WHICH APPLY TO THIS EVENT OR SITUATION.

REPORTER
- ☐ Captain
- ☐ First Officer
  - ☐ pilot flying
  - ☐ pilot not flying
- ☐ Other Crewmember
- ☐ _____

FLYING TIME
total _____ hrs.
last 90 days _____ hrs.
time in type _____ hrs.

CERTIFICATES/RATINGS
- ☐ student ☐ private
- ☐ commercial ☐ ATP
- ☐ instrument ☐ CFI
- ☐ multiengine ☐ F/E
- ☐ _____

ATC EXPERIENCE
- ☐ FPL ☐ Developmental
- radar _____ yrs.
- non-radar _____ yrs.
- supervisory _____ yrs.
- military _____ yrs.

AIRSPACE
- ☐ Class A (PCA)
- ☐ Class B (TCA)
- ☐ Class C (ARSA)
- ☐ Class D (Control Zone/ATA)
- ☐ Class E (General Controlled)
- ☐ Class G (Uncontrolled)
- ☐ Special Use Airspace
- ☐ airway/route _____
- ☐ unknown/other _____

WEATHER
- ☐ VMC ☐ ice
- ☐ IMC ☐ snow
- ☐ mixed ☐ turbulence
- ☐ marginal ☐ tstorm
- ☐ rain ☐ windshear
- ☐ fog ☐ _____

LIGHT/VISIBILITY
- ☐ daylight ☐ night
- ☐ dawn ☐ dusk
- ceiling _____ feet
- visibility _____ miles
- RVR _____ feet

ATC/ADVISORY SERV.
- ☐ local ☐ center
- ☐ ground ☐ FSS
- ☐ apch ☐ UNICOM
- ☐ dep ☐ CTAF
- Name of ATC Facility: _____

	AIRCRAFT 1	AIRCRAFT 2
Type of Aircraft (Make/Model)	(Your Aircraft) _____ ☐ EFIS ☐ FMS/FMC	(Other Aircraft) _____ ☐ EFIS ☐ FMS/FMC
Operator	☐ air carrier ☐ military ☐ corporate ☐ commuter ☐ private ☐ other _____	☐ air carrier ☐ military ☐ corporate ☐ commuter ☐ private ☐ other _____
Mission	☐ passenger ☐ training ☐ business ☐ cargo ☐ pleasure ☐ unk/other_____	☐ passenger ☐ training ☐ business ☐ cargo ☐ pleasure ☐ unk/other_____
Flight plan	☐ VFR ☐ SVFR ☐ none ☐ IFR ☐ DVFR ☐ unknown	☐ VFR ☐ SVFR ☐ none ☐ IFR ☐ DVFR ☐ unknown
Flight phases at time of occurrence	☐ taxi ☐ cruise ☐ landing ☐ takeoff ☐ descent ☐ missed apch/GAR ☐ climb ☐ approach ☐ other _____	☐ taxi ☐ cruise ☐ landing ☐ takeoff ☐ descent ☐ missed apch/GAR ☐ climb ☐ approach ☐ other _____
Control status	☐ visual apch ☐ on vector ☐ on SID/STAR ☐ controlled ☐ none ☐ unknown ☐ no radio ☐ radar advisories	☐ visual apch ☐ on vector ☐ on SID/STAR ☐ controlled ☐ none ☐ unknown ☐ no radio ☐ radar advisories

If more than two aircraft were involved, please describe the additional aircraft in the "Describe Event/Situation" section.

LOCATION
Altitude _____ ☐ MSL ☐ AGL
Distance and radial from airport, NAVAID, or other fix
_____
Nearest City/State _____

CONFLICTS
Estimated miss distance in feet: horiz _____ vert _____
Was evasive action taken? ○ Yes ○ No
Was TCAS a factor? ○ TA ○ RA ○ No
Did GPWS activate? Reset ○ Yes ○ No

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

NASA has established an Aviation Safety Reporting System (ASRS) to identify issues in the aviation system which need to be addressed. The program of which this system is a part is described in detail in FAA Advisory Circular 00-46D. Your assistance in informing us about such issues is essential to the success of the program. Please fill out this form as completely as possible, enclose in an sealed envelope, affix proper postage, and and send it directly to us.

The information you provide on the identity strip will be used only if NASA determines that it is necessary to contact you for further information. THIS IDENTITY STRIP WILL BE RETURNED DIRECTLY TO YOU. The return of the identity strip assures your anonymity.

AVIATION SAFETY REPORTING SYSTEM

Section 91.25 of the Federal Aviation Regulations (14 CFR 91.25) prohibits reports filed with NASA from being used for FAA enforcement purposes. This report will not be made available to the FAA for civil penalty or certificate actions for violations of the Federal Air Regulations. Your identity strip, stamped by NASA, is proof that you have submitted a report to the Aviation Safety Reporting System. We can only return the strip to you, however, if you have provided a mailing address. Equally important, we can often obtain additional useful information if our safety analysts can talk with you directly by telephone. For this reason, we have requested telephone numbers where we may reach you.

Thank you for your contribution to aviation safety.

NOTE: AIRCRAFT ACCIDENTS SHOULD NOT BE REPORTED ON THIS FORM. SUCH EVENTS SHOULD BE FILED WITH THE NATIONAL TRANSPORTATION SAFETY BOARD AS REQUIRED BY NTSB Regulation 830.5 (49CFR830.5).

If you want to mail this form, please fold both pages (and additional pages if required), enclose in a sealed, stamped envelope, and mail to:

NASA AVIATION SAFETY REPORTING SYSTEM
POST OFFICE BOX 189
MOFFETT FIELD, CALIFORNIA 94035-0189

If you wish to submit online, click the **Submit** button at the bottom of page 2 or 3 when complete.

DESCRIBE EVENT/SITUATION

Keeping in mind the topics shown below, discuss those which you feel are relevant and anything else you think is important. Include what you believe really caused the problem, and what can be done to prevent a recurrence, or correct the situation. (USE ADDITIONAL PAPER IF NEEDED)

CHAIN OF EVENTS
- How the problem arose
- How it was discovered
- Contributing factors
- Corrective actions

Page 2 of 3

HUMAN PERFORMANCE CONSIDERATIONS
- Perceptions, judgments, decisions
- Actions or inactions
- Factors affecting the quality of human performance

NASA ARC 277B (January 1994)

Click here to securely submit to NASA ➔ **Submit**

DESCRIBE EVENT/SITUATION, continued...

CHAIN OF EVENTS
- How the problem arose
- How it was discovered
- Contributing factors
- Corrective actions

Page 3 of 3

HUMAN PERFORMANCE CONSIDERATIONS
- Perceptions, judgments, decisions
- Actions or inactions
- Factors affecting the quality of human performance

NASA ARC 277B (January 1994)

Click here to securely submit to NASA ➔ **Submit**

Appendix C. *MATLAB* Syntax

MATLAB Syntax for SVD Calculations

The output term by document matrix from TMG was labeled **A** in which the rows constituted the term frequencies and the columns represented the documents. SVD was calculated on this **A** matrix. Syntax is shown here for both the LSA and non-LSA solutions. For each of the following sections, the commands will only be shown for the LSA solution with the understanding that equivalent calculations were done for the non-LSA solution.

```
[U, S, V] = svds (A, 150);    % Computing the term (U), document (V), and singular
                              % values (S) for the LSA solution
[U_noSVD, S_noSVD, V_noSVD] = svds (A, #);          % Computing the term
                              (U_noSVD),
                              % document (V_noSVD), and singular values (S_noSVD)
                              % for the non-LSA solution.  The number of dimensions
                              % entered was equal to the number of terms.
SV = V * S;                   % multiplying the S and V matrices before clustering steps.
SV_noSVD = V_noSVD * S_noSVD;       % multiplying the S and V matrices for the
                              % non-LSA solution.
```

MATLAB Syntax for Hierarchical Clustering

```
SV_cosine = pdist (SV, 'cosine');          % Calculating the cosine similarity between
                                           % the objects represented in the SV matrix.  MATLAB
                                           % treats everything as a distance, so the calculation of the
                                           % cosine is actually calculated as 1 – cosine.
SV_link_avg = linkage (SV_cosine, "average");          % Using the output from the
                                           % pdist command just used, this command groups the
                                           % objects into a binary, hierarchical cluster tree.
                                           % "Linkage" groups pairs of objects together that are
                                           % within close proximity.  Different methods of linkage
                                           % can be computed.  In this instance, the linkage is
                                           % calculated using the 'average' method.  Within this
                                           % study, the methods of 'single' and 'complete' were also
                                           % used.
coph_corr_avg = cophenet (SV_link_avg, SV_cosine);          % Using the
                                           % hierarchical cluster tree constructed in the previous step,
                                           % the cophenet correlation calculates how closely the
                                           % hierarchical tree relates to the original data generated by
                                           % the pdist command.
inconsistency = inconsistent (SV_link_avg);       % Calculates the inconsistency
                                           % coefficient which represents how the height of a link in
                                           % the hierarchical cluster tree compares to the average
                                           % height of links below it in the tree.
cluster_7071 = cluster (SV_link_avg, 'cutoff', 0.7071);          % Using
                                           % inconsistency coefficients determined in the previous
                                           % step, cutoff can be set to prune the hierarchical cluster
                                           % tree.  Here the cutoff was set to prune the tree if the
                                           % difference between the current link's height and the
                                           % average height of links below it is any greater than
                                           % 0.7071.
cluster_n_9 = cluster (SV_link_avg, 'maxclust', 9);       % An alternative to pruning
                                           % the hierarchical cluster tree is to set the maximum
                                           % number of clusters to be formed.  In this instance, the
                                           % number of clusters is set to be 9.
```

MATLAB Syntax for k-Means Clustering

First presentation is the calculation for the original clustering of k equal to four through nine clusters. When it was determined that a greater number of clusters should be evaluated, the second set of commands was constructed to calculate k for greater values. An evaluation was done to determine if 100 replications was necessary for each k-means calculation. K-means clustering was done of the first CA set and the values of replication were set to 10, 15, 25, 50, 75, and 100. The ratio of within to between variability was evaluated for each result and no difference was found between the performance of k-means after requesting 100 replications versus requesting 25 replications. Therefore, to save computing power, the k-means calculations for k set greater than 9 were set to 25 replications.

```
k4 = kmeans (SV, 4, 'distance', 'cosine', 'replicates', 100);
k5 = kmeans (SV, 5, 'distance', 'cosine', 'replicates', 100);
k6 = kmeans (SV, 6, 'distance', 'cosine', 'replicates', 100);
k7 = kmeans (SV, 7, 'distance', 'cosine', 'replicates', 100);
k8 = kmeans (SV, 8, 'distance', 'cosine', 'replicates', 100);
k9 = kmeans (SV, 9, 'distance', 'cosine', 'replicates', 100);
k9_SV = [k9 SV];
k8_SV = [k8 SV];
k7_SV = [k7 SV];
k6_SV = [k6 SV];
k5_SV = [k5 SV];
k4_SV = [k4 SV];
k4_SV_sort = sortrows (k4_SV, 1);
k5_SV_sort = sortrows (k5_SV, 1);
k6_SV_sort = sortrows (k6_SV, 1);
k7_SV_sort = sortrows (k7_SV, 1);
k8_SV_sort = sortrows (k8_SV, 1);
k9_SV_sort = sortrows (k9_SV, 1);
k4_SV_sort (:,1) =[];
k5_SV_sort (:,1) =[];
k6_SV_sort (:,1) =[];
k7_SV_sort (:,1) =[];
k8_SV_sort (:,1) =[];
```

```
k9_SV_sort (:,1) =[];
k4_SV_dist = squareform (pdist (k4_SV_sort, 'cosine'));
k5_SV_dist = squareform (pdist (k5_SV_sort, 'cosine'));
k6_SV_dist = squareform (pdist (k6_SV_sort, 'cosine'));
k7_SV_dist = squareform (pdist (k7_SV_sort, 'cosine'));
k8_SV_dist = squareform (pdist (k8_SV_sort, 'cosine'));
k9_SV_dist = squareform (pdist (k9_SV_sort, 'cosine'));
```

Next, the k-means clustering was calculated for values of k greater than 9. Various values were chosen for k and used to target into a range. The performance of the clustering was evaluated using the average silhouette value. The k-means and silhouette commands are presented here:

```
k25 = kmeans (SV, 25, 'distance', 'cosine', 'replicates', 25);
[silh25,h] = silhouette(SV, k25, 'cosine');
```

MATLAB Syntax for Within:Between Ratio

```
function Result = Count(data,condition)
% COUNT (A,B), created by Medlock, R. (2001)
% Counts the number of elements in A that match the criteria specified in B.
nElements = length(data);
IndexIDs = 1:nElements;
Result = eval(['data' condition]);
Result = IndexIDs(Result);
Result = length(Result);

function wbRatio = wbRatioCalc(k4, k4_SV_dist, k5, k5_SV_dist, k6, k6_SV_dist,
        k7, k7_SV_dist, k8, k8_SV_dist, k9, k9_SV_dist)
% SYNTAX:
% wbRatio = wbRatioCalc(k4, k4_SV_dist, k5, k5_SV_dist, k6, k6_SV_dist, k7,
% k7_SV_dist, k8, k8_SV_dist, k9, k9_SV_dist);
% DESCRIPTION:
% Returns matrix of 3 rows in which the first row is the cluster assignment,
% the second row contains the within variances, and the third row contains
% the between variances.
% PARAMETERS:
% k#_SV_dist are the cosine similarity matrices produces for each of the clusters
% RETURN VALUE:
% wbRatio is a matrix in which the columns hold the following values:
% 1. k = 4 cluster grouping
% 2. k = 4 within variabilities
% 3. k = 4 between variabilities
% 4. k = 5 cluster grouping
% 5. k = 5 within variabilities
% 6. k = 5 between variabilities
% 7. k = 6 cluster grouping
% 8. k = 6 within variabilities
% 9. k = 6 between variabilities
% 10. k = 7 cluster grouping
% 11. k = 7 within variabilities
% 12. k = 7 between variabilities
% 13. k = 8 cluster grouping
% 14. k = 8 within variabilities
% 15. k = 8 between variabilities
% 16. k = 9 cluster grouping
% 17. k = 9 within variabilities
% 18. k = 9 between variabilities

% k = 4
% calculate the size of each cluster
```

```
count1 = Count(k4, '==1');
count2 = Count(k4, '==2');
count3 = Count(k4, '==3');
count4 = Count(k4, '==4');
% figure out where each cluster begins and ends within the similarity matrix
m2begin = count1 + 1;
m2end = count1 + count2;
m3begin = m2end + 1;
m3end = m2end + count3;
m4begin = m3end + 1;
m4end = m3end + count4;
% calculate components of the WB matrix
m11_4 = k4_SV_dist (1:count1, 1:count1);
m12_4 = k4_SV_dist (1:count1, m2begin:m2end);
m13_4 = k4_SV_dist (1:count1, m3begin:m3end);
m14_4 = k4_SV_dist (1:count1, m4begin:m4end);
m21_4 = k4_SV_dist (m2begin:m2end, 1:count1);
m22_4 = k4_SV_dist (m2begin:m2end, m2begin:m2end);
m23_4 = k4_SV_dist (m2begin:m2end, m3begin:m3end);
m24_4 = k4_SV_dist (m2begin:m2end, m4begin:m4end);
m31_4 = k4_SV_dist (m3begin:m3end, 1:count1);
m32_4 = k4_SV_dist (m3begin:m3end, m2begin:m2end);
m33_4 = k4_SV_dist (m3begin:m3end, m3begin:m3end);
m34_4 = k4_SV_dist (m3begin:m3end, m4begin:m4end);
m41_4 = k4_SV_dist (m4begin:m4end, 1:count1);
m42_4 = k4_SV_dist (m4begin:m4end, m2begin:m2end);
m43_4 = k4_SV_dist (m4begin:m4end, m3begin:m3end);
m44_4 = k4_SV_dist (m4begin:m4end, m4begin:m4end);
m11_4_wiAvg = sum(m11_4,2)/(size(m11_4,2)-1);
m22_4_wiAvg = sum(m22_4,2)/(size(m22_4,2)-1);
m33_4_wiAvg = sum(m33_4,2)/(size(m33_4,2)-1);
m44_4_wiAvg = sum(m44_4,2)/(size(m44_4,2)-1);
m1_4_bw = [m12_4 m13_4 m14_4];
m2_4_bw = [m21_4 m23_4 m24_4];
m3_4_bw = [m31_4 m32_4 m34_4];
m4_4_bw = [m41_4 m42_4 m43_4];
m1_4_bwAvg = mean(m1_4_bw,2);
m2_4_bwAvg = mean(m2_4_bw,2);
m3_4_bwAvg = mean(m3_4_bw,2);
m4_4_bwAvg = mean(m4_4_bw,2);
k4_1 = ones(count1,1);
k4_2 = 2*ones(count2,1);
k4_3 = 3*ones(count3,1);
k4_4 = 4*ones(count4,1);
m1_4_WB = [k4_1 m11_4_wiAvg m1_4_bwAvg];
m2_4_WB = [k4_2 m22_4_wiAvg m2_4_bwAvg];
```

```
m3_4_WB = [k4_3 m33_4_wiAvg m3_4_bwAvg];
m4_4_WB = [k4_4 m44_4_wiAvg m4_4_bwAvg];
k4_WB = [m1_4_WB; m2_4_WB; m3_4_WB; m4_4_WB];
% clear the memory of unneeded variables.  To save space, this portion of the
% program is not included here.  It is simply a series of 'clear' commands to clear
% the memory of all variables except the final one created "k4_WB".

% k = 5
% calculate the size of each cluster
count1 = Count(k5, '==1');
count2 = Count(k5, '==2');
count3 = Count(k5, '==3');
count4 = Count(k5, '==4');
count5 = Count(k5, '==5');
% figure out where each cluster begins and ends within the similarity matrix
m2begin = count1 + 1;
m2end = count1 + count2;
m3begin = m2end + 1;
m3end = m2end + count3;
m4begin = m3end + 1;
m4end = m3end + count4;
m5begin = m4end + 1;
m5end = m4end + count5;
% calculate components of the WB matrix
m11_5 = k5_SV_dist (1:count1 ,1:count1);
m12_5 = k5_SV_dist (1:count1, m2begin:m2end);
m13_5 = k5_SV_dist (1:count1, m3begin:m3end);
m14_5 = k5_SV_dist (1:count1, m4begin:m4end);
m15_5 = k5_SV_dist (1:count1, m5begin:m5end);
m21_5 = k5_SV_dist (m2begin:m2end, 1:count1);
m22_5 = k5_SV_dist (m2begin:m2end, m2begin:m2end);
m23_5 = k5_SV_dist (m2begin:m2end, m3begin:m3end);
m24_5 = k5_SV_dist (m2begin:m2end, m4begin:m4end);
m25_5 = k5_SV_dist (m2begin:m2end, m5begin:m5end);
m31_5 = k5_SV_dist (m3begin:m3end, 1:count1);
m32_5 = k5_SV_dist (m3begin:m3end, m2begin:m2end);
m33_5 = k5_SV_dist (m3begin:m3end, m3begin:m3end);
m34_5 = k5_SV_dist (m3begin:m3end, m4begin:m4end);
m35_5 = k5_SV_dist (m3begin:m3end, m5begin:m5end);
m41_5 = k5_SV_dist (m4begin:m4end, 1:count1);
m42_5 = k5_SV_dist (m4begin:m4end, m2begin:m2end);
m43_5 = k5_SV_dist (m4begin:m4end, m3begin:m3end);
m44_5 = k5_SV_dist (m4begin:m4end, m4begin:m4end);
m45_5 = k5_SV_dist (m4begin:m4end, m5begin:m5end);
m51_5 = k5_SV_dist (m5begin:m5end, 1:count1);
m52_5 = k5_SV_dist (m5begin:m5end, m2begin:m2end);
```

```
m53_5 = k5_SV_dist (m5begin:m5end, m3begin:m3end);
m54_5 = k5_SV_dist (m5begin:m5end, m4begin:m4end);
m55_5 = k5_SV_dist (m5begin:m5end, m5begin:m5end);
m11_5_wiAvg = sum(m11_5,2)/(size(m11_5,2)-1);
m22_5_wiAvg = sum(m22_5,2)/(size(m22_5,2)-1);
m33_5_wiAvg = sum(m33_5,2)/(size(m33_5,2)-1);
m44_5_wiAvg = sum(m44_5,2)/(size(m44_5,2)-1);
m55_5_wiAvg = sum(m55_5,2)/(size(m55_5,2)-1);
m1_5_bw = [m12_5 m13_5 m14_5 m15_5];
m2_5_bw = [m21_5 m23_5 m24_5 m25_5];
m3_5_bw = [m31_5 m32_5 m34_5 m35_5];
m4_5_bw = [m41_5 m42_5 m43_5 m45_5];
m5_5_bw = [m51_5 m52_5 m53_5 m54_5];
m1_5_bwAvg = mean(m1_5_bw,2);
m2_5_bwAvg = mean(m2_5_bw,2);
m3_5_bwAvg = mean(m3_5_bw,2);
m4_5_bwAvg = mean(m4_5_bw,2);
m5_5_bwAvg = mean(m5_5_bw,2);
k5_1 = ones(count1,1);
k5_2 = 2*ones(count2,1);
k5_3 = 3*ones(count3,1);
k5_4 = 4*ones(count4,1);
k5_5 = 5*ones(count5,1);
m1_5_WB = [k5_1 m11_5_wiAvg m1_5_bwAvg];
m2_5_WB = [k5_2 m22_5_wiAvg m2_5_bwAvg];
m3_5_WB = [k5_3 m33_5_wiAvg m3_5_bwAvg];
m4_5_WB = [k5_4 m44_5_wiAvg m4_5_bwAvg];
m5_5_WB = [k5_5 m55_5_wiAvg m5_5_bwAvg];
k5_WB = [m1_5_WB; m2_5_WB; m3_5_WB; m4_5_WB; m5_5_WB];
% clear the memory of unneeded variables

% k = 6
% calculate the size of each cluster
count1 = Count(k6, '==1');
count2 = Count(k6, '==2');
count3 = Count(k6, '==3');
count4 = Count(k6, '==4');
count5 = Count(k6, '==5');
count6 = Count(k6, '==6');
% figure out where each cluster begins and ends within the similarity matrix
m2begin = count1 + 1;
m2end = count1 + count2;
m3begin = m2end + 1;
m3end = m2end + count3;
m4begin = m3end + 1;
m4end = m3end + count4;
```

```
m5begin = m4end + 1;
m5end = m4end + count5;
m6begin = m5end + 1;
m6end = m5end + count6;
% calculate components of the WB matrix
m11_6 = k6_SV_dist (1:count1, 1:count1);
m12_6 = k6_SV_dist (1:count1, m2begin:m2end);
m13_6 = k6_SV_dist (1:count1, m3begin:m3end);
m14_6 = k6_SV_dist (1:count1, m4begin:m4end);
m15_6 = k6_SV_dist (1:count1, m5begin:m5end);
m16_6 = k6_SV_dist (1:count1, m6begin:m6end);
m21_6 = k6_SV_dist (m2begin:m2end, 1:count1);
m22_6 = k6_SV_dist (m2begin:m2end, m2begin:m2end);
m23_6 = k6_SV_dist (m2begin:m2end, m3begin:m3end);
m24_6 = k6_SV_dist (m2begin:m2end, m4begin:m4end);
m25_6 = k6_SV_dist (m2begin:m2end, m5begin:m5end);
m26_6 = k6_SV_dist (m2begin:m2end, m6begin:m6end);
m31_6 = k6_SV_dist (m3begin:m3end, 1:count1);
m32_6 = k6_SV_dist (m3begin:m3end, m2begin:m2end);
m33_6 = k6_SV_dist (m3begin:m3end, m3begin:m3end);
m34_6 = k6_SV_dist (m3begin:m3end, m4begin:m4end);
m35_6 = k6_SV_dist (m3begin:m3end, m5begin:m5end);
m36_6 = k6_SV_dist (m3begin:m3end, m6begin:m6end);
m41_6 = k6_SV_dist (m4begin:m4end, 1:count1);
m42_6 = k6_SV_dist (m4begin:m4end, m2begin:m2end);
m43_6 = k6_SV_dist (m4begin:m4end, m3begin:m3end);
m44_6 = k6_SV_dist (m4begin:m4end, m4begin:m4end);
m45_6 = k6_SV_dist (m4begin:m4end, m5begin:m5end);
m46_6 = k6_SV_dist (m4begin:m4end, m6begin:m6end);
m51_6 = k6_SV_dist (m5begin:m5end, 1:count1);
m52_6 = k6_SV_dist (m5begin:m5end, m2begin:m2end);
m53_6 = k6_SV_dist (m5begin:m5end, m3begin:m3end);
m54_6 = k6_SV_dist (m5begin:m5end, m4begin:m4end);
m55_6 = k6_SV_dist (m5begin:m5end, m5begin:m5end);
m56_6 = k6_SV_dist (m5begin:m5end, m6begin:m6end);
m61_6 = k6_SV_dist (m6begin:m6end, 1:count1);
m62_6 = k6_SV_dist (m6begin:m6end, m2begin:m2end);
m63_6 = k6_SV_dist (m6begin:m6end, m3begin:m3end);
m64_6 = k6_SV_dist (m6begin:m6end, m4begin:m4end);
m65_6 = k6_SV_dist (m6begin:m6end, m5begin:m5end);
m66_6 = k6_SV_dist (m6begin:m6end, m6begin:m6end);
m11_6_wiAvg = sum(m11_6,2)/(size(m11_6,2)-1);
m22_6_wiAvg = sum(m22_6,2)/(size(m22_6,2)-1);
m33_6_wiAvg = sum(m33_6,2)/(size(m33_6,2)-1);
m44_6_wiAvg = sum(m44_6,2)/(size(m44_6,2)-1);
m55_6_wiAvg = sum(m55_6,2)/(size(m55_6,2)-1);
```

```
m66_6_wiAvg = sum(m66_6,2)/(size(m66_6,2)-1);
m1_6_bw = [m12_6 m13_6 m14_6 m15_6 m16_6];
m2_6_bw = [m21_6 m23_6 m24_6 m25_6 m26_6];
m3_6_bw = [m31_6 m32_6 m34_6 m35_6 m36_6];
m4_6_bw = [m41_6 m42_6 m43_6 m45_6 m46_6];
m5_6_bw = [m51_6 m52_6 m53_6 m54_6 m56_6];
m6_6_bw = [m61_6 m62_6 m63_6 m64_6 m65_6];
m1_6_bwAvg = mean(m1_6_bw,2);
m2_6_bwAvg = mean(m2_6_bw,2);
m3_6_bwAvg = mean(m3_6_bw,2);
m4_6_bwAvg = mean(m4_6_bw,2);
m5_6_bwAvg = mean(m5_6_bw,2);
m6_6_bwAvg = mean(m6_6_bw,2);
k6_1 = ones(count1,1);
k6_2 = 2*ones(count2,1);
k6_3 = 3*ones(count3,1);
k6_4 = 4*ones(count4,1);
k6_5 = 5*ones(count5,1);
k6_6 = 6*ones(count6,1);
m1_6_WB = [k6_1 m11_6_wiAvg m1_6_bwAvg];
m2_6_WB = [k6_2 m22_6_wiAvg m2_6_bwAvg];
m3_6_WB = [k6_3 m33_6_wiAvg m3_6_bwAvg];
m4_6_WB = [k6_4 m44_6_wiAvg m4_6_bwAvg];
m5_6_WB = [k6_5 m55_6_wiAvg m5_6_bwAvg];
m6_6_WB = [k6_6 m66_6_wiAvg m6_6_bwAvg];
k6_WB = [m1_6_WB; m2_6_WB; m3_6_WB; m4_6_WB; m5_6_WB; m6_6_WB];
% clear the memory of unneeded variables

% k = 7
% calculate the size of each cluster
count1 = Count(k7, '==1');
count2 = Count(k7, '==2');
count3 = Count(k7, '==3');
count4 = Count(k7, '==4');
count5 = Count(k7, '==5');
count6 = Count(k7, '==6');
count7 = Count(k7, '==7');
% figure out where each cluster begins and ends within the similarity matrix
m2begin = count1 + 1;
m2end = count1 + count2;
m3begin = m2end + 1;
m3end = m2end + count3;
m4begin = m3end + 1;
m4end = m3end + count4;
m5begin = m4end + 1;
m5end = m4end + count5;
```

```
m6begin = m5end + 1;
m6end = m5end + count6;
m7begin = m6end + 1;
m7end = m6end + count7;
% calculate components of the WB matrix
m11_7 = k7_SV_dist (1:count1, 1:count1);
m12_7 = k7_SV_dist (1:count1, m2begin:m2end);
m13_7 = k7_SV_dist (1:count1, m3begin:m3end);
m14_7 = k7_SV_dist (1:count1, m4begin:m4end);
m15_7 = k7_SV_dist (1:count1, m5begin:m5end);
m16_7 = k7_SV_dist (1:count1, m6begin:m6end);
m17_7 = k7_SV_dist (1:count1, m7begin:m7end);
m21_7 = k7_SV_dist (m2begin:m2end, 1:count1);
m22_7 = k7_SV_dist (m2begin:m2end, m2begin:m2end);
m23_7 = k7_SV_dist (m2begin:m2end, m3begin:m3end);
m24_7 = k7_SV_dist (m2begin:m2end, m4begin:m4end);
m25_7 = k7_SV_dist (m2begin:m2end, m5begin:m5end);
m26_7 = k7_SV_dist (m2begin:m2end, m6begin:m6end);
m27_7 = k7_SV_dist (m2begin:m2end, m7begin:m7end);
m31_7 = k7_SV_dist (m3begin:m3end, 1:count1);
m32_7 = k7_SV_dist (m3begin:m3end, m2begin:m2end);
m33_7 = k7_SV_dist (m3begin:m3end, m3begin:m3end);
m34_7 = k7_SV_dist (m3begin:m3end, m4begin:m4end);
m35_7 = k7_SV_dist (m3begin:m3end, m5begin:m5end);
m36_7 = k7_SV_dist (m3begin:m3end, m6begin:m6end);
m37_7 = k7_SV_dist (m3begin:m3end, m7begin:m7end);
m41_7 = k7_SV_dist (m4begin:m4end, 1:count1);
m42_7 = k7_SV_dist (m4begin:m4end, m2begin:m2end);
m43_7 = k7_SV_dist (m4begin:m4end, m3begin:m3end);
m44_7 = k7_SV_dist (m4begin:m4end, m4begin:m4end);
m45_7 = k7_SV_dist (m4begin:m4end, m5begin:m5end);
m46_7 = k7_SV_dist (m4begin:m4end, m6begin:m6end);
m47_7 = k7_SV_dist (m4begin:m4end, m7begin:m7end);
m51_7 = k7_SV_dist (m5begin:m5end, 1:count1);
m52_7 = k7_SV_dist (m5begin:m5end, m2begin:m2end);
m53_7 = k7_SV_dist (m5begin:m5end, m3begin:m3end);
m54_7 = k7_SV_dist (m5begin:m5end, m4begin:m4end);
m55_7 = k7_SV_dist (m5begin:m5end, m5begin:m5end);
m56_7 = k7_SV_dist (m5begin:m5end, m6begin:m6end);
m57_7 = k7_SV_dist (m5begin:m5end, m7begin:m7end);
m61_7 = k7_SV_dist (m6begin:m6end, 1:count1);
m62_7 = k7_SV_dist (m6begin:m6end, m2begin:m2end);
m63_7 = k7_SV_dist (m6begin:m6end, m3begin:m3end);
m64_7 = k7_SV_dist (m6begin:m6end, m4begin:m4end);
m65_7 = k7_SV_dist (m6begin:m6end, m5begin:m5end);
m66_7 = k7_SV_dist (m6begin:m6end, m6begin:m6end);
```

```
m67_7 = k7_SV_dist (m6begin:m6end, m7begin:m7end);
m71_7 = k7_SV_dist (m7begin:m7end, 1:count1);
m72_7 = k7_SV_dist (m7begin:m7end, m2begin:m2end);
m73_7 = k7_SV_dist (m7begin:m7end, m3begin:m3end);
m74_7 = k7_SV_dist (m7begin:m7end, m4begin:m4end);
m75_7 = k7_SV_dist (m7begin:m7end, m5begin:m5end);
m76_7 = k7_SV_dist (m7begin:m7end, m6begin:m6end);
m77_7 = k7_SV_dist (m7begin:m7end, m7begin:m7end);
m11_7_wiAvg = sum(m11_7,2)/(size(m11_7,2)-1);
m22_7_wiAvg = sum(m22_7,2)/(size(m22_7,2)-1);
m33_7_wiAvg = sum(m33_7,2)/(size(m33_7,2)-1);
m44_7_wiAvg = sum(m44_7,2)/(size(m44_7,2)-1);
m55_7_wiAvg = sum(m55_7,2)/(size(m55_7,2)-1);
m66_7_wiAvg = sum(m66_7,2)/(size(m66_7,2)-1);
m77_7_wiAvg = sum(m77_7,2)/(size(m77_7,2)-1);
m1_7_bw = [m12_7 m13_7 m14_7 m15_7 m16_7 m17_7];
m2_7_bw = [m21_7 m23_7 m24_7 m25_7 m26_7 m27_7];
m3_7_bw = [m31_7 m32_7 m34_7 m35_7 m36_7 m37_7];
m4_7_bw = [m41_7 m42_7 m43_7 m45_7 m46_7 m47_7];
m5_7_bw = [m51_7 m52_7 m53_7 m54_7 m56_7 m57_7];
m6_7_bw = [m61_7 m62_7 m63_7 m64_7 m65_7 m67_7];
m7_7_bw = [m71_7 m72_7 m73_7 m74_7 m75_7 m76_7];
m1_7_bwAvg = mean(m1_7_bw,2);
m2_7_bwAvg = mean(m2_7_bw,2);
m3_7_bwAvg = mean(m3_7_bw,2);
m4_7_bwAvg = mean(m4_7_bw,2);
m5_7_bwAvg = mean(m5_7_bw,2);
m6_7_bwAvg = mean(m6_7_bw,2);
m7_7_bwAvg = mean(m7_7_bw,2);
k7_1 = ones(count1,1);
k7_2 = 2*ones(count2,1);
k7_3 = 3*ones(count3,1);
k7_4 = 4*ones(count4,1);
k7_5 = 5*ones(count5,1);
k7_6 = 6*ones(count6,1);
k7_7 = 7*ones(count7,1);
m1_7_WB = [k7_1 m11_7_wiAvg m1_7_bwAvg];
m2_7_WB = [k7_2 m22_7_wiAvg m2_7_bwAvg];
m3_7_WB = [k7_3 m33_7_wiAvg m3_7_bwAvg];
m4_7_WB = [k7_4 m44_7_wiAvg m4_7_bwAvg];
m5_7_WB = [k7_5 m55_7_wiAvg m5_7_bwAvg];
m6_7_WB = [k7_6 m66_7_wiAvg m6_7_bwAvg];
m7_7_WB = [k7_7 m77_7_wiAvg m7_7_bwAvg];
k7_WB = [m1_7_WB; m2_7_WB; m3_7_WB; m4_7_WB; m5_7_WB; m6_7_WB;
m7_7_WB];
% clear the memory of unneeded variables
```

```
% k = 8
% calculate the size of each cluster
count1 = Count(k8, '==1');
count2 = Count(k8, '==2');
count3 = Count(k8, '==3');
count4 = Count(k8, '==4');
count5 = Count(k8, '==5');
count6 = Count(k8, '==6');
count7 = Count(k8, '==7');
count8 = Count(k8, '==8');
% figure out where each cluster begins and ends within the similarity matrix
m2begin = count1 + 1;
m2end = count1 + count2;
m3begin = m2end + 1;
m3end = m2end + count3;
m4begin = m3end + 1;
m4end = m3end + count4;
m5begin = m4end + 1;
m5end = m4end + count5;
m6begin = m5end + 1;
m6end = m5end + count6;
m7begin = m6end + 1;
m7end = m6end + count7;
m8begin = m7end + 1;
m8end = m7end + count8;
% calculate components of the WB matrix
m11_8 = k8_SV_dist (1:count1, 1:count1);
m12_8 = k8_SV_dist (1:count1, m2begin:m2end);
m13_8 = k8_SV_dist (1:count1, m3begin:m3end);
m14_8 = k8_SV_dist (1:count1, m4begin:m4end);
m15_8 = k8_SV_dist (1:count1, m5begin:m5end);
m16_8 = k8_SV_dist (1:count1, m6begin:m6end);
m17_8 = k8_SV_dist (1:count1, m7begin:m7end);
m18_8 = k8_SV_dist (1:count1, m8begin:m8end);
m21_8 = k8_SV_dist (m2begin:m2end, 1:count1);
m22_8 = k8_SV_dist (m2begin:m2end, m2begin:m2end);
m23_8 = k8_SV_dist (m2begin:m2end, m3begin:m3end);
m24_8 = k8_SV_dist (m2begin:m2end, m4begin:m4end);
m25_8 = k8_SV_dist (m2begin:m2end, m5begin:m5end);
m26_8 = k8_SV_dist (m2begin:m2end, m6begin:m6end);
m27_8 = k8_SV_dist (m2begin:m2end, m7begin:m7end);
m28_8 = k8_SV_dist (m2begin:m2end, m8begin:m8end);
m31_8 = k8_SV_dist (m3begin:m3end, 1:count1);
m32_8 = k8_SV_dist (m3begin:m3end, m2begin:m2end);
m33_8 = k8_SV_dist (m3begin:m3end, m3begin:m3end);
```

```
m34_8 = k8_SV_dist (m3begin:m3end, m4begin:m4end);
m35_8 = k8_SV_dist (m3begin:m3end, m5begin:m5end);
m36_8 = k8_SV_dist (m3begin:m3end, m6begin:m6end);
m37_8 = k8_SV_dist (m3begin:m3end, m7begin:m7end);
m38_8 = k8_SV_dist (m3begin:m3end, m8begin:m8end);
m41_8 = k8_SV_dist (m4begin:m4end, 1:count1);
m42_8 = k8_SV_dist (m4begin:m4end, m2begin:m2end);
m43_8 = k8_SV_dist (m4begin:m4end, m3begin:m3end);
m44_8 = k8_SV_dist (m4begin:m4end, m4begin:m4end);
m45_8 = k8_SV_dist (m4begin:m4end, m5begin:m5end);
m46_8 = k8_SV_dist (m4begin:m4end, m6begin:m6end);
m47_8 = k8_SV_dist (m4begin:m4end, m7begin:m7end);
m48_8 = k8_SV_dist (m4begin:m4end, m8begin:m8end);
m51_8 = k8_SV_dist (m5begin:m5end, 1:count1);
m52_8 = k8_SV_dist (m5begin:m5end, m2begin:m2end);
m53_8 = k8_SV_dist (m5begin:m5end, m3begin:m3end);
m54_8 = k8_SV_dist (m5begin:m5end, m4begin:m4end);
m55_8 = k8_SV_dist (m5begin:m5end, m5begin:m5end);
m56_8 = k8_SV_dist (m5begin:m5end, m6begin:m6end);
m57_8 = k8_SV_dist (m5begin:m5end, m7begin:m7end);
m58_8 = k8_SV_dist (m5begin:m5end, m8begin:m8end);
m61_8 = k8_SV_dist (m6begin:m6end, 1:count1);
m62_8 = k8_SV_dist (m6begin:m6end, m2begin:m2end);
m63_8 = k8_SV_dist (m6begin:m6end, m3begin:m3end);
m64_8 = k8_SV_dist (m6begin:m6end, m4begin:m4end);
m65_8 = k8_SV_dist (m6begin:m6end, m5begin:m5end);
m66_8 = k8_SV_dist (m6begin:m6end, m6begin:m6end);
m67_8 = k8_SV_dist (m6begin:m6end, m7begin:m7end);
m68_8 = k8_SV_dist (m6begin:m6end, m8begin:m8end);
m71_8 = k8_SV_dist (m7begin:m7end, 1:count1);
m72_8 = k8_SV_dist (m7begin:m7end, m2begin:m2end);
m73_8 = k8_SV_dist (m7begin:m7end, m3begin:m3end);
m74_8 = k8_SV_dist (m7begin:m7end, m4begin:m4end);
m75_8 = k8_SV_dist (m7begin:m7end, m5begin:m5end);
m76_8 = k8_SV_dist (m7begin:m7end, m6begin:m6end);
m77_8 = k8_SV_dist (m7begin:m7end, m7begin:m7end);
m78_8 = k8_SV_dist (m7begin:m7end, m8begin:m8end);
m81_8 = k8_SV_dist (m8begin:m8end, 1:count1);
m82_8 = k8_SV_dist (m8begin:m8end, m2begin:m2end);
m83_8 = k8_SV_dist (m8begin:m8end, m3begin:m3end);
m84_8 = k8_SV_dist (m8begin:m8end, m4begin:m4end);
m85_8 = k8_SV_dist (m8begin:m8end, m5begin:m5end);
m86_8 = k8_SV_dist (m8begin:m8end, m6begin:m6end);
m87_8 = k8_SV_dist (m8begin:m8end, m7begin:m7end);
m88_8 = k8_SV_dist (m8begin:m8end, m8begin:m8end);
m11_8_wiAvg = sum(m11_8,2)/(size(m11_8,2)-1);
```

```
m22_8_wiAvg = sum(m22_8,2)/(size(m22_8,2)-1);
m33_8_wiAvg = sum(m33_8,2)/(size(m33_8,2)-1);
m44_8_wiAvg = sum(m44_8,2)/(size(m44_8,2)-1);
m55_8_wiAvg = sum(m55_8,2)/(size(m55_8,2)-1);
m66_8_wiAvg = sum(m66_8,2)/(size(m66_8,2)-1);
m77_8_wiAvg = sum(m77_8,2)/(size(m77_8,2)-1);
m88_8_wiAvg = sum(m88_8,2)/(size(m88_8,2)-1);
m1_8_bw = [m12_8 m13_8 m14_8 m15_8 m16_8 m17_8 m18_8];
m2_8_bw = [m21_8 m23_8 m24_8 m25_8 m26_8 m27_8 m28_8];
m3_8_bw = [m31_8 m32_8 m34_8 m35_8 m36_8 m37_8 m38_8];
m4_8_bw = [m41_8 m42_8 m43_8 m45_8 m46_8 m47_8 m48_8];
m5_8_bw = [m51_8 m52_8 m53_8 m54_8 m56_8 m57_8 m58_8];
m6_8_bw = [m61_8 m62_8 m63_8 m64_8 m65_8 m67_8 m68_8];
m7_8_bw = [m71_8 m72_8 m73_8 m74_8 m75_8 m76_8 m78_8];
m8_8_bw = [m81_8 m82_8 m83_8 m84_8 m85_8 m86_8 m87_8];
m1_8_bwAvg = mean(m1_8_bw,2);
m2_8_bwAvg = mean(m2_8_bw,2);
m3_8_bwAvg = mean(m3_8_bw,2);
m4_8_bwAvg = mean(m4_8_bw,2);
m5_8_bwAvg = mean(m5_8_bw,2);
m6_8_bwAvg = mean(m6_8_bw,2);
m7_8_bwAvg = mean(m7_8_bw,2);
m8_8_bwAvg = mean(m8_8_bw,2);
k8_1 = ones(count1,1);
k8_2 = 2*ones(count2,1);
k8_3 = 3*ones(count3,1);
k8_4 = 4*ones(count4,1);
k8_5 = 5*ones(count5,1);
k8_6 = 6*ones(count6,1);
k8_7 = 7*ones(count7,1);
k8_8 = 8*ones(count8,1);
m1_8_WB = [k8_1 m11_8_wiAvg m1_8_bwAvg];
m2_8_WB = [k8_2 m22_8_wiAvg m2_8_bwAvg];
m3_8_WB = [k8_3 m33_8_wiAvg m3_8_bwAvg];
m4_8_WB = [k8_4 m44_8_wiAvg m4_8_bwAvg];
m5_8_WB = [k8_5 m55_8_wiAvg m5_8_bwAvg];
m6_8_WB = [k8_6 m66_8_wiAvg m6_8_bwAvg];
m7_8_WB = [k8_7 m77_8_wiAvg m7_8_bwAvg];
m8_8_WB = [k8_8 m88_8_wiAvg m8_8_bwAvg];
k8_WB = [m1_8_WB; m2_8_WB; m3_8_WB; m4_8_WB; m5_8_WB; m6_8_WB;
m7_8_WB; m8_8_WB];
% clear the memory of unneeded variables

% k = 9
% calculate the size of each cluster
count1 = Count(k9, '==1');
```

```
count2 = Count(k9, '==2');
count3 = Count(k9, '==3');
count4 = Count(k9, '==4');
count5 = Count(k9, '==5');
count6 = Count(k9, '==6');
count7 = Count(k9, '==7');
count8 = Count(k9, '==8');
count9 = Count(k9, '==9');
% figure out where each cluster begins and ends within the similarity matrix
m2begin = count1 + 1;
m2end = count1 + count2;
m3begin = m2end + 1;
m3end = m2end + count3;
m4begin = m3end + 1;
m4end = m3end + count4;
m5begin = m4end + 1;
m5end = m4end + count5;
m6begin = m5end + 1;
m6end = m5end + count6;
m7begin = m6end + 1;
m7end = m6end + count7;
m8begin = m7end + 1;
m8end = m7end + count8;
m9begin = m8end + 1;
m9end = m8end + count9;
% calculate components of the WB matrix
m11_9 = k9_SV_dist (1:count1, 1:count1);
m12_9 = k9_SV_dist (1:count1, m2begin:m2end);
m13_9 = k9_SV_dist (1:count1, m3begin:m3end);
m14_9 = k9_SV_dist (1:count1, m4begin:m4end);
m15_9 = k9_SV_dist (1:count1, m5begin:m5end);
m16_9 = k9_SV_dist (1:count1, m6begin:m6end);
m17_9 = k9_SV_dist (1:count1, m7begin:m7end);
m18_9 = k9_SV_dist (1:count1, m8begin:m8end);
m19_9 = k9_SV_dist (1:count1, m9begin:m9end);
m21_9 = k9_SV_dist (m2begin:m2end, 1:count1);
m22_9 = k9_SV_dist (m2begin:m2end, m2begin:m2end);
m23_9 = k9_SV_dist (m2begin:m2end, m3begin:m3end);
m24_9 = k9_SV_dist (m2begin:m2end, m4begin:m4end);
m25_9 = k9_SV_dist (m2begin:m2end, m5begin:m5end);
m26_9 = k9_SV_dist (m2begin:m2end, m6begin:m6end);
m27_9 = k9_SV_dist (m2begin:m2end, m7begin:m7end);
m28_9 = k9_SV_dist (m2begin:m2end, m8begin:m8end);
m29_9 = k9_SV_dist (m2begin:m2end, m9begin:m9end);
m31_9 = k9_SV_dist (m3begin:m3end, 1:count1);
m32_9 = k9_SV_dist (m3begin:m3end, m2begin:m2end);
```

```
m33_9 = k9_SV_dist (m3begin:m3end, m3begin:m3end);
m34_9 = k9_SV_dist (m3begin:m3end, m4begin:m4end);
m35_9 = k9_SV_dist (m3begin:m3end, m5begin:m5end);
m36_9 = k9_SV_dist (m3begin:m3end, m6begin:m6end);
m37_9 = k9_SV_dist (m3begin:m3end, m7begin:m7end);
m38_9 = k9_SV_dist (m3begin:m3end, m8begin:m8end);
m39_9 = k9_SV_dist (m3begin:m3end, m9begin:m9end);
m41_9 = k9_SV_dist (m4begin:m4end, 1:count1);
m42_9 = k9_SV_dist (m4begin:m4end, m2begin:m2end);
m43_9 = k9_SV_dist (m4begin:m4end, m3begin:m3end);
m44_9 = k9_SV_dist (m4begin:m4end, m4begin:m4end);
m45_9 = k9_SV_dist (m4begin:m4end, m5begin:m5end);
m46_9 = k9_SV_dist (m4begin:m4end, m6begin:m6end);
m47_9 = k9_SV_dist (m4begin:m4end, m7begin:m7end);
m48_9 = k9_SV_dist (m4begin:m4end, m8begin:m8end);
m49_9 = k9_SV_dist (m4begin:m4end, m9begin:m9end);
m51_9 = k9_SV_dist (m5begin:m5end, 1:count1);
m52_9 = k9_SV_dist (m5begin:m5end, m2begin:m2end);
m53_9 = k9_SV_dist (m5begin:m5end, m3begin:m3end);
m54_9 = k9_SV_dist (m5begin:m5end, m4begin:m4end);
m55_9 = k9_SV_dist (m5begin:m5end, m5begin:m5end);
m56_9 = k9_SV_dist (m5begin:m5end, m6begin:m6end);
m57_9 = k9_SV_dist (m5begin:m5end, m7begin:m7end);
m58_9 = k9_SV_dist (m5begin:m5end, m8begin:m8end);
m59_9 = k9_SV_dist (m5begin:m5end, m9begin:m9end);
m61_9 = k9_SV_dist (m6begin:m6end, 1:count1);
m62_9 = k9_SV_dist (m6begin:m6end, m2begin:m2end);
m63_9 = k9_SV_dist (m6begin:m6end, m3begin:m3end);
m64_9 = k9_SV_dist (m6begin:m6end, m4begin:m4end);
m65_9 = k9_SV_dist (m6begin:m6end, m5begin:m5end);
m66_9 = k9_SV_dist (m6begin:m6end, m6begin:m6end);
m67_9 = k9_SV_dist (m6begin:m6end, m7begin:m7end);
m68_9 = k9_SV_dist (m6begin:m6end, m8begin:m8end);
m69_9 = k9_SV_dist (m6begin:m6end, m9begin:m9end);
m71_9 = k9_SV_dist (m7begin:m7end, 1:count1);
m72_9 = k9_SV_dist (m7begin:m7end, m2begin:m2end);
m73_9 = k9_SV_dist (m7begin:m7end, m3begin:m3end);
m74_9 = k9_SV_dist (m7begin:m7end, m4begin:m4end);
m75_9 = k9_SV_dist (m7begin:m7end, m5begin:m5end);
m76_9 = k9_SV_dist (m7begin:m7end, m6begin:m6end);
m77_9 = k9_SV_dist (m7begin:m7end, m7begin:m7end);
m78_9 = k9_SV_dist (m7begin:m7end, m8begin:m8end);
m79_9 = k9_SV_dist (m7begin:m7end, m9begin:m9end);
m81_9 = k9_SV_dist (m8begin:m8end, 1:count1);
m82_9 = k9_SV_dist (m8begin:m8end, m2begin:m2end);
m83_9 = k9_SV_dist (m8begin:m8end, m3begin:m3end);
```

```
m84_9 = k9_SV_dist (m8begin:m8end, m4begin:m4end);
m85_9 = k9_SV_dist (m8begin:m8end, m5begin:m5end);
m86_9 = k9_SV_dist (m8begin:m8end, m6begin:m6end);
m87_9 = k9_SV_dist (m8begin:m8end, m7begin:m7end);
m88_9 = k9_SV_dist (m8begin:m8end, m8begin:m8end);
m89_9 = k9_SV_dist (m8begin:m8end, m9begin:m9end);
m91_9 = k9_SV_dist (m9begin:m9end, 1:count1);
m92_9 = k9_SV_dist (m9begin:m9end, m2begin:m2end);
m93_9 = k9_SV_dist (m9begin:m9end, m3begin:m3end);
m94_9 = k9_SV_dist (m9begin:m9end, m4begin:m4end);
m95_9 = k9_SV_dist (m9begin:m9end, m5begin:m5end);
m96_9 = k9_SV_dist (m9begin:m9end, m6begin:m6end);
m97_9 = k9_SV_dist (m9begin:m9end, m7begin:m7end);
m98_9 = k9_SV_dist (m9begin:m9end, m8begin:m8end);
m99_9 = k9_SV_dist (m9begin:m9end, m9begin:m9end);
m11_9_wiAvg = sum(m11_9,2)/(size(m11_9,2)-1);
m22_9_wiAvg = sum(m22_9,2)/(size(m22_9,2)-1);
m33_9_wiAvg = sum(m33_9,2)/(size(m33_9,2)-1);
m44_9_wiAvg = sum(m44_9,2)/(size(m44_9,2)-1);
m55_9_wiAvg = sum(m55_9,2)/(size(m55_9,2)-1);
m66_9_wiAvg = sum(m66_9,2)/(size(m66_9,2)-1);
m77_9_wiAvg = sum(m77_9,2)/(size(m77_9,2)-1);
m88_9_wiAvg = sum(m88_9,2)/(size(m88_9,2)-1);
m99_9_wiAvg = sum(m99_9,2)/(size(m99_9,2)-1);
m1_9_bw = [m12_9 m13_9 m14_9 m15_9 m16_9 m17_9 m18_9 m19_9];
m2_9_bw = [m21_9 m23_9 m24_9 m25_9 m26_9 m27_9 m28_9 m29_9];
m3_9_bw = [m31_9 m32_9 m34_9 m35_9 m36_9 m37_9 m38_9 m39_9];
m4_9_bw = [m41_9 m42_9 m43_9 m45_9 m46_9 m47_9 m48_9 m49_9];
m5_9_bw = [m51_9 m52_9 m53_9 m54_9 m56_9 m57_9 m58_9 m59_9];
m6_9_bw = [m61_9 m62_9 m63_9 m64_9 m65_9 m67_9 m68_9 m69_9];
m7_9_bw = [m71_9 m72_9 m73_9 m74_9 m75_9 m76_9 m78_9 m79_9];
m8_9_bw = [m81_9 m82_9 m83_9 m84_9 m85_9 m86_9 m87_9 m89_9];
m9_9_bw = [m91_9 m92_9 m93_9 m94_9 m95_9 m96_9 m97_9 m98_9];
m1_9_bwAvg = mean(m1_9_bw,2);
m2_9_bwAvg = mean(m2_9_bw,2);
m3_9_bwAvg = mean(m3_9_bw,2);
m4_9_bwAvg = mean(m4_9_bw,2);
m5_9_bwAvg = mean(m5_9_bw,2);
m6_9_bwAvg = mean(m6_9_bw,2);
m7_9_bwAvg = mean(m7_9_bw,2);
m8_9_bwAvg = mean(m8_9_bw,2);
m9_9_bwAvg = mean(m9_9_bw,2);
k9_1 = ones(count1,1);
k9_2 = 2*ones(count2,1);
k9_3 = 3*ones(count3,1);
k9_4 = 4*ones(count4,1);
```

```
k9_5 = 5*ones(count5,1);
k9_6 = 6*ones(count6,1);
k9_7 = 7*ones(count7,1);
k9_8 = 8*ones(count8,1);
k9_9 = 9*ones(count9,1);
m1_9_WB = [k9_1 m11_9_wiAvg m1_9_bwAvg];
m2_9_WB = [k9_2 m22_9_wiAvg m2_9_bwAvg];
m3_9_WB = [k9_3 m33_9_wiAvg m3_9_bwAvg];
m4_9_WB = [k9_4 m44_9_wiAvg m4_9_bwAvg];
m5_9_WB = [k9_5 m55_9_wiAvg m5_9_bwAvg];
m6_9_WB = [k9_6 m66_9_wiAvg m6_9_bwAvg];
m7_9_WB = [k9_7 m77_9_wiAvg m7_9_bwAvg];
m8_9_WB = [k9_8 m88_9_wiAvg m8_9_bwAvg];
m9_9_WB = [k9_9 m99_9_wiAvg m9_9_bwAvg];
k9_WB = [m1_9_WB; m2_9_WB; m3_9_WB; m4_9_WB; m5_9_WB; m6_9_WB;
m7_9_WB; m8_9_WB; m9_9_WB];
% clear the memory of unneeded variables

% Combine all the WB ratios into the final result
wbRatio = [k4_WB k5_WB k6_WB k7_WB k8_WB k9_WB];
```

Appendix D. Commercial Aviation Key Words

Top 20 Keywords for each of the 54 clusters from the 1st sample of CA documents:

CA1 Set 1	CA1 Set 2	CA1 Set 3	CA1 Set 4	CA1 Set 5	CA1 Set 6
FREQ	MY	RAMP	LAX	HOLDING	RPTR
COM	I	GATE	APCH	HOLD	CALLBACK
RADIO	THIS	GND	COMPLEX	PATTERN	REVEALED
VOLUME	YOU	TRUCK	SMO	PUBLISHED	CONVERSATION
VHF	IT	WINGTIP	CIVET	TURNS	FOLLOWING
CONTACT	CAPT	WING	LOC	FIX	INFO
HF	HIS	PARKED	BASE	FMS	WITH
FREQS	OEF	TAXI	SIGHT	VOR	HE
STANDBY	AIRLINE	PARKING	FINAL	INTXN	HAS
MIKE	AIRMAN	TAXIWAY	VISUAL	LEGS	FLC
OKC	DAY	ACFT	VIS	EFC	STATES
GSO	CHKLIST	PERSONNEL	N	OUTBOUND	THEY
NICOSIA	MGMNT	TUG	S	SPUDS	FUELR
CASPER	HE	FORWARD	BRASILIA	RYANN	STATED
PANEL	COCKPIT	AREA	RWY	RADIAL	TAHITI
PASRO	RESOURCE	PUSHBACK	RWYS	INBOUND	PLT
SWITCH	JOB	TAXIING	STADIUM	ENTERED	FAA
RADIOS	STALL	PUSH	ILS	POPPS	JUMPERS
POWAL	IS	MARSHALLER	SOCAL	ENTER	THAT
TUNED	HIM	STOPPED	SNORKEL	INSTRUCTIONS	FAIRY

CA1 Set 7	CA1 Set 8	CA1 Set 9	CA1 Set 10	CA1 Set 11
KTS	BRAKE	CABIN	WT	HOLD
SPD	SNOW	PRESSURIZATION	LBS	SHORT
AIRSPD	BRAKING	PACK	BAL	RWY
KIAS	DAMAGE	BLEED	LOAD	LINE
SLOW	LIGHTS	SWITCHES	FUEL	TXWY
SLOWED	BRAKES	AIR	TKOF	TAXI
#	PARKING	CONDITIONING	GROSS	TWR
FT	THE	APU	PAX	LINES
FLAPS	TIRES	PACKS	PAPERWORK	STOPPED
APCH	EDGE	OXYGEN	MANIFEST	GND
MARKER	CTRLINE	BLEEDS	MAX	ONTO
SLOWING	ACFT	PRESSURE	CARGO	TAXIING
AUTOTHROTTLES	TIRE	MASKS	GRAVITY	ACROSS
RESTR	MAIN	HORN	WTS	STOP
ATC	PRESSURE	NORMAL	DATA	TAXIED
DSCNT	RWY	AUTO	FORM	POS
THROTTLES	NOSE	CHKLIST	BAGS	PAST
HIGH	NOSEWHEEL	EMER	OVERWT	CROSSED
COMPLY	STOP	TKOF	LIMITS	JLN
BELOW	ACCUMULATOR	OFF	INDEX	CLRED

CA1 Set 12	CA1 Set 13	CA1 Set 14	CA1 Set 15	CA1 Set 16
HDG	RWY	TAIL	SFO	RWY
DEGS	TAXI	TAILSKID	BRIDGE	SHORT
TURN	GND	LNDG	BAY	HOLD
DEP	CLRNC	STRIKE	APCH	TWR
DEG	CROSS	TOUCHDOWN	VISUAL	LAND
L	SHORT	ROTATION	MATEO	LAHSO
#	ACTIVE	NOSE	SAN	ORD
R	HOLD	NORMAL	RWY	PAPA
CTLR	INSTRUCTIONS	PITCH	VIS	LAIKE
TURNED	ACROSS	INSPECTION	PRM	LNDG
HEADING	XING	WIND	QUIET	FOREIGN
ZMA	TAXIING	DAMAGE	ARCHI	TAXI
ASSIGNED	TXWY	AFT	STUDENT	RWYS
CLB	CROSSED	KTS	TOE	INSTRUCTIONS
TURNING	TWR	MAINT	TIPP	ALT
CLRNC	RAMP	NOISE	HWD	FT
MAINTAIN	CTL	WINDS	FMS	
GYRO	METERING	SKID	SAMUL	
DIRECTION	TAXIED	SCRAPED	APCHS	
STINSON	READ	ATTITUDE	BRASILIA	

CA1 Set 17	CA1 Set 18	CA1 Set 19	CA1 Set 20	CA1 Set 21	CA1 Set 22
FT	HELI	DOOR	TCASII	FLAPS	TKOF
CLRNC	BRIDGE	ATTENDANT	RA	TKOF	TWR
CTLR	VENT	FLT	TFC	FLAP	RWY
READ	VOLCANO	HER	CLB	CHKLIST	POS
ALT	WING	ATTENDANTS	FT	WARNING	ROLL
#	TOUR	PAX	TA	HORN	CLRED
BACK	CHOPPER	SHE	O'CLOCK	THRUST	HOLD
DSND	KETCHIKAN	COCKPIT	#	DEGS	CLRNC
READBACK	FLOAT	SEAT	TARGET	TRIM	FOR
MAINTAIN	HILO	THE	CLBING	CONFIGN	ACR
CLRED	AREA	CABIN	AT	HANDLE	ABORT
SAID	OPERATORS	PA	ATC	NORMAL	TAXIED
CALL	ANNOUNCED	OPEN	VISUALLY	RETRACTED	READY
HEARD	PERMISSION	DOCTOR	DSND	LEADING	ONTO
DISCRETION	MY	PAIN	CONFLICT	SLATS	ABORTED
CTR	SFAR	DOORS	VERT	LEVER	BBA
HE	MONUMENT	SEATED	FPM	EXTENDED	HEARD
ARR	MOVIE	AGENT	RECEIVED	THROTTLES	END
RAMMS	COMMUNITY	SLIDE	FOLLOWED	EDGE	INTO
LEVEL	TONGASS	EMER	US	SPOILERS	Z

CA1 Set 23	CA1 Set 24	CA1 Set 25	CA1 Set 26	CA1 Set 27	CA1 Set 28
DEP	GEAR	TWR	RESTR	ENG	INSPECTOR
SID	PIN	APCH	DSCNT	START	SEAT
TURN	LNDG	LNDG	XING	CHKLIST	FAA
HDG	DOWN	LAND	CROSS	SHUT	JUMP
PROC	NOSE	RWY	FT	COWLING	COCKPIT
NOISE	PINS	FREQ	ARR	THE	MASK
DME	MAINT	LANDED	INTXN	MAINT	PAX
ABATEMENT	DOORS	CTL	AT	ICE	MEDICAL
TKOF	MAIN	CLRED	MAKE	EGT	OXYGEN
DEG	RETRACT	CONTACT	FMC	LEVER	HE
DEGS	WARNING	FINAL	RESTRS	SWITCH	LICENSE
SIDS	HANDLE	SWITCH	FMS	ANTI	HIS
LAS	THE	SWITCHED	NM	FIRE	AGENT
ROPPR	HORN	GAR	PROFILE	RESTART	CERTIFICATE
DEPS	CIRCUIT	CLRNC	CROSSED	IGNITION	MY
TEB	HATCH	MARKER	#	PWR	FLT
MOONY	EMER	VISUAL	ALT	HEAT	RIDE
CLRNC	CHKLIST	CONTACTED	VNAV	PROBE	PAPERWORK
INITIAL	EXTENSION	AGL	FO	RUNNING	COMPANY
BRIEFED	LOCKED	WITHOUT	ATC	TEMP	POI

CA1 Set 29	CA1 Set 30	CA1 Set 31	CA1 Set 32	CA1 Set 33	CA1 Set 34
RADIAL	RWY	SMA	TERRAIN	CENTER	ALTIMETER
DEG	UNICOM	TFC	GPWS	LARAMIE	SETTING
INTERCEPT	CTAF	X	WARNING	MONCTON	ALTIMETERS
HDG	DOWNWIND	Y	APCH	SMT	#
VOR	PATTERN	LEFT	PULL	BADDY	RESET
COURSE	ANNOUNCED	EVASIVE	WHOOP	ALT	FT
DME	FINAL	O'CLOCK	EED	RIGHT	SET
ARR	ARPT	SMT	MSL	ARWY	ALT
DCA	MY	COLLISION	WARNINGS	ZDV	QNH
GEP	BASE	SINGLE	BULLHEAD	DEVIATION	INCHES
#	STUDENT	LTT	IFP	LEFT	LOW
OUTBOUND	CESSNA	ACTION	THRUST	LITKY	LEVEL
DEGS	IFR	HE	SLOPE	BKX	SETTINGS
MILAM	I	MI	ACTIVATION	CHEYENNE	MILLIBARS
TRANSITION	INTENTIONS	SAW	FT	ROUTE	ATIS
BUF	RADIO	PASSED	MANEUVER	WBND	DSCNT
VECTOR	GAR	NEAR	VISUAL	O	LEVELED
TURN	UNCTLED	DOWNWIND	RIDGE	CENTER'S	RESETTING
WAVEY	HEARD	PLT	EGPWS	VICTOR	QFE
DIRECT	ACFT	BASE	RISING	COURSE	HG

CA1 Set 35	CA1 Set 36	CA1 Set 37	CA1 Set 38	CA1 Set 39	CA1 Set 40
CLOSED	DIRECT	AIRSPACE	RPTR	FUEL	MAINT
RWY	VOR	CLASS	RWY	LBS	MEL
NOTAMS	COURSE	TCA	TXWY	TANK	LOGBOOK
ATIS	NAV	B	TAXI	PUMPS	WRITE
ARPT	INS	FLOOR	CALLBACK	DISPATCH	LOG
NOTAM	HDG	VFR	HOLD	QUANTITY	INOP
WX	GPS	I	SHORT	TANKS	PREFLT
LIGHTS	AIRWAY	MY	REVEALED	GAUGES	INSPECTION
FSS	FMS	IFR	CONVERSATION	BOOST	MECH
DISPLACED	OMEGA	CONTACT	FOLLOWING	PUMP	APU
CLOSURE	DEG	MSL	ONTO	GAUGE	ITEMS
AVAILABLE	NEEDLE	ME	TAXIWAY	RESERVE	SIGNED
THRESHOLD	TRACK	AREA	LINES	POUNDS	DISPATCH
CONSTRUCTION	INTXN	SQUAWK	SIGNS	BURN	UPS
TFR	WERE	ZZZ	INFO	RELEASE	ITEM
PLANT	CTR	SWF	GND	BOARD	UP
TEMPORARY	ATC	DPC	INCURSION	ALTERNATE	ACFT
LIGHTING	WE	RADAR	HE	LOAD	OPEN
SDM	CDI	GCN	LIGHTS	FLT	ZZZ
LNDG	ROBRT	SFRA	TAXIING	FUELER	WRITTEN

CA1 Set 41	CA1 Set 42	CA1 Set 43	CA1 Set 44	CA1 Set 45
TFC	RTE	AUTOPLT	RESTRICTION	LOC
O'CLOCK	FILED	ALT	XING	APCH
VFR	ROUTING	FT	PROFILE	ILS
EVASIVE	DIRECT	ENGAGED	RESTRICTIONS	INTERCEPT
PASSED	PLAN	CAPTURE	DSNT	DME
ACFT	FLT	MODE	DME	GS
SIGHT	TRANSITION	DISCONNECTED	EFB	COURSE
AVOID	PAGE	DSCNT	FMC	PLATE
SAW	WAYPOINT	LEVELOFF	DSCNT	APCHS
SEPARATION	FMC	SELECT	MAKE	INTERCEPTED
ACTION	CLRNC	TRIM	CROSS	RWY
COLLISION	OCEAN	CLB	CIVET	FAF
NEAR	FIX	SELECTED	CENTER	ESTABLISHED
MISS	ROUTE	FPM	POM	MDA
OTHER	RELEASE	DIRECTOR	FIX	TUNED
APPEARED	FMS	DISENGAGED	CHART	NAV
PASS	LOADED	LEVEL	SPEED	VECTORED
CLBING	SID	ALERTER	VNAV	VOR
TWIN	DISPATCH	HAND	SPD	FOR
GLIDER	OUR	WHEEL	FIM	INTERCEPTING

CA1 Set 46	CA1 Set 47	CA1 Set 48	CA1 Set 49	CA1 Set 50
RVR	TXWY	ACR	PDC	TURB
VISIBILITY	TAXI	X	CODE	MODERATE
MINIMUMS	RWY	Y	DEP	ENCOUNTERED
APCH	GND	SECTOR	XPONDER	SEVERE
WX	ONTO	TCASII	SQUAWK	WAKE
CAT	RAMP	TFC	ACARS	WINDSHEAR
FOG	K	RA	CLRNC	ICE
LIGHTS	TAXIING	SEPARATION	PLAN	TSTMS
RPTED	TXWYS	HIM	EAGUL	CHOP
SPECS	B	ISSUED	TRANSPONDER	STORM
CTRLINE	SHORT	ATX	FLT	ENCOUNTER
III	H	HE	CLUE	AUTOPLT
ILS	INSTRUCTIONS	CLBING	RTE	WINDS
MI	HOLD	VFR	RECEIVED	FT
LEGAL	DIAGRAM	SMT	JOHNS	LIGHT
HT	L	HIS	CORRECT	CONTINUOUS
II	Q	WORKING	PDC'S	MICROBURST
RWY	VIA	WAS	ROUTING	ICING
ATIS	TAXIED	PLT	ST	CLOUDS
DECISION	P	MLG	CHK	CELLS

CA1 Set 51	CA1 Set 52	CA1 Set 53	CA1 Set 54
CLB	VISUAL	ALT	HRS
#	APCH	FT	DUTY
ALT	SIGHT	ALERTER	DAY
AT	RWY	ASSIGNED	REST
CLRNC	FINAL	#	TRIP
CTR	ARPT	ALERT	SCHEDULED
ATC	FIELD	WINDOW	HR
CRUISE	BASE	FLYING	FATIGUE
CLBING	US	SET	SCHEDULING
MACH	WE	DSNDED	CREW
WE	DOWNWIND	MSL	SLEEP
LEVEL	FOR	COPLT	DAYS
CGA	TWR	LEVEL	MINS
BACK	TFC	PNF	PERIOD
MOATT	VECTORED	PF	BLOCK
REQUESTED	LAND	DSCNT	TIME
MINS	LINED	LEVELOFF	HOTEL
CLBED	CLRED	FO	NIGHT
FANS	FOLLOW	THROUGH	TIRED
READ	ELLINGTON	DEV	LEGS

Top 20 Keywords for each of the 54 clusters from the 2nd sample of CA documents:

CA2 Set 1	CA2 Set 2	CA2 Set 3	CA2 Set 4	CA2 Set 5
SFO	ICE	TCASII	KTS	CABIN
VISUAL	ANTI	TFC	SPD	PRESSURIZATION
BRIDGE	HEAT	RA	AIRSPD	OXYGEN
VIS	SNOW	FT	SLOW	PRESSURE
APCH	DEICE	CLB	KIAS	MASKS
SIGHT	DEICING	O'CLOCK	#	VALVE
BAY	WINDSHIELD	TA	SLOWED	PACKS
SEP	CONTAMINATION	#	FT	OUTFLOW
RWY	WING	TARGET	SLOWING	MANUAL
TOE	PITOT	CLBING	RESTR	DIFFERENTIAL
CTRLINE	DEICED	FPM	DSCNT	PACK
MATEO	ENG	ATC	KT	MASK
QUIET	FREEZING	AT	ATC	PORTABLE
MAINTAIN	WINGS	VISUALLY	LIMIT	LIGHT
ARCHI	FLUID	DSND	BUFFET	PROB
SAN	COMPRESSOR	CONFLICT	ASSIGNED	EMER
MILL	TEMP	BELOW	MACH	BLEED
APCHS	PWR	RECEIVED	ACCELERATED	AUTO
BRIJJ	STALL	FOLLOWED	MAINTAIN	STARTER
TIP	WINDSCREEN	ADVISED	SPACING	STANDBY

CA2 Set 6	CA2 Set 7	CA2 Set 8	CA2 Set 9	CA2 Set 10
PDC	RVR	AUTOPLT	DIRECT	RESTR
DEP	VISIBILITY	ALT	VOR	XING
CLRNC	APCH	CAPTURE	GPS	DSCNT
SQUAWK	FOG	MODE	NAV	CROSS
XPONDER	MINIMUMS	FT	COURSE	MAKE
ACARS	CAT	SPD	OMEGA	AT
CODE	WX	ENGAGED	HDG	INTXN
OBTAINED	TOUCHDOWN	VERT	CHART	FT
PRE	LEGAL	IAS	DEGS	ARR
MESSAGE	MVY	DSCNT	INTXN	RESTRICTION
IAD	MI	SELECTED	CHESTER	FIX
SID	LIGHTS	DISCONNECTED	ADF	CTR
FORMAT	MID	FMA	ERROR	MI
DELIVERY	ROLLOUT	CLB	ENRTE	VNAV
FILED	BOS	DISENGAGED	ARWYS	GIVEN
MIA	HT	AUTOTHROTTLES	GIJ	SPD
PDC'S	III	MANUALLY	SLT	VIBES
OUR	DECISION	ACFT	DIR	HARTY
UPLINKED	RPTED	LEVEL	YNG	DME
REVISED	GS	AUTOTHROTTLE	PROCEEDING	PROFILE

CA2 Set 11	CA2 Set 12	CA2 Set 13	CA2 Set 14	CA2 Set 15	CA2 Set 16
PUSHBACK	ATTENDANT	MEL	RADIAL	CTLR	#
BRAKES	FLT	APU	DEG	FT	ATC
BRAKE	DOOR	MAINT	INTERCEPT	CLRNC	CTR
PUSH	SEAT	INOP	VOR	READ	CLB
TUG	PAX	ITEM	OUTBOUND	ALT	CLRNC
GND	HER	PLACARD	COURSE	READBACK	DSCNT
PARKING	COCKPIT	LOGBOOK	HDG	#	ALT
GATE	SHE	PACK	AIRWAY	BACK	ACR
START	ATTENDANTS	GENERATOR	DEGS	DSND	LEVEL
TOW	CABIN	ONS	DEP	DSCNT	DSND
CREW	MEDICAL	DISPATCH	#	SAID	READ
FORWARD	JUMP	DEFERRED	INTERCEPTED	HEARD	AT
ENGS	SEATS	RELEASE	INBOUND	RESPONDED	BACK
BAR	DEADHEADING	DEFERRAL	LNAV	MAINTAIN	SIMILAR
RELEASED	AFT	WRITE	ARR	CLRED	CLBING
JETWAY	PURSER	YAW	NAV	SIGN	ZLA
HYD	FAA	DAMPER	DQN	REPLIED	CRUISE
TAXI	DOCTOR	DC	CREPE	CALL	CGA
DAMAGE	INSPECTOR	INLET	MQO	CTR	ASKED
CHOCKS	CREW	BUS	SID	WE	REQUESTED

CA2 Set 17	CA2 Set 18	CA2 Set 19	CA2 Set 20	CA2 Set 21
RAMP	MAINT	ENG	FLAPS	FILED
GATE	LOGBOOK	START	FLAP	PLAN
TAXI	WRITE	FIRE	TKOF	RTE
GND	INSPECTION	SHUT	CHKLIST	DIRECT
PARKED	LOG	PWR	WARNING	ROUTING
WINGTIP	MECH	APU	HORN	CLRNC
TRUCK	AIRWORTHINESS	CHKLIST	TRIM	FLT
WING	ZZZ	PARKING	DEGS	PDC
PARKING	ENTRY	SHUTDOWN	HANDLE	GANDER
TAXIING	ACFT	IGNITION	CONFIGN	COMPUTER
ACFT	FERRY	FUEL	EXTENDED	ROUTE
STOP	DEFERRED	EGT	SOUNDED	AIRWAY
TAXIED	PREFLT	RUNNING	PWR	VIA
MARSHALLER	DISPATCH	LEVER	ABORT	TOPPS
DAMAGE	CLIPBOARD	GATE	RETRACTED	COORDINATES
SPOT	PERMIT	THE	CHKLISTS	CENTER
THE	OPEN	EVAC	DETENT	FMS
TIP	THAT	EXTERNAL	WINDSHEAR	NRT
BLAST	RELEASE	SWITCH	OVERSPD	OUR
AREA	MECHS	RPM	THROTTLES	SJU

CA2 Set 22	CA2 Set 23	CA2 Set 24	CA2 Set 25	CA2 Set 26
TAXI	ALTIMETER	SMA	LOC	TURB
RWY	SETTING	Y	APCH	WX
TAXIWAY	ALTIMETERS	TWR	ILS	MODERATE
GND	FT	RIGHT	INTERCEPT	RADAR
SHORT	#	TFC	GS	DEV
HOLD	SET	LGB	COURSE	TSTMS
CROSS	ATIS	TVC	CRP	CELLS
INSTRUCTIONS	ALT	TCA	INTERCEPTED	DEVIATING
ACTIVE	RESET	O'CLOCK	TUNED	CELL
ACROSS	METERS	EVASIVE	ESTABLISHED	TSTM
RWYS	LEVEL	LEFT	HDG	MACH
CROSSED	INCHES	MISS	CAPTURE	ZJX
OUTER	TRANSITION	COAST	REINTERCEPT	DAB
TAXIING	LCL	AN	NAV	ENCOUNTERED
TAXIWAYS	SETTINGS	PASSED	MSL	RETURNS
INNER	ELEVATION	DOWNWIND	ALIVE	AREA
ONTO	QNH	DRO	DOT	BUFFET
RAMP	HPA	X	RWY	UNABLE
XING	LEVELED	MLG	STADIUM	TOPS
PAPA	LOW	APPROX	FOR	HAIL

CA2 Set 27	CA2 Set 28	CA2 Set 29	CA2 Set 30	CA2 Set 31	CA2 Set 32
CTAF	FMC	TWR	TKOF	ACR	RWY
UNICOM	FIX	APCH	TWR	X	TWR
ANNOUNCED	PAGE	FREQ	RWY	Y	LAND
STUDENT	DATA	LAND	POS	#	FINAL
DOWNWIND	HOLDING	LNDG	CLRED	TFC	LNDG
RWY	SKEBR	CONTACT	FOR	CPR	GAR
INTENTIONS	LEGS	SWITCH	ROLL	XYZ	APCH
FINAL	ENTERED	LANDED	HOLD	TCASII	LINED
HELI	RAW	CLRED	READY	RA	DOWNWIND
CESSNA	MISEN	CTL	ABORT	HE	ON
BASE	RTE	CLRNC	CHKLIST	Z	LANDED
RADIO	BEENO	SWITCHED	TAXI	CLBING	BASE
TOUCH	PATTERN	WITHOUT	TAXIED	PLT	CLRED
ARPT	PROGRAMMED	GND	CLRNC	SECTOR	ILS
PATTERN	ARR	FREQUENCY	ONTO	ISSUED	VISUAL
COMMUTER	LOCKE	MARKER	ABORTED	FT	LIGHTS
CALLS	LKT	RADIO	ITEMS	OBSERVED	THRESHOLD
SBP	WAYPOINT	FORGOT	TOOK	WORKING	FOR
ANNOUNCING	DIRECT	ORH	IMMEDIATE	CLBED	EXECUTED
HELIJET	MGW	NEVER	TAXIING	LEVEL	WINDS

CA2 Set 33	CA2 Set 34	CA2 Set 35	CA2 Set 36	CA2 Set 37
GEAR	FMS	IFR	RPTR	WT
PIN	RTE	VFR	CALLBACK	AGENT
PINS	WAYPOINT	WX	REVEALED	RELEASE
NOSE	DIRECT	I	CONVERSATION	OPS
HANDLE	PROGRAMMED	CONDITIONS	FOLLOWING	PAX
RETRACT	LOADED	CLOUDS	INFO	FLT
DOWN	TRANSITION	VISIBILITY	FAA	BAGS
FLAG	COL	AIRSPACE	HAS	DISPATCH
LOCK	NAV	PLAN	HE	PAPERWORK
MAINT	COORDINATES	CEILING	WITH	DISPATCHER
MAIN	ENTERED	CLASS	JUMPERS	LBS
THE	ARR	ARPT	STATES	BAG
PREFLT	MAYAH	TCA	OWNER	INSPECTOR
INSTALLED	BTG	LAYER	THE	LOAD
DOOR	DEP	FSS	IS	CARGO
LNDG	WAYPOINTS	MY	STATED	MANIFEST
LOCKED	DATA	ME	HIS	SKID
WALKAROUND	PROGRAMMING	CANCEL	FEELS	FAA
RETRACTION	RNAV	OVCST	PLT	WTS
REMOVE	INS	NY	JOESS	BOARDED

CA2 Set 38	CA2 Set 39	CA2 Set 40	CA2 Set 41	CA2 Set 42	CA2 Set 43
DME	TCAS	I	VISUAL	DEP	TXWY
DEG	TFC	ME	APCH	SID	RWY
TURN	RA	HE	SIGHT	CLRNC	TAXI
DEP	CLB	THIS	ARPT	TURN	GND
#	CPR	IT	FINAL	PROC	ONTO
SID	DSNT	IS	RWY	HDG	RAMP
HDG	CLANG	YOU	BASE	SJC	TAXIING
RADIAL	AT	MY	DOWNWIND	TKOF	B
SEL	ATC	HIS	FIELD	CLB	TXWYS
VOR	FPM	CAPT	WE	POMONA	TAXIED
DEGS	AURAL	HIM	FOR	POM	END
SRP	ADVISORY	DO	TFC	DELIVERY	SHORT
NARRATIVE	WARNING	THEY	US	RESTRICTION	D
R	TA	HAVE	CLRED	CONEY	M
STEFE	O'CLOCK	JOB	FOLLOW	KARYN	INSTRUCTIONS
KIP	SECS	AM	VECTORED	ABATEMENT	P
SEA	RESOLUTION	DIDN'T	TWR	LCA	VIA
MOUNTAIN	LGT	DON'T	WERE	MAINTAIN	H
ABATEMENT	ADVISED	OUT	STEWART	TEB	C
AT	#	FMN	MAULE	RESTRICTIONS	DIAGRAM

CA2 Set 44	CA2 Set 45	CA2 Set 46	CA2 Set 47	CA2 Set 48	CA2 Set 49
LNDG	CIVET	TFC	HDG	ALT	HOLD
DAMAGE	ARR	VFR	DEGS	FT	SHORT
TOUCHDOWN	LAX	SMT	TURN	ASSIGNED	RWY
BRAKING	PROFILE	O'CLOCK	DEG	ALERTER	LINE
RUDDER	ARNES	NEAR	DEP	#	TAXI
NOSE	RESTRS	PASSED	#	CLBING	TXWY
WIND	BREMR	MISS	ASSIGNED	CLB	TWR
GEAR	MITTS	EVASIVE	COMPASS	ALERT	LINES
MAIN	FUELR	COLLISION	L	SET	INSTRUCTIONS
FLARE	DSNT	AVOID	TURNED	WINDOW	ACROSS
THE	ALTS	ACFT	R	THROUGH	STOPPED
KTS	INTXN	RADAR	NORCAL	FLYING	CROSSED
REVERSE	FT	SAW	GIVEN	LEVELOFF	CROSS
NORMAL	SOCAL	AIRSPACE	CLB	PF	TAXIING
BRAKES	HUNDA	TCA	UGA	MSL	STOP
AUTO	GS	IFR	ILE	KNOB	HOLDING
PWR	SNRKL	CLASS	CTLR	OUR	PAST
TAIL	ALT	PASS	VECTOR	DEV	POS
TAILSKID	LUXOR	MSL	MAINTAIN	LEVEL	TAXIED
RWY	SUZZI	TWIN	GAVE	LEVELED	INCURSION

CA2 Set 50	CA2 Set 51	CA2 Set 52	CA2 Set 53	CA2 Set 54
LAX	DME	HRS	FREQ	FUEL
RWY	APCH	DUTY	RADIO	LBS
APCH	ARC	TRIP	CENTER	TANK
VISUAL	VOR	DAY	CONTACT	GAUGES
COMPLEX	MDA	REST	COM	WT
SIGHT	TERRAIN	HR	VESAR	LOAD
SMO	ILS	SCHEDULING	FREQS	QUANTITY
SOCAL	PUBLISHED	SCHEDULED	NICOSIA	TANKS
LOC	PLATE	SLEEP	COMMERCIAL	XFEED
FINAL	FIX	FATIGUE	OCEANIC	DISPATCH
BRASILIA	GPWS	PERIOD	MIKE	GALLONS
VIS	TLC	TIME	INTERCOM	BURN
SMT	FAF	LEGAL	TRIED	GAUGE
RWYS	LDA	FLT	VOLLS	SLIP
MLG	MINIMUM	LEG	COMS	FUELER
BASE	ESTABLISHED	TIRED	VHF	POUNDS
PARALLEL	EXECUTED	DAYS	CHANGE	PUMP
HAZE	LOC	NIGHT	FIR	BOARD
WDB	ALT	LEGS	RADIOS	FUELED
SAAB	MSL	SCHEDULE	MINS	BAL

Top 20 Keywords for each of the 54 clusters from the 3rd sample of CA documents:

CA3 Set 1	CA3 Set 2	CA3 Set 3	CA3 Set 4	CA3 Set 5	CA3 Set 6
PDC	O'CLOCK	LAX	HRS	TCAS	SID
DEP	TFC	RWY	REST	CLB	DEP
CLRNC	EVASIVE	CIVET	DAY	TFC	DME
CODE	SAW	APCH	DUTY	TARGET	TURN
SQUAWK	PASSED	MITTS	SCHEDULED	RA	HDG
ACARS	ACTION	SOCAL	SLEEP	ATC	SJC
XPONDER	MISS	LOC	HR	CLBING	DEG
DELIVERY	COLLISION	ARR	FATIGUE	DSNT	#
SID	TWIN	FINAL	TRIP	O'CLOCK	ABATEMENT
RECEIVED	NEAR	COMPLEX	DAYS	TCASI	RESTRICTION
FLT	AT	PDZ	NIGHT	GULFSTREAM	NOISE
PRE	SMT	ARNES	SCHEDULING	RATE	PROC
PREDEP	RADAR	SMO	CREW	#	DEGS
FILED	GLIDER	VISUAL	REDUCED	LEVEL	CLB
PLAN	ACFT	SIGHT	LEGAL	FPM	BRIEFED
SFO	TCA	OVERSHOOT	MINS	TA	PUBLISHED
REVISED	AVOID	ILS	TIRED	ADVISORY	SIDS
PREFLT	SPC	SNRKL	PERIOD	II	LOUPE
WYLYY	SPOTTED	VIS	SCHEDULE	ISP	HEADING
OBTAINED	WING	HVT	HOTEL	FGT	ILSQ

CA3 Set 7	CA3 Set 8	CA3 Set 9	CA3 Set 10	CA3 Set 11
HDG	HOLDING	APCH	AUTOPLT	VOR
DEGS	HOLD	DME	ALT	COURSE
TURN	PATTERN	VOR	ENGAGED	DIRECT
DEP	TURNS	ILS	MODE	NAV
DEG	PUBLISHED	TERRAIN	FT	OMEGA
HEADING	BANK	PLATE	CAPTURE	INS
COMPASS	FIX	ARPT	LEVEL	ARR
BUG	EFC	MDA	DISCONNECTED	AIRWAY
CTLR	ENTERED	GPWS	ARMED	WAYPOINT
TURNED	PYE	ARC	PITCH	BLISS
ASSIGNED	TEDDY	GS	DISENGAGED	OUTBND
COMPASSES	TURN	APCHS	KNOB	TRACK
R	MERUE	FAF	SELECTED	HDG
US	ENTRY	FIELD	ARM	ESL
ATC	GPS	MISSED	TRIM	ERROR
L	INSTRUCTIONS	VISUAL	SELECT	DEG
DG	OUTBOUND	DSCNT	LEVELOFF	GTF
ESC	PRUNN	ZLC	ENGAGE	ARWY
ASKED	CWK	BRIEFED	RE	INTXN
ORF	ATC	PAPI	RATE	TUNED

CA3 Set 12	CA3 Set 13	CA3 Set 14	CA3 Set 15	CA3 Set 16
CABIN	TAXIWAY	RESTR	TOUCHDOWN	TCASII
PRESSURIZATION	OUTER	XING	BRAKING	RA
OXYGEN	GND	DSCNT	LNDG	TFC
HORN	TAXIING	CROSS	DAMAGE	CLB
PRESSURE	ZA	MAKE	NORMAL	FT
CHKLIST	TAXI	FT	SKID	#
MASKS	AUTHORITY	DME	TAIL	O'CLOCK
MASK	PORT	RESTRICTION	REVERSE	TARGET
BLEEDS	VEHICLES	INTXN	KTS	TA
PACKS	ONTO	AT	NOSE	ATC
OUTFLOW	UNIFORM	MI	MAIN	AT
SWITCH	GATE	ARR	WIND	CONFLICT
VALVE	CAR	FIX	GEAR	DSND
PANEL	TAXIWAYS	FMS	SPD	FPM
WARNING	CONCOURSE	RATE	ROTATION	CLBING
PACK	WDB	VOR	BRAKE	COMMAND
SMOKE	INNER	CROSSED	TAILSKID	CLBED
VALVES	ROAD	ATC	XWIND	ADVISED
MANUAL	MARKED	DYLIN	INSPECTION	BELOW
ALT	ACTIVE	PUTTZ	ACFT	FOLLOWED

CA3 Set 17	CA3 Set 18	CA3 Set 19	CA3 Set 20	CA3 Set 21	CA3 Set 22
FLT	FMC	MSL	FREQ	CLASS	#
PAX	DSCNT	FT	RADIO	AIRSPACE	CLB
ATTENDANT	VNAV	ALT	CALL	B	CENTER
COCKPIT	RESTR	#	COM	C	MACH
SEAT	XING	ASSIGNED	SIGN	ME	ALT
CAPT	PROGRAMMED	APCH	FREQS	VFR	ZAB
SHE	ARR	DSNDING	CONTACT	KENAI	CTLR
FAA	PAGE	CLRNC	CTLR	PWA	MINOW
HER	FIX	DSND	XXY	BFI	ZHU
INSPECTOR	MCP	ICING	ZTL	D	READ
DISPATCH	CROSS	DSNDED	COMS	MY	AIRSPD
ATTENDANTS	JAKSN	QUONSETT	ZXY	SHORTS	REQUESTED
THE	SKEBR	SANGSTER	FRA	DATE	ATC
REVISION	PROFILE	ALERTER	HEARD	I	WE
THAT	LEGS	DSCNT	XMISSIONS	ADIZ	BACK
CHK	SELECTED	BEARR	COMPANY	CERTIFICATE	BUFFET
DEICING	MODE	LEVEL	OCEANIC	TAC	MINS
STATION	BRUSR	TO	XMISSION	ANC	TO
BOARD	MISEN	THROUGH	MAIQUETIA	CLEVELAND	READBACK
COMPANY	DATA	SMO	SIMILAR	FLOOR	SHANNON

CA3 Set 23	CA3 Set 24	CA3 Set 26	CA3 Set 27	CA3 Set 28	CA3 Set 25
TXWY	GEAR	CTAF	LOC	KTS	ALTIMETER
TAXI	PINS	RWY	APCH	SPD	SETTING
RWY	PIN	UNICOM	ILS	AIRSPD	ALTIMETERS
RAMP	NOSE	ANNOUNCED	INTERCEPT	SLOW	#
GND	LNDG	ARPT	GS	SLOWED	FT
TXWYS	INSTALLED	UNCTLED	COURSE	#	RESET
ONTO	DOWN	PATTERN	TUNED	KIAS	SET
E	MAIN	FSS	RAW	KT	ATIS
B	LOCKED	INTENTIONS	FREQ	SLOWING	ALT
K	WARNING	CLOSED	HDG	STAR	DSCNT
TAXIING	HORN	SARATOGA	CDI	SPACING	LOW
MARKINGS	RETRACT	RADIO	IDENT	FT	INCHES
LIGHTS	MAINT	TAXIING	CAPTURED	TURB	CHKLIST
C	HANDLE	DEPART	INTERCEPTED	OVERSPD	SETTINGS
D	EXTENDED	LIGHTS	INTERCEPTING	DSCNT	CAPT'S
J	GREEN	DEPARTING	G	DIETZ	QNH
P	CHKLIST	CALLS	VECTOR	PWR	STANDBY
VIA	RED	DUCHESS	DATA	REDUCTION	LEVELED
INSTRUCTIONS	THE	VOID	VECTORED	ASSIGNED	LEVEL
ON	INSPECTION	OTHER	VECTORS	GRUNZ	EGPWS

CA3 Set 29	CA3 Set 30	CA3 Set 31	CA3 Set 32	CA3 Set 33	CA3 Set 34
TCASII	RVR	MAINT	PUSHBACK	TWR	SMA
RA	VISIBILITY	MEL	TUG	APCH	Y
TFC	MINIMUMS	LOGBOOK	START	LNDG	X
VISUAL	APCH	WRITE	GND	LAND	SMA'S
FT	CAT	ITEM	BRAKE	FREQ	EVASIVE
CLB	WX	DISPATCH	BRAKES	SWITCH	MLG
SEPARATION	FOG	MECH	PUSH	SWITCHED	STUDENT
SIGHT	RVV	DEFERRED	PARKING	CONTACT	PASSED
TA	RPTED	LOG	TOW	CLRED	ACFT
O'CLOCK	II	RELEASE	GATE	CTL	PATRICK
US	MI	PREFLT	BAR	LANDED	TFC
MAINTAIN	ILS	INOP	CREW	WE	RIGHT
BELOW	LEGAL	ENTRY	ENG	US	PRC
DSND	TKOF	UP	DRIVER	VISUAL	VIS
#	TWR	DISCREPANCY	SALUTE	GND	AVOID
MI	DECISION	ITEMS	ENGS	MARKER	CBE
CLBING	ALTERNATE	WRITTEN	RAMP	WITHOUT	ATA
TARGET	ATIS	THAT	PUSHED	FINAL	VFR
ISSUED	SPECS	ACFT	THE	RWY	SINGLE
APCH	REPORTED	FLT	RELEASED	OM	O'CLOCK

CA3 Set 35	CA3 Set 36	CA3 Set 37	CA3 Set 38	CA3 Set 39	CA3 Set 40
HOLD	FT	DOOR	ENG	FINAL	IFR
SHORT	CLRNC	OPEN	FIRE	BASE	VFR
RWY	ALT	CARGO	START	DOWNWIND	ARPT
LINE	DSND	CLOSED	APU	RWY	WX
TXWY	#	SLIDE	PWR	VISUAL	SCATTERED
TWR	READ	CABIN	OIL	SIGHT	VISIBILITY
LINES	DSCNT	THE	MAINT	TFC	ME
TAXI	CTLR	ATTENDANT	SHUT	APCH	REPORT
STOPPED	ARR	DOORS	SHUTDOWN	FOLLOW	CLOUDS
CROSSED	CTR	OPENED	COWLING	KING	I
HOLDING	CROSS	GALLEY	OVERHEAT	TWR	TCA
TAXIING	BACK	COCKPIT	BLEED	PATTERN	IMPERIAL
PAST	MAINTAIN	FLT	CHKLIST	FOR	SMT
GND	CIVET	SECURED	THE	ARPT	AWOS
Q	CLRED	PERSONNEL	ENGS	TURN	FLT
ACROSS	AT	AFT	LEVER	CESSNA	PLAN
MARKINGS	READBACK	SVC	RUNNING	AIR	DEP
TAXIED	WE	PALLETS	CAUTION	WE	SJD
SIGN	ZDV	PAX	LOOP	FIELD	LAYER
H	SINCA	CREW	EMER	PIPER	TEMPORARY

CA3 Set 41	CA3 Set 42	CA3 Set 43	CA3 Set 44	CA3 Set 45	CA3 Set 46
RADIAL	RAMP	HELI	FLAPS	RWY	ALT
DEG	GATE	HELIS	FLAP	TAXI	FT
INTERCEPT	PARKING	EMS	TKOF	GND	ALERTER
VOR	WINGTIP	PLT	HORN	SHORT	ASSIGNED
DEP	PARKED	ROTOR	WARNING	INSTRUCTIONS	CLB
OUTBOUND	DAMAGE	TCA	TRIM	TXWY	ALERT
HDG	WING	CANYON	CHKLIST	HOLD	FLYING
COURSE	AREA	MAUI	DEGS	TAXIING	LEVELOFF
TURN	BLAST	PATTERN	HANDLE	CROSS	SET
DME	TAXI	ZZZ	CONFIGN	RAMP	CLBING
ARR	L	POLICE	SETTING	ACTIVE	#
SID	STRUCK	PROPERTY	PWR	TWR	THROUGH
JOIN	LINE	HOVER	STABILIZER	ACROSS	PNF
INTERCEPTED	ACFT	STATUE	SPD	CLRNC	ATTN
DCA	THE	HELIPAD	RETRACTED	ONTO	DEV
MZB	LIGHT	LNDG	THRUST	TAXIED	HAND
ANPU	BUILDING	HOUSE	THROTTLES	CROSSED	LEVEL
SBJ	SHUT	MIDFIELD	VR	CTL	TURB
DEGS	HIT	DIRECTLY	ADVANCED	STOPBAR	CLBED
#	TERMINAL	ORBIT	THE	INSTRUCTED	DSNDED

CA3 Set 47	CA3 Set 48	CA3 Set 49	CA3 Set 50
RPTR	RTE	RWY	CIRCUIT
CALLBACK	DIRECT	LNDG	BREAKER
REVEALED	PLAN	LAND	BREAKERS
CONVERSATION	ROUTING	GAR	MAINT
FOLLOWING	FILED	TWR	POPPED
HE	FMS	APCH	TECHNICIAN
STATES	FMC	FINAL	PULLED
HAS	PDC	AGL	PROC
INFO	TRANSITION	VISUAL	RESET
THEY	FLT	TAILWIND	CONTRACT
STATED	CLRNC	WINDS	CHKLIST
THAT	GRANN	LINED	AUTOSLAT
IS	PROGRAMMED	LANDED	HORN
HIS	DISCONTINUITY	MOONEY	PANEL
FEELS	LOADED	BUCKLEY	THRUST
WITH	CRI	SIDESTEP	PRIOR
WAKE	COURSE	KTS	MECH
FLC	WYLYY	BASE	COLLARED
ARSA	RTING	AROUND	RWY
THE	MLF	STABILIZED	#

CA3 Set 51	CA3 Set 52	CA3 Set 53	CA3 Set 54
SFO	TKOF	ACR	FUEL
VISUAL	TWR	X	LBS
APCH	RWY	Y	TANK
RWY	POS	#	WT
BRIDGE	ROLL	SECTOR	TANKS
US	HOLD	Z	LOAD
BRIJJ	TAXIED	TFC	GAUGES
SIGHT	FOR	MTR	XFEED
BAY	CLRED	ISSUED	QUANTITY
MENLO	ONTO	TCASII	BAL
TOE	ABORT	HE	BURN
TIPTOE	CLRNC	RA	RELEASE
FINAL	READY	DSNDING	IMBAL
ILS	END	CLIMB	PUMP
HE	TAXI	HIM	PUMPS
PARALLEL	INTO	OBSERVED	FUELED
LAND	ABORTED	RADAR	ALTERNATE
CTRLINE	WE	HIS	FLT
WE	CHKLIST	PLT	DISPATCH
CENTERLINE	ITEMS	SIMILAR	GAUGE

Top 20 Keywords for each of the 54 clusters from the 4th sample of CA documents:

CA4 Set 1	CA4 Set 2	CA4 Set 3	CA4 Set 4	CA4 Set 5	CA4 Set 6
AUTOPLT	RVR	FINAL	ALTIMETER	CLASS	I
ALT	VISIBILITY	BASE	SETTING	AIRSPACE	HE
CAPTURE	CAT	DOWNWIND	FT	B	MY
FT	MINIMUMS	TWR	#	VNY	IT
MODE	WX	PATTERN	ALTIMETERS	VFR	HIM
ENGAGED	APCH	RWY	RESET	AREA	FLYING
ARMED	FOG	LAND	SET	ZZZ	IS
SELECTED	II	SIGHT	ALT	HELI	DO
DSCNT	RPTED	HELI	INCHES	SHORELINE	ARE
LEVELOFF	MI	TFC	ATIS	BUR	AIRPLANE
DISENGAGED	LIGHTS	GAR	LOW	CFI	ME
DISCONNECTED	ATIS	I	LEVELED	MSL	DORNIER
LEVEL	ILS	R	DSCNT	C	HIS
SELECT	FORECAST	FOLLOW	LEVEL	MUGU	GET
PITCH	RVV	APCH	SETTINGS	JUMPERS	DOING
RATE	PREVAILING	TURNING	QNH	FLOOR	FFDO
VERT	RWY	VISUAL	PASSING	DELIVERY	MOST
MANUALLY	III	CRJ	CHKLIST	ABCDE	THIS
DEV	SHOOT	MI	AT	IFR	YOU
SPD	SPECS	L	ERROR	DEPARTING	KNOW

CA4 Set 7	CA4 Set 8	CA4 Set 9	CA4 Set 10	CA4 Set 11	CA4 Set 12
DEGS	TXWY	TAXIWAY	MAINT	MEL	TKOF
HDG	TAXI	TAXI	LOGBOOK	MAINT	RWY
TURN	RWY	OUTER	WRITE	INOP	TWR
DEP	GND	GND	LOG	RELEASE	POS
HEADING	RAMP	RWY	ITEM	DISPATCH	ROLL
R	ONTO	INNER	MECH	LOGBOOK	CLRED
DEG	TXWYS	PARALLEL	UP	FLT	HOLD
#	TAXIING	ONTO	OPEN	TRU	FOR
L	INSTRUCTIONS	TANGO	SYS	BREAKER	CLRNC
CTLR	TAXIED	SIGN	LEAK	CIRCUIT	READY
TURNING	GATE	TAXIWAYS	ENTRY	DEFERRED	TAXI
TKOF	B	SHORT	INSPECTION	MEL'S	ABORT
COMPASS	J	AREA	DISCREPANCY	ITEM	ACR
TURNED	K	ECHO	ZZZ	NUMBER	DATA
CLB	D	PROBLEM	PREFLT	SYS	ONTO
MM	VIA	ON	WRITTEN	PLACARD	TAXIED
WHITESTONE	DIAGRAM	EXIT	UPS	CREW	RWYS
BUG	CTL	E	CHKS	DEFERRAL	TAXIING
DME	P	SIGNS	ARRIVED	ANTI	Y
GYRO	EXIT	RAMP	REPAIR	LIST	ABORTED

CA4 Set 13	CA4 Set 14	CA4 Set 15	CA4 Set 16	CA4 Set 17
LOC	HRS	RWY	CLOSED	RADIAL
APCH	DUTY	HIGHSPD	NOTAM	DEG
ILS	DAY	DFW	NOTAMS	INTERCEPT
INTERCEPT	REST	TWR	RWY	VOR
GS	TRIP	APCH	CLOSURE	HDG
COURSE	SCHEDULING	ILS	FSS	OUTBOUND
INTERCEPTED	HR	VISUAL	WRITER	DEGS
GLIDE	CREW	MCO	ARPT	COURSE
SLOPE	DAYS	WE	ATIS	#
VECTORED	SCHEDULED	LNDG	CLOSURES	DME
ESTABLISHED	FATIGUE	SIDE	CONES	NAV
FALSE	SCHEDULE	EM	LNSAY	DEP
HDG	LEGAL	SHORT	CONSTRUCTION	INBOUND
RWY	SLEEP	TPA	CTAF	INTERCEPTED
AUTOPLT	PERIOD	LINED	WX	TURN
WERE	TIME	YANKEE	MENTION	TRANSITION
VECTOR	COMPANY	EXPECT	TXWYS	DCA
VECTORS	NIGHT	CLB	LIGHTS	AIRWAY
ALIVE	FLT	ALT	UNICOM	TUNED
CAPTURE	ASSIGNMENT		ORL	PALEO

CA4 Set 18	CA4 Set 19	CA4 Set 20	CA4 Set 21	CA4 Set 22
HOLD	FUEL	ALT	DOOR	MAIN
SHORT	LBS	FT	ATTENDANT	BRAKES
RWY	TANK	ASSIGNED	FLT	BRAKING
LINE	WT	ALERTER	PAX	APPLIED
TAXI	GAUGES	ALERT	SHE	TIRES
TXWY	BAL	LEVELOFF	HER	RWY
LINES	TANKS	FLYING	COCKPIT	STEERING
TWR	PUMPS	CLBING	CABIN	TIRE
CROSS	RELEASE	#	SEAT	BRAKE
GND	DISPATCH	ATC	ATTENDANTS	NOSE
ACROSS	ALTERNATE	ATTN	OPEN	GEAR
CROSSED	LOAD	CAPT	FORWARD	NOSEWHEEL
TAXIED	QUANTITY	DEV	AFT	WHEEL
STOPPED	FUELER	THROUGH	OPENED	THE
INSTRUCTIONS	BURN	LEVEL	GALLEY	ACFT
HOLDING	MINIMUM	DISTR	JUMP	REVERSE
TAXIING	IMBAL	NOTICED	CARGO	DAMAGE
MARKINGS	BOARD	SET	BAG	RUDDER
STOP	MINS	AURAL	ALCOHOL	TOUCHDOWN
TXWYS	SLIP	DSNDED	AGENT	SKID

CA4 Set 23	CA4 Set 24	CA4 Set 25	CA4 Set 26	CA4 Set 27	CA4 Set 28
SMA	GEAR	SMT	FMC	APCH	PUSH
Y	PIN	TFC	LNAV	VISUAL	PUSHBACK
X	PINS	O'CLOCK	FMS	ARPT	GND
TFC	NOSE	EVASIVE	FIX	FINAL	GATE
PLT	PREFLT	GLIDER	DEP	SIGHT	START
RIGHT	INSTALLED	PASSED	NAV	RWY	BRAKE
TCA	REMOVED	VFR	RNAV	WE	BRAKES
LTT	DOWN	SAW	DIRECT	FIELD	TUG
SMT	LNDG	AVOID	COURSE	TERRAIN	PARKING
SBP	MAIN	COLLISION	DATA	GPWS	RAMP
LGB	HANDLE	NEAR	ARR	LNDG	PUSHED
AN	MAINT	MI	SPL	ILS	ENG
O'CLOCK	RETRACT	ACFT	RTE	CONFIGURED	SALUTE
EVASIVE	INSPECTION	MISS	PAGE	VECTORED	CREW
MISS	FLAG	ACTION	LAS	FAF	DRIVER
LEFT	INDICATION	SPOTTED	PROGRAMMED	FOR	TOW
SIGHT	STALL	RIGHT	SKEBR	RIVER	RELEASED
HE	CHKLIST	HE	LOADED	HIGH	PERSONNEL
AKN	GREEN	WING	ROSUN	BASE	MECH
NEAR	HORN	LEFT	RAW	WERE	PUSHING

CA4 Set 29	CA4 Set 30	CA4 Set 31	CA4 Set 32	CA4 Set 33	CA4 Set 34
CIVET	TKOF	DME	VFR	RESTR	FT
ARR	FLAPS	VOR	IFR	DSCNT	CTLR
LAX	FLAP	INTXN	CONDITIONS	XING	CLRNC
PROFILE	CHKLIST	ARR	WX	CROSS	DSND
ARNES	WARNING	FIX	ARPT	MAKE	ALT
MITTS	HORN	#	FSS	VNAV	READ
FMS	THROTTLES	XING	PLAN	AT	DSCNT
SOCAL	PWR	APCH	AWOS	FT	READBACK
DSCNT	CONFIGN	LDA	FLT	FMS	BACK
CLRED	TRIM	WHIGG	CANCEL	ARR	ATC
BREMR	SLATS	AT	CLOUDS	INTXN	#
VIA	HANDLE	PLATE	CLOUD	FMC	DISCRETION
VNAV	ABORT	DSCNT	VISIBILITY	RESTRS	SAID
KTS	SLAT	MISREAD	LAYER	OLYMPIA	CTR
PDZ	ABORTED	SSR	CANCELLED	RATE	HEARD
DENAY	LEVERS	SOBER	BMG	MI	CALL
ILS	ADVANCED	RESTRICTION	ARLINGTON	ATC	WHAT
FT	CHKLISTS	FT	I	NM	MAINTAIN
DSND	THE	ARC	TCA	PROGRAMMED	TO
FMC	THRUST	CHART	SCATTERED	PROFILE	PLT'S

CA4 Set 35	CA4 Set 36	CA4 Set 37	CA4 Set 38	CA4 Set 39	CA4 Set 40
ACR	FAA	DEP	ENG	RTE	TOUCHDOWN
X	INSPECTOR	FT	START	PDC	TAIL
Y	JUMP	CLB	FIRE	PLAN	TAILSKID
XYZ	SEAT	ALT	APU	ROUTING	LNDG
SECTOR	CHK	CLBING	PWR	FILED	DAMAGE
#	MY	TFC	SHUT	FLT	WINDS
TCASII	RATING	MAINTAIN	SHUTDOWN	DEP	ROTATION
ISSUED	LETTER	#	OIL	CLRNC	STRIKE
MLT	CERTIFICATE	ASSIGNED	LEVER	DIRECT	NORMAL
TFC	REVISION	TCAS	SWITCH	OUR	KTS
RADAR	PLT	THROUGH	MAINT	ACARS	SINK
OBSERVED	CHIEF	INITIAL	EGT	CODE	WIND
SEPARATION	COMPANY	METERS	NORMAL	LOADED	FLARE
FT	I	PNF	IGNITION	FMS	TAILWIND
RA	MEDICAL	CLBED	LIGHT	WAYPOINT	SCRAPED
C	TRAINING	SEPARATION	SWITCHES	OAL	SKID
VFR	MANUAL	O'CLOCK	CHKLIST	NEW	HARD
CLBING	TRNING	COPLT	FUEL	SQUAWK	PWR
PLT	QUALIFIED	CTLR	BOOST	STORED	STABILIZED
WORKING	RECORDS	HDG	THE	ERROR	COMPONENT

CA4 Set 41	CA4 Set 42	CA4 Set 43	CA4 Set 44	CA4 Set 45
RPTR	TWR	MARSHALLER	DEP	LAX
CALLBACK	APCH	RAMP	SID	RWY
REVEALED	FREQ	WINGTIP	TURN	LOC
CONVERSATION	LNDG	WING	HDG	COMPLEX
FOLLOWING	LAND	PARKING	SJC	FINAL
INFO	CONTACT	DAMAGE	PROC	VIS
HAS	SWITCHED	PARKED	DEGS	BASE
STATES	SWITCH	JETWAY	DEG	SIGHT
HE	RADIO	ACFT	DME	APCH
THEY	LANDED	GATE	HEADING	COMMUTER
SAILPLANE	VISUAL	SIGNAL	VOR	FOR
SAYS	CLRNC	STOP	BRIEFED	SMO
WITH	CTL	THE	RESTRICTION	OVERSHOT
FAA	MARKER	TAXIING	#	VISUAL
FEELS	CLRED	SPOT	BRIEFING	RWYS
HIS	COM	L	LOUPE	HAZE
ANALYST	FREQS	SHUT	DEPS	ROMEN
REPORTER	WITHOUT	PROP	TKOF	OVERSHOOT
IS	WE	PERSONNEL	TEXT	SUN
THINKS	GND	STRUCK	MEAD	TWR

CA4 Set 46	CA4 Set 47	CA4 Set 48	CA4 Set 49	CA4 Set 50
TURB	TCASII	CENTER	SNOW	SFO
MODERATE	RA	O	DEICE	VISUAL
ENCOUNTERED	TFC	F	DEICING	BRIDGE
AIRSPD	FT	CLIMB	ICE	APCH
ICING	CLB	PROBLEM	WINGS	RWY
ALT	O'CLOCK	MFR	FROST	SIGHT
SEVERE	TA	CLRNC	DEICED	BAY
WX	TARGET	DEVIATION	BRAKING	QUIET
CLB	#	SITUATION	NIL	TWR
STORM	AT	TPA	FLUID	MLG
TSTMS	CONFLICT	REPORTED	POOR	PARALLEL
ICE	ATC	ROUTE	BLOWING	MATEO
RADAR	CLBING	PIE	FREEZING	BRIJJ
LIGHTNING	ADVISED	DI	SURFACES	SAMUL
TSTM	VERT	MADRID	TKOF	CONTACT
FT	BELOW	TO	COLD	MILL
WAKE	FPM	ABC	STG	CLOUDS
NIMBUS	COMMAND	KINGSTON	POLICY	SPD
CUMULO	DSND	DPK	SNOWING	APCHS
CLOUD	FOLLOWED	DCT	PLOWED	SPACED

CA4 Set 51	CA4 Set 52	CA4 Set 53	CA4 Set 54
#	DIRECT	CABIN	KTS
CLB	VOR	OXYGEN	SPD
X	COURSE	PACKS	AIRSPD
ALT	NAV	BLEEDS	SLOW
CTR	HDG	PRESSURIZATION	KIAS
DSND	CTR	MASKS	SLOWED
CLRNC	OMEGA	PRESSURE	MACH
LEVEL	PNT	HORN	FT
ACR	NEEDLES	BOTTLES	SLOWING
ATC	HVQ	EMER	APCH
AT	ADF	PORTABLE	DSCNT
BACK	INTXN	DONNED	MENLO
LINK	BOY	MASK	#
REQUESTED	INS	CHKLIST	MAINTAIN
ZOA	WERE	AUTO	REDUCTION
FANS	WE	DOCTOR	POVOC
ZID	BPT	PRESSURIZED	AT
MINS	BEARING	BLEED	DEGAN
DISCRETION	WAYPOINT	SWITCHES	CTLR
CLBING	TRACK	PAX	RESTRS

Top 20 Keywords for each of the 54 clusters from the 5th sample of CA documents:

CA5 Set 1	CA5 Set 2	CA5 Set 3	CA5 Set 4	CA5 Set 5
TKOF	DME	CTAF	ACR	HOLD
CHKLIST	VOR	UNICOM	X	SHORT
FLAPS	SJC	PATTERN	Y	RWY
HORN	RESTRICTION	FSS	TCASII	LINE
ROLL	XING	DOWNWIND	TFC	TWR
WINDOW	ARC	ANNOUNCED	#	TAXI
ABORTED	FIX	RWY	CLBING	TXWY
PWR	VORTAC	RADIO	CGA	CROSSED
WARNING	CROSS	IFR	RA	POS
NUMBERS	LOUPE	STUDENT	SEPARATION	STOP
WT	#	COMMUTER	HIM	LINES
ABORT	AT	INTENTIONS	WDB	STOPPED
NORMAL	DLF	I	FOREIGN	PAST
REJECTED	APCH	ON	FT	TAXIING
HANDLE	RESTRICTIONS	HIM	HE	CLRED
THROTTLES	RMI	BASE	Z	ACROSS
ROTATION	BAY	PLT	PLT	GND
COMPLETED	READOUT	BEECH	SECTOR	CROSS
ADVANCED	FNT	UNCTLED	ISSUED	RWYS
FLAP	MSO	FINAL	DSNDING	TAXIED

CA5 Set 6	CA5 Set 7	CA5 Set 8	CA5 Set 9	CA5 Set 10	CA5 Set 11
RPTR	RVR	TCASII	SFO	FILED	RADIAL
CALLBACK	VISIBILITY	RA	BRIDGE	RTE	DEG
REVEALED	MINIMUMS	TFC	VISUAL	PDC	INTERCEPT
CONVERSATION	WX	CLB	APCH	PLAN	VOR
FOLLOWING	APCH	FT	WDB	CLRNC	COURSE
HE	MI	TA	BAY	ROUTING	HDG
INFO	FOG	O'CLOCK	MATEO	DEP	DEGS
STATED	ILS	TARGET	VIS	FLT	NAV
MGM	RPTED	#	MLG	ACARS	ARR
WITH	CEILING	CLBING	SEP	DIRECT	OUTBOUND
STATES	ATIS	FPM	RWYS	CODE	#
HAS	IFR	AT	APCHS	OUR	DIRECT
TCA	OVCST	ATC	SIGHT	SQUAWK	FMS
FAA	CONDITIONS	BELOW	PARALLEL	OCEANIC	INTERCEPTED
BELIEVES	CURRENT	CLBED	QUIET	XPONDER	INSTEAD
FEELS	MISSED	DSND	MAINTAIN	DELIVERY	FWA
BETA	LEGAL	CONFLICT	Y	PRE	TURN
DUPAGE	BELOW	VISUALLY	RWY	VIA	AIRWAY
THE	DECISION	ADVISED	FOSTER	RTES	HEADING
HIS	ROLLOUT	INTRUDER	PASS	RECEIVED	HFD

CA5 Set 12	CA5 Set 13	CA5 Set 14	CA5 Set 15	CA5 Set 16	CA5 Set 17
PUSHBACK	TKOF	ALT	RWY	TWR	APCH
BRAKE	TWR	FT	TAXI	LNDG	VISUAL
TUG	RWY	ALERTER	GND	LAND	ARPT
GND	POS	ASSIGNED	SHORT	APCH	RWY
BRAKES	HOLD	ALERT	HOLD	FREQ	SIGHT
PARKING	CLRED	MSL	ACTIVE	CONTACT	FINAL
PUSH	ROLL	LEVELOFF	CROSS	RWY	FIELD
DRIVER	READY	SET	TAXIWAY	CLRED	FOR
TOW	CLRNC	WARNING	INSTRUCTIONS	FINAL	BASE
START	FOR	CAPT	TAXIING	SWITCH	DOWNWIND
CREW	TAXI	LEVEL	CROSSED	LANDED	CLRED
GATE	ABORT	LEVELING	ACROSS	CLRNC	LINED
SET	WAIT	#	CLRNC	SWITCHED	WE
PUSHED	HEARD	WINDOW	TWR	CONTACTED	VECTORED
BAR	TAXIING	DSNDED	ONTO	CTL	TERRAIN
MOVEMENT	ONTO	DEV	XING	ON	TWR
ENG	INTO	ATTN	RWYS	OM	LAND
DISCONNECT	AIRBORNE	PNF	BRAVO	WITHOUT	GAR
THE	LUBBOCK	CLBING	HEARD	VISUAL	CTRLINE
CHOCKS	SHORT	THROUGH	CTL	CUB	ILS

CA5 Set 18	CA5 Set 19	CA5 Set 20	CA5 Set 21	CA5 Set 22
KTS	HRS	FMC	ARR	ICE
SPD	DUTY	FMS	FT	ICING
AIRSPD	DAY	DIRECT	DSCNT	ANTI
SLOWED	REST	PAGE	CROSS	SNOW
SLOW	HR	RTE	INTXN	WING
#	SLEEP	DEP	PROFILE	WINGS
KIAS	FATIGUE	FIX	STAR	DEICING
DSCNT	NIGHT	SKEBR	DSND	ENG
SLOWING	TRIP	ARR	OLM	TEMP
RESTRICTION	SCHEDULED	LEGS	CIVET	STALL
REDUCTION	HOTEL	LNAV	TONTO	ACD
FT	DAYS	CDU	AT	ELEVATOR
ATC	TIRED	RNAV	RESTRS	AIRSPD
CIVET	SCHEDULING	DISPLAY	XING	VANES
KT	LEG	ENTERED	ALT	COMPRESSOR
FMC	CREW	NAV	#	HEAT
MSL	SCHEDULE	THE	CLRED	DEICED
ALT	TRIPS	PROGRAMMING	CHART	ADHERING
HIGH	PHF	FIXES	YYZ	CONDITIONS
ACCELERATED	AM	SPANE	KORRY	MODERATE

CA5 Set 23	CA5 Set 24	CA5 Set 25	CA5 Set 26	CA5 Set 27
WIND	AUTOPLT	LOC	DSCNT	MAINT
KTS	ALT	APCH	ALT	MEL
BRAKING	CAPTURE	ILS	#	LOGBOOK
PWR	ENGAGED	GS	HEATT	ITEM
TOUCHDOWN	MODE	INTERCEPT	CROSS	MECH
FLARE	DISENGAGED	ESTABLISHED	BOG	INOP
LNDG	LEVEL	COURSE	ATC	WRITE
DAMAGE	DISCONNECTED	VECTORED	DSND	LOG
TOUCHED	FT	THE	CTR	DISPATCH
ACFT	PITCH	BRIEFED	DISCRETION	DEFERRED
WINDS	TRIM	MARKER	PMD	RELEASE
RWY	WHEEL	FREQ	DESCENT	APU
RUDDER	AUTOTHROTTLES	FT	LANDR	CREW
SINK	LEVELOFF	RWY	WINDOW	DISPATCHER
APPLIED	SELECTED	TUNED	AT	INSPECTION
SKID	ARMED	CAPTURED	HEC	PREFLT
MAIN	LNAV	INTERCEPTED	MCP	FLT
TAIL	VNAV	DME	LEVELED	PLACARD
NORMAL	ALTDEV	FOR	CLRNC	THE
CTRLINE	KNOB	PLATE	AUTOPLT	REQUIRED

CA5 Set 28	CA5 Set 29	CA5 Set 30	CA5 Set 31	CA5 Set 32	CA5 Set 33
DEP	FUEL	CLB	SMA	VFR	RAMP
SID	TANK	#	Y	TFC	PARKED
PROC	LBS	ALT	X	EVASIVE	GATE
DME	XFEED	REQUESTED	TFC	COLLISION	WINGTIP
TURN	PUMP	MACH	O'CLOCK	O'CLOCK	WING
HDG	CTR	CENTER	HSV	PASSED	TRUCK
NOISE	IMBAL	CRUISING	EVASIVE	GLIDER	TAXI
ABATEMENT	TANKS	CLRNC	TWR	NEAR	AREA
PDC	BOOST	WE	SPOTTED	RADAR	MARSHALLER
BRIEFED	PUMPS	AT	LEFT	CESSNA	ACFT
TKOF	QUANTITY	DJB	SAW	IFR	SIGNAL
CLRNC	VALVE	FGT	SMA'S	GLIDERS	STRUCK
DELIVERY	GAUGE	CRUISE	LCL	MISS	WALKER
DEGS	IMBALANCE	MINS	NEAR	TCA	PARKING
HEADING	ENGINE	LEVEL	RIGHT	ACTION	STOP
FILED	AUX	ATC	PATROL	TARGET	THE
CLB	LB	YYY	HE	AVOID	DAMAGE
OBSTACLE	GALLONS	ASKED	ADS	ACFT	TAXIED
MEAD	CHKLIST	GANDER	ACTION	XPONDER	MARSHALER
BRIEFING	DEACTIVATED	X	MSL	TARGETS	POLE

CA5 Set 34	CA5 Set 35	CA5 Set 36	CA5 Set 37	CA5 Set 38	CA5 Set 39
GEAR	FUEL	LIGHTS	TXWY	HDG	CLASS
PIN	WT	RWY	TAXI	DEGS	AIRSPACE
PINS	LBS	TXWY	GND	TURN	B
NOSE	LOAD	THRESHOLD	RAMP	DEP	ME
MAINT	RELEASE	DISPLACED	RWY	DEG	MY
RETRACT	DISPATCH	END	ONTO	#	FLOOR
LNDG	FLT	EDGE	TXWYS	CTLR	I
PREFLT	BOARD	CLOSED	K	R	VFR
INSTALLED	BAL	LIGHTING	VIA	TURNED	MYF
FLAGS	BURN	TKOF	D	GIVEN	BURBANK
LOCKED	PAPERWORK	THE	GATE	CLB	WHITEMAN
EMER	GROSS	VISIBLE	EXIT	ASSIGNED	D
DOWN	MANIFEST	BRIGHT	G	COMPASS	C
STOWED	MINIMUM	MARKINGS	TAXIING	US	STUDENT
RED	POUNDS	NIGHT	TAXIED	L	OUTSIDE
FLAG	PAX	WHITE	L	TURNING	SOLO
HANDLE	DEST	BLUE	SIGN	GAVE	INSTRUCTOR
HANGAR	PAYLOAD	TAXI	SIGNS	TOLD	XCOUNTRY
DOOR	MINS	CTRLINE	E	TKOF	GAGGLE
HYD	SLIP	TAXIED	H	BLD	MIRAMAR

CA5 Set 40	CA5 Set 41	CA5 Set 42	CA5 Set 43	CA5 Set 44	CA5 Set 45
VISUAL	SMT	FREQ	ALTIMETER	ENG	COURSE
TFC	TFC	RADIO	SETTING	START	NAV
SEPARATION	MDT	COM	ALTIMETERS	SHUT	OMEGA
SIGHT	O'CLOCK	CONTACT	RESET	PWR	INS
TCASII	MFR	DULLES	#	OIL	VOR
O'CLOCK	TWIN	TAHITI	FT	ENGS	TRACK
US	HE	CTR	INCHES	APU	COORDINATES
MAINTAIN	TCA	VHF	ATIS	CHKLIST	DIRECT
RA	Y	IAD	SET	MAINT	SALEM
FOLLOW	DEBRIS	ME	ALT	IGNITION	TOKYO
BEHIND	VFR	I	SETTINGS	STRAP	ETA
Y	O	XMISSIONS	QFE	EGT	PLAN
FT	LOVE	SELCAL	QNH	THE	POS
CLB	SEPARATION	CHANGE	LOW	SWITCH	AIRWAY
APCH	SIGHT	AUDIO	MB	PROP	ERROR
WAKE	BHM	HF	RESETTING	RUNNING	LONGITUDE
LTT	RIGHT	MDW	TRANSITION	OVERHEAT	ONS
HE	ME	PAIN	LEVEL	XBLEED	USING
HELI	BIRD	ATIS	MILLIBARS	AIR	PLOTTING
PASS	PASSED	PCT	HG	MECH	CHART

CA5 Set 46	CA5 Set 47	CA5 Set 48	CA5 Set 49	CA5 Set 50
FAA	FLAPS	PAX	TXWY	RESTR
INSPECTOR	FLAP	ATTENDANT	RWY	XING
CERTIFICATE	GEAR	DOOR	SHORT	DSCNT
MY	LNDG	CABIN	TAXI	CROSS
FLT	SPD	FLT	HOLD	AT
TEMPORARY	KTS	COCKPIT	GND	MAKE
CHIEF	CHKLIST	SEAT	P	FT
THAT	EXTENDED	ATTENDANTS	ONTO	INTXN
MEDICAL	AGL	HER	INSTRUCTIONS	FMS
AKL	DEGS	OXYGEN	TAXIING	RATE
HIS	SPOILERS	SHE	B	FMC
I	LOWERED	JUMP	CROSS	MI
MGR	WARNING	GALLEY	TWR	MEET
LICENSE	DOWN	BOARDING	EXIT	HIGH
TRIP	RETRACTED	CAPT	C	FIX
MASK	CONFIGN	AGENT	CROSSED	PROFILE
CERTIFICATES	LIGHT	EMER	E	ATC
CHK	BRAKES	SEATED	Q	VNAV
PAD	SELECTED	PURSER	TXWYS	ARR
MKG	HANDLE	MASKS	ACROSS	RESTRS

CA5 Set 51	CA5 Set 52	CA5 Set 53	CA5 Set 54
LAX	HOLDING	FT	CENTER
COMPLEX	PATTERN	READ	LEFT
APCH	HOLD	CLRNC	DEVIATION
ILS	PUBLISHED	READBACK	JLN
VISUAL	EFC	CALL	CLIMB
RWY	RADIAL	CTLR	DEVIATING
SMO	FMS	BACK	F
HAZE	FIX	SIMILAR	UKIAH
SIGHT	TURNS	HEARD	DESCENDED
FINAL	OUTBOUND	SIGN	CONTACT
LOC	LEGS	ALT	JUMPERS
RWYS	INBOUND	#	HOUSTON
S	TURN	SIGNS	MKK
BASE	VOR	ATC	REPORTER
N	ENTRY	DSND	BRADLEY
SUN	RLG	CTR	DSNT
STADIUM	ENTER	MAINTAIN	ATLANTA
SOCAL	SBV	ZBW	IFR
TFC	FLO	SOUNDING	DISCRETION
SADDE	SSR	FREQ	SCOTT

Top 20 Keywords for each of the 54 clusters from the 6th sample of CA documents:

CA6 Set 1	CA6 Set 2	CA6 Set 3	CA6 Set 4	CA6 Set 5
LAX	MAINT	DOOR	HOLD	CABIN
APCH	MEL	ATTENDANT	SHORT	PRESSURIZATION
RWY	LOGBOOK	FLT	LINE	PACKS
LOC	WRITE	PAX	RWY	APU
CIVET	ITEM	ATTENDANTS	TAXI	SWITCH
ARNES	LOG	CABIN	TXWY	PACK
VISUAL	INOP	SHE	TWR	HYD
VIS	RELEASE	COCKPIT	LINES	MASKS
MITTS	ITEMS	HER	TAXIING	BLEED
SOCAL	ENTRY	SEATED	TAXIED	EMER
ILS	DISPATCH	OPEN	STOP	SYS
DENAY	SYS	SEAT	ACROSS	PRESSURE
ARR	SIGNED	LAVATORY	PAST	AIR
FUELR	PREFLT	AFT	PAINTED	FLOW
COMPLEX	DEFERRED	SMOKING	CROSSED	ADG
STADIUM	MECH	ANNOUNCEMENT	STOPPED	HEAT
GS	MISSING	SUITCASE	GND	SWITCHES
PDZ	PACK	PREPARE	MARKINGS	CHKLIST
SMO	OPEN	PUSHBACK	OF	QRH
WE	DOCUMENT	SEATBELT	SIGNS	PANEL

CA6 Set 7	CA6 Set 8	CA6 Set 9	CA6 Set 10	CA6 Set 11	CA6 Set 12
LIGHTS	TCASII	HDG	GEAR	HRS	TFC
TXWY	RA	DEGS	PIN	REST	O'CLOCK
RWY	TFC	TURN	PINS	DUTY	VFR
EDGE	CLB	DEG	NOSE	DAY	TWIN
LIGHT	FT	DEP	MAIN	TRIP	EVASIVE
LIGHTING	TA	CTLR	LNDG	SLEEP	SMT
PAVEMENT	O'CLOCK	#	FERRY	CREW	ACFT
RAMP	CLBING	ASSIGNED	INSPECTION	SCHEDULING	NEAR
CTRLINE	ATC	TURNED	MAINT	FATIGUE	PASSED
MAIN	#	L	DOWN	HR	MISS
DAMAGE	COMMAND	BACK	PERMIT	SCHEDULED	SPOTTED
RAIN	TARGET	COMPASS	SKID	NIGHT	AVOID
TAXI	CLBED	HOOVER	PROP	LEGAL	RIGHT
END	CONFLICT	R	LOCKED	HOTEL	MI
PROP	VISUALLY	GAVE	COVERS	DAYS	AT
R	AT	TO	ATTACHED	SCHEDULE	COLLISION
THE	FPM	READ	INSTALLED	TIME	TARGET
GRASS	ADVISED	HEADING	REMOVED	BLOCK	CLOSE
TAXIWAY	VERT	US	RETRACTION	PERIOD	F
SIDE	DSND	SAID	GREEN	REDUCED	SAW

CA6 Set 13	CA6 Set 14	CA6 Set 15	CA6 Set 16	CA6 Set 17	CA6 Set 18
ENG	RWY	TWR	RADIAL	FLAPS	SFO
START	SHORT	APCH	DEG	TKOF	APCH
THE	TAXI	LNDG	INTERCEPT	FLAP	VISUAL
IGNITION	HOLD	LAND	VOR	HORN	BRIDGE
CHKLIST	TXWY	CONTACT	HDG	CHKLIST	VIS
ENGS	GND	FREQ	DEP	WARNING	TOE
FIRE	INSTRUCTIONS	RWY	OUTBOUND	TRIM	ARCHI
PWR	CROSS	SWITCH	DEGS	STABILIZER	TIPP
LEVER	TWR	LANDED	SID	SETTING	RWY
OIL	ACROSS	CLRED	COURSE	HANDLE	QUIET
MAINT	CROSSED	CTL	INSTEAD	THROTTLES	SEP
SHUT	TAXIING	VISUAL	SBJ	PWR	BAY
EGT	ACTIVE	FINAL	FMS	DETENT	TRDOW
RESTART	RWYS	OM	NAV	CONFIGN	BRIJJ
TEMP	ONTO	BUSY	TRANSITION	SHAKER	GAROW
TKOF	CLRNC	GAR	SRP	ADVANCED	ALTITUDE
FUEL	CTL	WE	CRI	SOUNDED	FINAL
FORWARD	XING	ON	SLC	COMPLETED	RWYS
PLUGS	TAXIWAY	CLRING	DPK	ITEM	APCHS
SHUTDOWN	STOP	CLRNC	INBOUND	ABORTED	FMGC

CA6 Set 19	CA6 Set 20	CA6 Set 21	CA6 Set 22	CA6 Set 23
HOLDING	ACR	RPTR	TCASII	DEP
PATTERN	X	CALLBACK	SEPARATION	SID
HOLD	Y	REVEALED	VISUAL	TURN
FIX	TCASII	CONVERSATION	RA	PROC
PUBLISHED	RA	FOLLOWING	TFC	HDG
TURNS	FT	HAS	SIGHT	DME
INBOUND	TFC	INFO	FT	TKOF
VOR	#	STATES	CLB	SJC
RIC	CPR	STATED	O'CLOCK	DEG
INSTRUCTIONS	CLBING	HE	CESSNA	ABATEMENT
DME	MLT	HIS	TARGET	BRIEFED
EFC	SECTOR	THE	MAINTAIN	NOISE
ENTRY	SEPARATION	INCIDENT	US	DEGS
ENTERING	HIM	WITH	TA	OBSTACLE
PXT	MIL	IS	KING	LOUPE
DDM	CLB	FEELS	HIM	PROCS
ROBRT	ISSUED	FAA	VFR	RESTRICTION
RADIAL	GULFSTREAM	THAT	APCH	R
OUTBOUND	Z	FLC	ADVISED	HEADING
LEGS	DALAS	BOG	ARROW	TEB

CA6 Set 24	CA6 Set 25	CA6 Set 26	CA6 Set 27	CA6 Set 28
TKOF	PARKING	PDC	INS	RVR
TWR	PARKED	DEP	WAYPOINT	VISIBILITY
HOLD	BRAKE	XPONDER	TRACK	MINIMUMS
POS	GATE	CODE	GANDER	APCH
RWY	TUG	CLRNC	COORDINATES	CAT
ACR	THE	ACARS	WAYPOINTS	FOG
CLRED	RAMP	SQUAWK	ROUTE	II
CLRNC	JETWAY	AACES	INS'S	ILS
SIGN	MARSHALLER	FILED	W	III
READY	WINGTIP	DELIVERY	OCEANIC	RWY
CALL	BRAKES	FLT	UNITS	ATIS
INTO	ACFT	RNAV	ISANI	RPTED
SHORT	WING	READ	COURSE	TOUCHDOWN
FOR	TRUCK	CORRECT	MERLY	MI
HEARD	STOP	PDC'S	N	WX
Y	PARK	DVC	DITCH	RVV
MOVER	SHUT	PLAN	#	ROLLOUT
XYZ	PERSONNEL	TRANSPONDER	INSERTED	RVR'S
SIMILAR	AREA	AGENT	ACCURACY	HT
X	ENGS	AMENDMENT	ESTIMATE	LIGHTS

CA6 Set 29	CA6 Set 30	CA6 Set 31	CA6 Set 32	CA6 Set 33	CA6 Set 34
SMA	CLB	RTE	FMC	FUEL	AUTOPLT
Y	FT	PLAN	DSCNT	LBS	ALT
TFC	TCASII	FILED	VNAV	TANK	CAPTURE
PLT	TFC	ROUTING	RESTRICTION	PUMP	MODE
EVASIVE	#	FLT	LNAV	GAUGE	ENGAGED
STUDENT	RA	DIRECT	PROGRAMMED	TANKS	FT
X	DEP	FMS	XING	ALTERNATE	SELECTED
ACR	O'CLOCK	COMPUTER	RESTR	EMER	DSCNT
LEFT	CLBING	FMC	PROGRAMMING	BURN	AIRSPD
TUPELO	TA	LOADED	PAGE	LOAD	SPD
COLLISION	RATE	CLRNC	MODE	FUELING	LEVELOFF
TWR	FPM	ORIGINAL	LAS	QUANTITY	DISCONNECTED
ACTION	HDG	PDC	SELECTED	FUELER	VERT
HE	ISSUED	AIRWAY	FIX	XFEED	AUTOTHROTTLES
NEAR	DEGS	DEP	ORVIL	PUMPS	WINDOW
VFR	LEVEL	CHANGE	ARR	POUNDS	VNAV
O	TARGET	PAGE	CROSS	WT	FMA
INSTR	PASSING	ACARS	RESTRICTIONS	GAUGES	LEVEL
ATX	AT	LWB	THE	STL	SELECT
HIM	TURN	ETOPS	RNAV	DISPATCH	MANUALLY

CA6 Set 35	CA6 Set 36	CA6 Set 37	CA6 Set 38	CA6 Set 39	CA6 Set 40
DIRECT	INSPECTOR	VFR	RAMP	DOWNWIND	AIRSPACE
VOR	FAA	IFR	GND	PATTERN	CLASS
COURSE	JUMP	CONDITIONS	TAXI	RWY	B
NAV	SEAT	WX	GATE	FINAL	TCA
OMEGA	MGR	CLOUDS	TXWY	CESSNA	FLOOR
SIE	MDW	HOSPITAL	CTL	UNICOM	MSL
INTXN	CHK	I	PUSH	CTAF	C
OTU	FLT	VISIBILITY	SPOT	BASE	FREEWAY
ERROR	MY	MI	PUSHBACK	LAND	VFR
AIRWAY	OXYGEN	CLOUD	ALPHA	STUDENT	HAYWARD
CKB	CERTIFICATES	SCATTERED	TAXIING	ANNOUNCED	SHELF
DEG	CERTIFICATE	OVCST	OUTER	TFC	CORRIDOR
SLIDR	MEDICAL	LAYER	INNER	ARPT	TFC
SJC	COCKPIT	BROKEN	PUSHED	HELI	PHX
CMK	MASTER	FSS	X	GAR	RING
RBS	ACI	ZZZ	TAXIED	LNDG	LUKE
CLUCK	FAX	ZOA	CLRNC	ON	ALPINE
ARR	AIRMAN	FOG	ACTIVE	I	SPORT
GFMS	JUMPSEAT	MY	ONTO	UPWIND	D
TRACK	MECH	PATIENT	AREA	DURANGO	I

CA6 Set 41	CA6 Set 42	CA6 Set 43	CA6 Set 44	CA6 Set 45
I	FREQ	ALT	TCAS	#
HE	COM	FT	CLB	ALT
IT	RADIO	ASSIGNED	RA	CTR
MY	VHF	ALERTER	TFC	CLB
THIS	CONTACT	#	ADVISORY	CLRNC
EXPERIENCE	CTR	ATC	RVSM	MACH
IS	BWI	ALERT	II	GANDER
HAS	LOST	PF	DSNT	DSCNT
CAPT	VOLUME	DSNDED	LEARJET	LEVEL
FLYING	MIKE	CLBED	CONFLICT	DSND
ME	ZAN	LEVELOFF	RESOLUTION	MERIDA
AIRPLANE	QUIET	PNF	WE	OCEANIC
SOME	KNOB	ALTDEV	DEVIATION	HAVANA
ONE	RADIOS	DEV	LEVEL	AT
WAY	CODE	SET	AN	DSNT
CAN	CENTER	AUTOPLT	DELLS	TO
DRUZZ	CHANGE	DISTR	VSI	CRUISE
YRS	ANCHORAGE	THROUGH	CLBING	REQUESTED
HIS	SWITCH	CLBING	RECEIVED	HIGHER
GET	CALL	LEVEL	OUR	BUFFET

CA6 Set 46	CA6 Set 47	CA6 Set 48	CA6 Set 49	CA6 Set 50
VISUAL	APCH	KTS	RESTR	FT
APCH	LOC	SPD	XING	CLRNC
SIGHT	ILS	AIRSPD	DSCNT	CTLR
ARPT	INTERCEPT	SLOWED	CROSS	ALT
RWY	GS	SLOW	FMS	READ
FIELD	DME	#	ARR	READBACK
BASE	COURSE	SLOWING	FT	#
FINAL	FAF	KIAS	AT	SAID
DOWNWIND	PLATE	SPACING	DME	CLRED
FOR	TUNED	FT	INTXN	BACK
LINED	RWY	RESTR	MAKE	MAINTAIN
LOC	ESTABLISHED	JAMMN	RESTRS	CALL
TWR	VECTORED	KT	LTOWN	DSCNT
LAND	VECTORS	KARLA	MI	HEARD
FOLLOW	INTERCEPTED	STABILIZED	CROSSED	DSND
VIS	VECTOR	REDUCTION	NM	ACKNOWLEDGED
MI	MISSED	KRENA	#	US
ALB	VOR	WE	KORRY	SIGN
WE	THE	ASKED	HIGH	ASKED
MLG	MDA	SPDS	STAR	ATC

CA6 Set 51	CA6 Set 52	CA6 Set 53	CA6 Set 54
ALTIMETER	KTS	WT	TXWY
SETTING	WIND	DATA	RWY
ALTIMETERS	WINDS	DISPATCH	TAXI
#	BRAKING	LOAD	ONTO
RESET	LNDG	PAPERWORK	GND
FT	REVERSE	FLT	TXWYS
ALT	FLARE	BAGS	K
SET	TOUCHDOWN	BAL	SIGN
ATIS	DAMAGE	MANIFEST	B
LEVELED	FLAPS	AGENT	H
LEVEL	NORMAL	MAX	M
INCHES	VREF	PAX	P
QNH	XWIND	TKOF	L
CAPT'S	TAILWIND	COUNT	E
LOW	THRUST	WTS	END
LCL	DOWN	NOTAM	TURN
QFE	TAIL	OPS	TAXIING
CORRECTED	THRESHOLD	FUEL	RAMP
READ	AGL	CLOSEOUT	VIA
SETTINGS	SINK	LBS	TAXIED

Appendix E. Division of Commercial Aviation Document Sets

Division of the commercial aviation sets of documents into each of the 31 categories

Categories	CA1	CA2	CA3	CA4	CA5	CA6
Wind	14	44	49	40	23	
			15			
Ice		2		49	22	
Weather	8	7	30	2	7	28
	46	26	40	32		37
	50	35		46		52
Air Collision / TCASII	20	3	2	23	4	8
	31	24	5	25	8	12
	32	31	16	35	32	20
	41	39	29	47	40	22
	48	46	53		41	44
Restricted Airspace	26	10	14	33	2	32
	44		18		39	40
			21		50	49
Flight Plan	42	21	48	39	10	31
Navigation	12	9	6	7	11	9
	23	14	7	17	20	16
	29	28	11	26	25	23
	36	34	41	44	28	27
		38		52	38	35
		47			45	
Altitude	34	23	19	4	14	43
	53	48	25	20	43	51
			36			
			46			
Speed	7	4	28	54	18	48
Landing Gear	24	33	24	24	47	10
					34	
Engine Issues	27	19	38	38	44	13
Autopilot	43	8	10	1	24	34
Weight	10	37	54	19	29	33
	39	54			35	53
FAA Inspection	28			36	46	36
Maintenance Inspection	40	13	31	10	27	2
		18	50	11		
Cabin & Passenger Issues	9	5	12	21	48	3
	19	12	17	53		5
			37			

Categories	CA1	CA2	CA3	CA4	CA5	CA6
ATC	17 49	6 15 16	1 22	34	53	26 50
Communication / Radio	1 30	53	20		3 42	42
Fatigue	54	52	4	14	19	11
Taxi	11 13 47	22 43	13 23 35 45	8 9 18	5 15 36 37 49	4 7 14 54
Runway Issues	35 38	49		16		
Parking / Pushback	3	11 17	32 42	28 43	12 33	38 25
Take-off	21 22 51	20 30	44 52	12 30 37 48 51	1 13 30 54	17 24 30
Landing	25	25 29 32	33	22 42	16	15
Visual Approach	52	41		15 27	17 51	46
Descent / Approach	45	51	9 27 39	3 13 31	21 26	47
Holding	5		8		52	19
Location Issues	4 15 16 33	1 42 45 50	3 51	29 45 50	9	1 18
Reporter Callback	6	36	47	41	6	21
Helicopter Issues	18	27	43	5		39
Miscellaneous (including reporter callback and helicopter issues)	2 37	40	26 34	6	31	6 29 41 45

Multiple document clusters were collapsed in the construction of the 31-category solution. A ratio of the number of repeated keywords to the total number of keywords represented within that selection of documents was calculated as a measure of the equivalence between the sets collapsed within the category. The following table presents the ratio for each of the categories in which multiple clusters were a part of its construction.

Categories	Commercial Aviation (CA) Sets					
	CA1	CA2	CA3	CA4	CA5	CA6
Wind			0.10			
Ice						
Weather	0.10	0.08	0.10	0.12		0.17
Air Collision / TCASII	0.40	0.47	0.48	0.49	0.43	0.44
Restricted Airspace	0.38		0.20		0.13	0.17
Flight Plan						
Navigation	0.35	0.42	0.39	0.38	0.38	0.35
Altitude	0.33	0.30	0.40	0.25	0.25	0.25
Speed						
Landing Gear					0.20	
Engine Issues						
Autopilot						
Weight	0.15	0.20			0.10	0.25
FAA Inspection						
Maintenance Inspection		0.30	0.10	0.20		
Cabin & Passenger Issues	0.10	0.05	0.20	0.10		0.05
ATC	0.05	0.28	0.00			0.10
Communication / Radio	0.05				0.10	
Fatigue						
Taxi	0.60	0.40	0.51	0.45	0.56	0.50
Runway Issues	0.10					
Parking / Pushback		0.25	0.20	0.20	0.15	0.15
Take-off	0.07	0.15	0.10	0.19	0.15	0.03
Landing		0.32		0.00		
Visual Approach				0.31	0.35	
Descent / Approach			0.22	0.08	0.30	
Holding						
Location Issues	0.18	0.20	0.30	0.20		0.20

Division of the commercial aviation sets of documents into each of the 9 categories

Categories	CA1	CA2	CA3	CA4	CA5	CA6
Weather	8	2	15	2	7	28
	14	7	30	32	22	37
	46	26	40	40	23	52
	50	35	49	46		
		44		49		
SA	12	3	2	7	2	8
	20	9	5	17	4	9
	23	10	6	23	8	12
	26	14	7	25	10	16
	29	21	11	26	11	20
	31	24	14	33	20	22
	32	28	16	35	25	23
	36	31	18	39	28	27
	41	34	21	44	32	31
	42	38	29	47	38	32
	44	39	41	52	39	35
	48	46	48		40	40
		47	53		41	44
					45	49
					50	
Attention / Monitoring	7	4	10	1	14	10
	24	8	19	4	18	13
	27	19	24	20	24	34
	34	23	25	24	34	43
	43	33	28	38	43	48
	53	48	36	54	44	51
			38		47	
			46			
Weight	10	37	54	19	29	33
	39	54			35	53
Inspection	28	13	31	10	27	2
	40	18	50	11	46	36
				36		
Interpersonal	9	5	12	21	48	3
	19	12	17	53		5
			37			
Communication	1	6	1	34	3	26
	17	15	20		42	42
	30	16	22		53	50
	49	53				

Categories	CA1	CA2	CA3	CA4	CA5	CA6
Physiological	54	52	4	14	19	11
Context	3	1	3	3	1	1
	4	11	8	8	5	4
	5	17	9	9	9	7
	11	20	13	12	12	14
	13	22	23	13	13	15
	15	25	27	15	15	17
	16	29	32	16	16	18
	21	30	33	18	17	19
	22	32	35	22	21	24
	25	41	39	27	26	25
	33	42	42	28	30	30
	35	43	44	29	33	38
	38	45	45	30	36	46
	45	49	51	31	37	47
	47	50	52	37	49	54
	51	51		42	51	
	52			43	52	
				45	54	
				48		
				50		
				51		

Similar to the calculations done to explore the similarity of keywords present in the document sets combined to create the 31-catgory solution, a ratio of the number of repeated keywords to the number of total terms was calculated for the 9-category solution. The following table presents these ratios.

Categories	Commercial Aviation (CA) Sets					
	CA1	CA2	CA3	CA4	CA5	CA6
Weather	0.15	0.09	0.15	0.09	0.03	0.17
Situation Awareness	0.46	0.45	0.44	0.48	0.46	0.41
Attention / Monitoring	0.33	0.30	0.40	0.25	0.25	0.25
Weight	0.15	0.20			0.10	0.25
Inspection	0.00	0.30	0.10	0.13	0.00	0.05
Interpersonal	0.10	0.05	0.20	0.10		0.05
Communication	0.08	0.21	0.03		0.13	0.13
Physiological Context	0.42	0.47	0.42	0.45	0.44	0.41

Appendix F. General Aviation Key Words

Top 20 Keywords for each of the 35 clusters from the 1st sample of GA documents:

GA1 Set 1	GA1 Set 2	GA1 Set 3	GA1 Set 4	GA1 Set 5
FORMATION	TFR	RPTR	IFR	TFC
AGL	RESTR	CALLBACK	PLAN	O'CLOCK
OVER	AREA	REVEALED	CLRNC	ACFT
LOW	ZZZ	HE	FLT	TCASII
AEROBATIC	AIRSPACE	CONVERSATION	FILED	OTHER
LAKE	FLT	INFO	VFR	EVASIVE
BEACH	TFR'S	FOLLOWING	FSS	FT
BOAT	NOTAMS	FAA	FILE	PASSED
PHOTO	TEMPORARY	STATES	CANCEL	COLLISION
AREA	RESTRS	HIS	HOUSTON	DSNDING
PASSES	NOTAM	HAS	CANCELLATION	MISS
WATER	BRIEFING	WITH	LFI	OUR
FLYING	CHART	STATED	DIRECT	SAW
POPULATED	YYY	IS	FILING	ACTION
PEOPLE	BRIEFER	FEELS	CUSTOMS	RA
AREAS	AREAS	FSDO	CTR	US
MANEUVERS	SECTIONAL	BELIEVES	FREQ	NEAR
RACE	DUATS	OFFICE	RTE	JET
AEROBATICS	FSS	ANALYST	VOID	#
AERIAL	MOA	MTR	CONTACT	SEPARATION

GA1 Set 6	GA1 Set 7	GA1 Set 8	GA1 Set 9	GA1 Set 10
PROP	LOC	BALLOON	SMA	TXWY
DOOR	APCH	BASKET	Y	RWY
ENG	ILS	LINES	X	TAXI
THE	GS	WIND	DOWNWIND	GND
PARKED	INTERCEPT	BALLOONS	HE	ONTO
DAMAGE	MISSED	ENVELOPE	SMT	TAXIING
RAMP	VECTORS	LAUNCH	TFC	RAMP
HANGAR	NEEDLE	HOT	Z	TAXIED
WING	COURSE	PWR	PLT	ACTIVE
WINGTIP	MARKER	BURNER	RIGHT	CROSS
ACFT	VECTOR	SITE	TWR	TXWYS
PARKING	FAF	PAX	LEFT	FBO
TAXIING	SET	LNDG	HIM	INSTRUCTIONS
STRUCK	NAV	PWRLINES	BASE	DIAGRAM
OIL	VECTORED	DAMAGE	FINAL	ACROSS
HIT	APCHS	WINDS	HIS	RWYS
FORWARD	HOBBO	TARGET	PATTERN	SHORT
L	DME	FIELD	LCL	CROSSED
BAR	MINIMUMS	INFLATION	TOUCH	PROGRESSIVE
INSPECTED	PLATE	WIRES	EVASIVE	INTXN

GA1 Set 11	GA1 Set 12	GA1 Set 13	GA1 Set 14	GA1 Set 15
MAINT	CLASS	ARSA	HELI	FUEL
FAA	AIRSPACE	BUR	HELIS	TANK
CERTIFICATE	B	ATA	HELI'S	TANKS
DUTY	C	VNY	NEWS	ENG
MEDICAL	D	OUTER	POLICE	GAUGES
INSPECTION	FT	CONTACT	HOVERING	GALLONS
CREW	FLOOR	OQU	MIL	GALS
LOG	VFR	SNA	PHOTO	HRS
AIRWORTHINESS	BAY	REPORTER	CHASE	GAUGE
OWNER	CLR	WHP	ROTOR	EMER
LIMITATIONS	REMAIN	PORTLAND	NAVY	EMPTY
REGISTRATION	MSL	SMYRNA	CHOPPER	FULL
FLT	SQUAWK	LAX	ULTRALIGHT	HR
LOGBOOK	MOFFETT	O	SCENE	QUIT
COMPANY	XPONDER	RADAR	AUTOS	MINS
FORM	CODE	I	PHOTOGRAPHER	GAS
INSPECTOR	RING	ABE	TV	PWR
LOGBOOKS	PHX	TROUTDALE	FIXED	BURN
PINNED	ATL	TRANSPONDER	A	RESERVE
DAY	NM	W	HOVER	AUX

GA1 Set 16	GA1 Set 17	GA1 Set 18	GA1 Set 19	GA1 Set 20
STUDENT	ADIZ	APCH	ICE	HOLD
INSTRUCTOR	POTOMAC	VISUAL	ICING	SHORT
SOLO	CODE	WE	FREEZING	RWY
HIS	DC	MISSED	RIME	LINE
STUDENT'S	PCT	SIGHT	CLOUDS	TAXI
TRAINING	WASHINGTON	CIRCLE	FT	TXWY
XCOUNTRY	SQUAWK	MDA	CONDITIONS	GND
HE	XPONDER	ARPT	MEA	LINES
STUDENTS	TRACON	IFR	TRACE	INSTRUCTIONS
FLT	JYO	CIRCLING	ENCOUNTERED	TAXIING
HIM	FLT	ANW	PICKING	STOPPED
WE	PLAN	MINIMUMS	TOPS	CROSS
SIMULATED	DISCRETE	VECTORED	ACCUMULATION	CROSSED
DUAL	THEY	ILS	PITOT	MARKINGS
MLT	BALTIMORE	PROC	CLOUD	HOLDING
SCHOOL	NY	VISIBILITY	FORECAST	ACTIVE
LNDGS	MARTIN	FORBES	MSL	SIGN
HER	B	RWY	TOP	STOP
COUNTRY	FREQUENCY	CAPT	LOWER	INCURSION
BLYTHE	LEESBURG	FINAL	WINDSHIELD	RWYS

GA1 Set 21	GA1 Set 22	GA1 Set 23	GA1 Set 24	GA1 Set 25
I	ALTIMETER	DEP	PATTERN	VOR
NAV	SETTING	SID	DOWNWIND	RADIAL
ARPT	FT	TEB	RWY	DME
GPS	ALT	PROC	FINAL	ARC
VOR	#	FT	UNICOM	APCH
MY	SET	CLB	TFC	NAV
LORAN	ELEVATION	#	ANNOUNCED	DEG
SECTIONAL	MSL	CLRNC	CTAF	DIRECT
NIGHT	RESET	WE	BASE	INTERCEPT
RADIO	SETTINGS	ABATEMENT	RADIO	FIX
BEACON	MODE	TURN	ACFT	COURSE
COM	COPLT'S	DEGS	OTHER	WE
CHART	C	PF	ON	FMS
FDK	ALTIMETERS	BRIEFED	HEARD	GPS
COORDINATES	LEVELED	CAPT	CESSNA	CTR
GUARD	VSI	HDG	HE	PROC
LIGHTS	STATIC	TKOF	LEG	OUTBOUND
WAYPOINT	READING	HEADING	UNCTLED	ARR
CITY	BKF	NOISE	TURNING	INTXN
FREQ	PRESSURE	CREW	L	COPLT

GA1 Set 26	GA1 Set 27	GA1 Set 28	GA1 Set 29	GA1 Set 30
CLOUDS	ALT	GEAR	TCA	WIND
VFR	FT	HORN	SAN	RUDDER
WX	WE	LNDG	CHART	LNDG
VISIBILITY	CLB	DOWN	ATA	RWY
CONDITIONS	CTR	WARNING	FLOOR	KTS
CLOUD	DSCNT	FLAPS	DIEGO	DAMAGE
LAYER	OUR	LEVER	SPIRIT	L
CEILING	ASSIGNED	CHKLIST	MONTGOMERY	BRAKES
OVCST	CLRNC	THE	MODE	ACFT
SCATTERED	CTLR	UP	GILLESPIE	GRASS
BROKEN	DSND	SWITCH	LAX	NOSE
IFR	#	PROP	BAY	XWIND
MI	US	RETRACTED	MIRAMAR	PLANE
IMC	WERE	THROTTLE	LOS	APPLIED
FOG	COPLT	FLAP	W	BRAKING
I	CAPT	GREEN	ANGELES	THE
RAIN	CLRED	EXTENDED	PHL	R
FORECAST	FGT	GUMP	O	WHEEL
HOLE	XING	HANDLE	XPONDER	PWR
FT	MAINTAIN	NOSE	VFR	BRAKE

GA1 Set 31	GA1 Set 32	GA1 Set 33	GA1 Set 34	GA1 Set 35
HDG	TKOF	AUTOPLT	CLOSED	TWR
DEGS	RWY	ALT	NOTAMS	RWY
TURN	TAXI	FT	RWY	LAND
DEP	TWR	CAPTURE	NOTAM	FINAL
DEG	CLRNC	ASSIGNED	ARPT	BASE
CLRNC	CLRED	SELECT	FSS	CLRED
COMPASS	READY	DSCNT	THRESHOLD	ME
ASSIGNED	DEP	ENGAGED	CLOSURE	TFC
#	HOLD	CAPT	DISPLACED	DOWNWIND
CLB	GND	DISCONNECTED	X'S	TOLD
CTLR	TAXIED	DIRECTOR	THERE	RPT
WE	ROLL	COPLT	LANDED	R
HEADING	ONTO	ALERTER	RUNWAY	I
HSI	FOR	BUTTON	X	TOUCH
MAINTAIN	INTXN	PRESELECT	UNICOM	MI
RADIAL	DEPART	FO	CONSTRUCTION	SIGHT
INSTRUCTIONS	POS	DISENGAGED	BULVERDE	FOLLOW
INTERCEPT	HEARD	SET	FIELD	L
DIRECT	I	TRIM	WILLIAMS	GO
US	RUNUP	LEVEL	NO	LNDG

Top 20 Keywords for each of the 35 clusters from the 2nd sample of GA documents:

GA2 Set 1	GA2 Set 2	GA2 Set 3	GA2 Set 4	GA2 Set 5
GLIDER	DEP	TXWY	VOR	I
TOW	SID	RWY	NAV	FREQ
GLIDERS	TEB	TAXI	GPS	ARPT
ROPE	FT	GND	COURSE	ME
TOWING	CLB	TAXIING	RADIAL	FTG
SARATOGA	WE	ONTO	DIRECT	TWR
BANNER	DME	DIAGRAM	ERROR	TOLD
THERMAL	#	ACTIVE	FMS	RADIO
SOARING	PROC	RAMP	FIX	THEM
PIM	TURN	TAXIED	DME	MY
HOOK	SIC	PROGRESSIVE	DEG	SAID
FARM	CLRNC	INTXN	INTERCEPT	CALL
PLANE	MAINTAIN	CROSSED	AIRWAY	PHONE
G	BOACH	CROSS	RNAV	FREQS
MANSFIELD	DEGS	INSTRUCTIONS	OUTBOUND	VERO
MNN	HDG	FBO	NEEDLE	SHE
JEAN	RESTRS	TXWYS	HSI	SO
LAUNCH	MDW	END	HDG	CALLED
SPOILERS	PNF	SIGN	ATC	CRYSTAL
LGC	ALT	SHORT	CDI	HWV

GA2 Set 6	GA2 Set 7	GA2 Set 8	GA2 Set 9	GA2 Set 10
TCA	GEAR	ICE	STUDENT	TFC
CHART	LNDG	ICING	INSTRUCTOR	O'CLOCK
VFR	FLAPS	CARB	SOLO	OTHER
LA	HORN	HEAT	STUDENT'S	ACFT
SAN	DOWN	RIME	HIS	EVASIVE
TCA'S	WARNING	PITOT	XCOUNTRY	PASSED
TRANSPONDER	CHKLIST	SNOW	STUDENTS	FT
MYF	PROP	CLOUDS	TRAINING	SAW
MODE	THE	FREEZING	HE	COLLISION
ATA	UP	ACCUMULATION	SHE	US
MSL	RETRACTED	TOPS	ENG	TCASII
FLOOR	NOSE	BOOTS	CTLS	OUR
NY	LEVER	WINDSHIELD	HER	ACTION
LAX	DAMAGE	IMC	ENDORSEMENT	APPROX
CORRIDOR	PWR	AIRFRAME	VERO	MISS
DET	CHK	PWR	LNDGS	RA
SQUAWK	NORMAL	PIREPS	THE	TWIN
PIT	HANDLE	LOWER	MULTI	WING
HHR	LOCKED	LEADING	CFI	PATH
BOS	GREEN	AIRSPD	HIM	CESSNA

GA2 Set 11	GA2 Set 12	GA2 Set 13	GA2 Set 14	GA2 Set 15
TWR	LIGHTS	IFR	MEDICAL	CLASS
RWY	RWY	PLAN	CERTIFICATE	AIRSPACE
LAND	NIGHT	CLRNC	FAA	B
FINAL	LIGHT	VFR	MAINT	C
BASE	LIGHTING	FLT	REGISTRATION	D
CLRED	BEACON	FILED	ANNUAL	FT
TFC	VISIBILITY	FSS	COMPANY	FLOOR
TOUCH	INTENSITY	CONTACT	INSPECTOR	CHART
DOWNWIND	EDGE	CANCEL	INSPECTION	MSL
CTLR	THE	VOID	PART	TERMINAL
FOR	ARPT	SQUAWK	MECH	VFR
LNDG	VASI	CENTER	PLT	CORRIDOR
GO	CTRLINE	RELEASE	PERMIT	ENTERED
WE	FOG	CTLR	FERRY	CLB
SIGHT	DARK	DEP	OWNER	CLR
US	DONALDSON	VMC	PAPERWORK	OUTER
RPT	LIT	CTR	AIRWORTHINESS	RING
ON	END	CANCELLED	LOGBOOKS	SECTIONAL
GAR	LIGHTED	DIRECT	LOGBOOK	INCURSION
TOLD	PAPI	RADAR	EXAMINER	AREA

GA2 Set 16	GA2 Set 17	GA2 Set 18	GA2 Set 19	GA2 Set 20
BALLOON	HDG	PATTERN	THE	TFR
BASKET	DEGS	DOWNWIND	RWY	ZZZ
LINES	TURN	CESSNA	RUDDER	TFR'S
ENVELOPE	DEG	RWY	WIND	NOTAMS
PWR	DEP	ANNOUNCED	L	NOTAM
BALLOONS	COMPASS	FINAL	NOSE	TFRS
LINE	CTLR	TFC	KTS	PLANT
WIRES	CLRNC	BASE	BRAKES	BRIEFING
HOT	DIRECTIONAL	UNICOM	DAMAGE	FSS
PAX	GYRO	ACFT	APPLIED	MILES
LAUNCH	WE	HE	MAIN	FDC
BURNER	HEADING	RADIO	GRASS	AIRSPACE
FABRIC	BUG	OTHER	PROP	FLT
RPTR	#	CTAF	R	STADIUM
VENT	ASSIGNED	ENTRY	LNDG	PLANTS
CREW	COURSE	HEARD	TOUCHDOWN	RESTR
POLE	VECTORS	XWIND	DOWN	BRIEFER
TARGET	DIRECT	HIS	ACFT	RADIUS
THE	INTERCEPT	L	WHEEL	AREA
AIR	RADIAL	LEG	PLANE	EFFECT

GA2 Set 21	GA2 Set 22	GA2 Set 23	GA2 Set 24	GA2 Set 25
HELI	ADIZ	RESTR	SMA	CLOSED
AGL	POTOMAC	AREA	Y	NOTAMS
AREA	CODE	AIRSPACE	X	NOTAM
HELIS	WASHINGTON	MOA	DOWNWIND	FSS
OVER	PLAN	WACO	Z	UNICOM
PASS	SQUAWK	AREAS	TFC	RWY
LOW	FLT	FLT	LEFT	ARPT
WATER	XPONDER	WEST	FINAL	CLOSURE
STADIUM	DC	MIL	PATTERN	BRIEFING
PHOTOGRAPHER	NY	PROHIBITED	SMT	X'S
PEOPLE	FDK	KEY	TWR	NOTAMED
OF	TRACON	ESN	BASE	ATIS
FLYING	FILE	TEMPORARY	AN	WX
JUMPERS	FSS	RESTRS	RIGHT	BRIEFER
SKYDIVERS	MANASSAS	LORAN	SAW	ASOS
DROP	LEESBURG	ALBUQUERQUE	HE	NO
SAFE	HEF	RTE	MTR	DISPLACED
AREAS	PCT	E	GYR	SHR
JUMP	NUMBER	ORLANDO	ACFT	LSZH
AERIAL	BALTIMORE	SERVICE	HIM	MEN

GA2 Set 26	GA2 Set 27	GA2 Set 28	GA2 Set 29	GA2 Set 30
RAMP	TKOF	ALT	DSCNT	APCH
PARKED	RWY	FT	WE	LOC
THE	TWR	AUTOPLT	FT	ILS
PROP	TAXI	ALTIMETER	ALT	MISSED
WING	CLRNC	ASSIGNED	CAPT	GS
ACFT	DEP	SETTING	DSND	INTERCEPT
DAMAGE	READY	ALERTER	US	NDB
ENG	CLRED	#	OUR	PLATE
TAXIING	HOLD	CLB	#	FAF
TXWY	FOR	CAPTURE	CTLR	MDA
PARKING	GND	PF	XING	VECTORED
HANGAR	TAXIED	SET	ARR	VOR
TIE	POS	PNF	CLRNC	OM
TAXI	ONTO	ATC	CLRED	DME
STRUCK	INTXN	LEVEL	WERE	COURSE
PLANE	ROLL	PRESELECT	TCASII	MINIMUMS
SMT	DEPART	ENGAGED	FO	WE
FBO	RUNUP	CLBING	ATC	VECTORS
WINGTIP	CTL	DSCNT	MAINTAIN	PROC
AIRPLANE	HEARD	TRIM	RESTR	APCHS

GA2 Set 31	GA2 Set 32	GA2 Set 33	GA2 Set 34	GA2 Set 35
ARSA	RPTR	FUEL	HOLD	CLOUDS
ATA	CALLBACK	TANK	SHORT	WX
BADER	REVEALED	TANKS	RWY	VFR
CTL	CONVERSATION	ENG	TAXI	VISIBILITY
TRANSPONDER	FOLLOWING	GALLONS	LINE	CONDITIONS
RENTON	HE	GAUGES	TXWY	CLOUD
ZONE	INFO	GALS	GND	SCATTERED
BARBARA	STATES	HR	INSTRUCTIONS	LAYER
CONTACT	HAS	MINS	CROSS	CEILING
EVANSVILLE	FAA	GAUGE	CROSSED	OVCST
SANTA	IS	EMER	TAXIED	IFR
ARSA'S	FEELS	HRS	TAXIING	BROKEN
CHARLESTON	STATED	BURN	ACROSS	MI
WV	WITH	CAP	INSTRUCTION	FOG
PDX	HIS	PUMP	RWYS	RAIN
TTD	ANALYST	RESERVE	RUN	IMC
BOEING	LETTER	CONSUMPTION	CLRED	BELOW
CLARKSBURG	PLTS	RAN	ONTO	HOLE
ACY	THEY	PWR	STOPPED	ENCOUNTERED
TIPTON	REPORTER	QUIT	LINES	CEILINGS

Top 20 Keywords for each of the 35 clusters from the 3rd sample of GA documents:

GA3 Set 1	GA3 Set 2	GA3 Set 3	GA3 Set 4	GA3 Set 5
CLOSED	INSPECTION	ALT	TFR	COMPASS
ARPT	MAINT	FT	NOTAMS	HDG
UNICOM	DOOR	ALTIMETER	ZZZ	GYRO
NOTAM	ENG	#	NOTAM	DEGS
RWY	OIL	ASSIGNED	BRIEFING	INDICATOR
NOTAMS	PREFLT	SETTING	FLT	DIRECTIONAL
CTAF	HR	DSND	BRIEFER	MAGNETIC
LANDED	COMPARTMENT	WE	PLANT	HSI
CLOSURE	LOG	DSCNT	FSS	DG
X'S	ANNUAL	ALERTER	TFR'S	GYROSCOPE
FIELD	XPONDER	CLB	FDC	SLAVED
ON	REMOVED	OUR	POWER	TSTMS
FREQ	BLEED	CTR	MILES	HEBER
CONSTRUCTION	MECH	CTLR	WACO	SPOUSE
NOTAMED	THE	ATC	AIRSPACE	INSTS
NO	COMPLETED	US	PLANTS	I
SOLDOTNA	COVER	SET	LOCAL	RDU
SFZ	APU	CLRED	DUATS	FORT
VBT	GRACE	CLRNC	INFO	MOULTRIE
BRIEFING	PROGRAM	MAINTAIN	YYY	COURSE

GA3 Set 6	GA3 Set 7	GA3 Set 8	GA3 Set 9	GA3 Set 10
ADIZ	CLASS	STUDENT	TCA	RESTR
CODE	AIRSPACE	INSTRUCTOR	ARSA	AREA
POTOMAC	B	HE	LAX	AIRSPACE
WASHINGTON	C	HIS	CHART	MOA
PLAN	D	HIM	VFR	AREAS
SQUAWK	FT	XCOUNTRY	FLOOR	CHART
JYO	MSL	SOLO	ATA	SECTIONAL
PCT	FLOOR	CFI	LA	CHARTS
FLT	PHX	STUDENT'S	SMO	NAV
XPONDER	I	TRAINING	W	MAP
TRACON	VFR	SIMULATED	MONTE	GPS
DC	GPS	LNDGS	TCA'S	SALISBURY
LEESBURG	RING	TEACHING	THROUGH	WAC
HEF	CLR	ENG	MODE	FOLLOWING
DVFR	CHART	TOUCH	CORRIDOR	FLT
NY	BRIDGE	GFL	MSL	NELLIS
DISCRETE	REMAIN	STUDENTS	ANGELES	HOT
FILED	BELOW	PROCS	BURBANK	DIRECT
BRIEFER	CORRIDOR	THROTTLE	ESSEX	RTE
B	ASH	EMER	LOS	PHELPS

GA3 Set 11	GA3 Set 12	GA3 Set 13	GA3 Set 14	GA3 Set 15
HELI	FREQ	BALLOON	RPTR	WIND
ROTOR	RADIO	BASKET	CALLBACK	DAMAGE
HELIS	CONTACT	LINES	REVEALED	RWY
COLLECTIVE	I	ENVELOPE	CONVERSATION	LNDG
HOVER	COM	BALLOONS	HE	NOSE
POLICE	THEY	PWR	FOLLOWING	PWR
WIRE	TWR	SITE	FAA	XWIND
ACCIDENT	ARSA	LAUNCH	HAS	KTS
AUTOROTATIONS	ME	BURNER	INFO	PROP
AGL	APCH	WINDS	HIS	WINDS
TRAVIS	FRG	HOT	STATES	BOUNCED
SHERIFF'S	SIGNALS	WIND	WITH	FLARE
METRO	RADIOS	VENT	STATED	HARD
NEWS	FREQS	LAUNCHED	FEELS	AIRSPD
SCENE	RADAR	FIELD	H	FLAPS
RIVER	CTLR	LNDG	IS	BOUNCE
PHOTOGRAPHER	XT	LINE	O'KNIGHT	PLANE
MEDIA	THEM	DAMAGE	PETER	GUST
MILFAC	HDOF	CREW	BELIEVES	TOUCHED
ORBIT	COMS	ASCENT	OWNER	DOWN

GA3 Set 16	GA3 Set 17	GA3 Set 18	GA3 Set 19	GA3 Set 20
TXWY	SHE	GEAR	CAPT	BRAKES
RWY	HER	LNDG	WE	RUDDER
TAXI	CTLR	DOWN	OUR	APPLIED
GND	SAID	HORN	COPLT	BRAKE
RAMP	MIA	FLAPS	FO	BRAKING
TAXIING	SUSII	THE	US	GRASS
ONTO	FXE	NOSE	HE	L
ACTIVE	DEER	HANDLE	FMS	RWY
TAXIED	SUPVR	WARNING	CAPT'S	R
INTXN	TWR	CHKLIST	COCKPIT	THE
FBO	US	UP	WERE	FULL
SIGN	DULLES	DAMAGE	CREW	PLANE
DIAGRAM	BISCAYNE	PWR	PF	WHEEL
INSTRUCTIONS	WE	GREEN	CO	STOP
TXWYS	MKY	LEVER	CHKLISTS	TOUCHDOWN
CROSS	LVL	SWITCH	SIC	KTS
MARKINGS	NAMPA	PROP	SEAT	DAMAGE
ACROSS	DADE	RETRACTED	HIS	MAIN
VIA	ME	ENG	PNF	TAIL
CROSSED	SPOKE	LOWERED	MARES	ACFT

GA3 Set 21	GA3 Set 22	GA3 Set 23	GA3 Set 24	GA3 Set 25
APCH	SMA	FUEL	DEP	TFC
LOC	Y	TANK	SID	O'CLOCK
ILS	X	TANKS	HDG	ACFT
GS	TFC	ENG	DEGS	OTHER
MISSED	LTT	GAUGES	CLB	EVASIVE
INTERCEPT	ACFT	GALLONS	FT	FT
COURSE	Z	GALS	PROC	#
VECTORED	RIGHT	QUIT	TEB	SEPARATION
NDB	HE	HRS	WE	CLBING
OM	FINAL	GAUGE	TURN	COLLISION
NEEDLE	SAW	BURN	CLRNC	AVOID
VECTORS	INBND	EMPTY	HEADING	PASSED
WE	EVASIVE	PUMP	#	TCASII
DME	BEHIND	CARB	TKOF	TARGET
VECTOR	TWR	SWITCHED	FO	ACTION
INBOUND	ISSUED	GPH	MAINTAIN	US
MARKER	CGA	LBS	READ	RADAR
INTERCEPTED	HIS	HR	CPR	SIGHT
VISUAL	TRNING	TOPPED	BRIEFED	MSL
PLATE	TURN	RESERVE	DEG	APPEARED

GA3 Set 26	GA3 Set 27	GA3 Set 28	GA3 Set 29	GA3 Set 30
IFR	VOR	OVER	AUTOPLT	TWR
PLAN	RADIAL	FAA	ALT	RWY
CLRNC	DIRECT	AGL	TRIM	LAND
VFR	DME	LOW	DIRECTOR	BASE
FLT	FIX	SHOW	CAPTURE	DOWNWIND
FILED	APCH	FLY	FT	FINAL
FSS	FMS	PEOPLE	DSCNT	TFC
CANCEL	DEG	PASS	ASSIGNED	SIGHT
WX	INTXN	AEROBATICS	ENGAGED	CLRED
CONDITIONS	ARR	INSPECTOR	PRESELECT	L
IMC	GPS	FLYING	SELECT	INSTRUCTED
VOID	OUTBOUND	HOUSE	SELECTED	TURN
FILE	WE	WAIVER	VNAV	R
ATC	NAV	AREA	CLB	MI
DEP	COURSE	AEROBATIC	FMS	TOLD
ENHANCED	STAR	BEACH	MODE	FOR
HQM	RNAV	PASSES	RATE	RPTED
DIRECT	RTE	JUMPERS	FPM	FOLLOW
AN	INTERCEPT	WATER	LEVEL	US
MGW	DSCNT	BANNER	PITCH	VISUAL

GA3 Set 31	GA3 Set 32	GA3 Set 33	GA3 Set 34	GA3 Set 35
PATTERN	HOLD	CLOUDS	PROP	TKOF
DOWNWIND	SHORT	WX	WING	RWY
FINAL	RWY	VFR	PARKED	TWR
RWY	TAXI	CLOUD	DAMAGE	TAXI
ANNOUNCED	LINE	CONDITIONS	TIP	HOLD
BASE	TXWY	LAYER	RAMP	TAXIED
TFC	LINES	VISIBILITY	WINGTIP	CLRNC
CESSNA	GND	CEILING	ACFT	ROLL
RADIO	CROSS	SCATTERED	THE	READY
ACFT	INSTRUCTIONS	BROKEN	HANGAR	DEP
CTAF	CROSSED	IMC	STRUCK	ONTO
LEG	TAXIING	FOG	TIE	POS
OTHER	ACROSS	IFR	EDGE	CLRED
UNICOM	STOPPED	ICING	TAXI	RUN
ON	TAXIED	CEILINGS	TAXIING	FOR
L	TWR	I	PARKING	GND
HEARD	ACTIVE	OVCST	LEADING	TAXIING
TURNING	XING	ICE	INCH	JET
XWIND	RAMP	FT	TXWY	ACTIVE
HE	CLRNC	TOP	L	END

Top 20 Keywords for each of the 35 clusters from the 4th sample of GA documents:

GA4 Set 1	GA4 Set 2	GA4 Set 3	GA4 Set 4	GA4 Set 5
PATTERN	PROP	RWY	I	FAA
DOWNWIND	DAMAGE	TAXI	TWR	CERTIFICATE
RWY	ENG	TKOF	FREQ	MAINT
FINAL	THE	CLRNC	ARPT	MEDICAL
TFC	HANGAR	GND	RADIO	INSPECTOR
ANNOUNCED	DOOR	TWR	FREQS	PERMIT
UNICOM	STRUCK	HOLD	ARSA	AIRWORTHINESS
CTAF	ACFT	CLRED	TRIED	LICENSE
OTHER	WING	INSTRUCTIONS	ME	PART
RADIO	OIL	TAXIWAY	MY	LOGBOOKS
BASE	COWLING	TAXIED	CONTACT	ANNUAL
ACFT	BAR	TAXIING	SECTIONAL	LOGBOOK
HEARD	PREFLT	READY	THEY	MR
CALLS	BRAKE	ACTIVE	THEM	FERRY
ENTRY	TOW	ONTO	ATIS	OWNER
CESSNA	BRAKES	SHORT	RADIOS	MECH
ON	PARKED	INTXN	MARYSVILLE	INSPECTION
POS	AIRPLANE	RUN	RESPONSE	COMPENSATION
INTENTIONS	PLANE	RUNUP	ISM	OFFICE
CHEROKEE	PARKING	CTL	COM	COMPANY

GA4 Set 6	GA4 Set 7	GA4 Set 8	GA4 Set 9	GA4 Set 10
CLASS	CTLR	FUEL	TFR	SMA
AIRSPACE	FT	TANK	TFR'S	Y
B	ALT	TANKS	NOTAMS	X
D	WE	ENG	FLT	Z
C	MAINTAIN	GALLONS	ZZZ	SMT
FLOOR	DSND	GAUGES	BRIEFING	RIGHT
MSL	#	HRS	BRIEFER	ACFT
REMAIN	CLRED	GALS	P	TFC
FT	DSCNT	HR	FSS	FINAL
VFR	CLRNC	GAUGE	WACO	LEFT
CLR	OUR	MINS	SVC	ATA
SQUAWK	CLB	QUIT	TFRS	HE
FFZ	CENTER	EMER	PLANT	AN
SDL	APCH	RESERVE	VIOLATED	PLT
BURKE	RESPONDED	PUMP	MILES	BEHIND
ENTERED	ASSIGNED	BURN	VEGAS	BASE
BOUNDARY	US	GAS	WX	PATTERN
DVT	SAID	FULL	EFFECT	HIS
AREA	HDG	EMPTY	RESTRS	OBSERVED
TRAVIS	ATC	CONSUMPTION	AREA	MOFFETT

GA4 Set 11	GA4 Set 12	GA4 Set 13	GA4 Set 14	GA4 Set 15
ICE	ADIZ	TXWY	AGL	TWR
ICING	POTOMAC	RWY	OVER	DOWNWIND
HEAT	CODE	TAXI	LOW	RWY
CARB	WASHINGTON	GND	HOUSE	LAND
PITOT	XPONDER	RAMP	LAKE	FINAL
RIME	DC	ONTO	AREA	BASE
CONDITIONS	PLAN	DIAGRAM	TOWN	TFC
FREEZING	FLT	TAXIING	POPULATED	CLRED
TOPS	TRACON	TXWYS	FLYING	L
INCH	DISCRETE	TAXIED	BOAT	TOUCH
STATIC	GAI	ACTIVE	PARAMOTOR	INSTRUCTED
TRACE	FSS	FBO	PEOPLE	R
FT	SQUAWK	SIGNS	AEROBATIC	CTLR
CLOUDS	NY	MARKINGS	BANNER	FOR
CLB	PCT	RWYS	FT	PATTERN
ROUGH	MILES	END	BEACH	ENTER
BOOTS	SQUAWKING	PARALLEL	ALT	FOLLOW
ENCOUNTERED	B	INTXN	AREAS	ME
CTR	FDK	ON	FORMATION	MI
TOP	LEESBURG	CROSS	TOW	CESSNA

GA4 Set 16	GA4 Set 17	GA4 Set 18	GA4 Set 19	GA4 Set 20
DEP	ALT	JUMPERS	GEAR	STUDENT
SID	AUTOPLT	JUMP	HORN	SOLO
TEB	FT	DROP	LNDG	INSTRUCTOR
FT	ALTIMETER	PARACHUTE	DOWN	STUDENT'S
HDG	ASSIGNED	ZONE	FLAPS	HIS
TURN	SETTING	SKYDIVERS	WARNING	SHE
CLB	DSCNT	MOORE	HANDLE	HER
PROC	#	FITCHBURG	CHKLIST	STUDENTS
DME	CLB	OPS	THE	SIMULATED
DEGS	SET	LOAD	UP	ENDORSEMENT
#	WE	FIT	GREEN	LNDGS
WE	OUR	SKYDIVING	PROP	XCOUNTRY
TKOF	ALERTER	PARACHUTISTS	DAMAGE	TRAINING
CLRNC	CAPTURE	TEMPLE	NOSE	HIM
MAINTAIN	ENGAGED	JUMPING	RETRACTED	CFI
BRIEFED	ALTIMETERS	SKYDIVE	EXTENDED	THE
ALT	LEVELOFF	PORTLAND	PWR	HE
US	COPLT	YY	LOCKED	PWR
FO	LEVEL	GRAY	FLAP	DUAL
CAPT	ATC	MOXEE	LEVER	LNDG

GA4 Set 21	GA4 Set 22	GA4 Set 23	GA4 Set 24	GA4 Set 25
LAX	APCH	RESTR	HOLD	RPTR
CORRIDOR	ILS	AREA	SHORT	CALLBACK
ANGELES	LOC	AIRSPACE	RWY	REVEALED
SMO	MISSED	ZZZ	LINE	CONVERSATION
LOS	GS	TEMPORARY	TAXI	FOLLOWING
SHORELINE	INTERCEPT	SECTIONAL	TXWY	HE
SANTA	APCHS	GPS	GND	INFO
MONICA	MDA	FLT	LINES	STATES
SOCAL	MINIMUMS	SFR	CROSS	FAA
SPECIAL	OM	AREAS	CROSSED	HAS
RTE	COURSE	BRIEFING	INSTRUCTIONS	STATED
DIEGO	INBOUND	HSI	STOPPED	WITH
RULES	VOR	PLANT	TAXIING	HIS
BURBANK	WE	NOTAMS	INTXN	ANALYST
CLASS	VECTORED	LORAN	TAXIED	SWITCHES
LGB	PLATE	FIRE	TWR	FEELS
FREEWAY	PROC	CHART	SIGN	RVSM
SAN	DME	YYY	ACTIVE	MLT
VNY	NDB	NOTAM	RAMP	BELIEVES
AIRSPACE	MARKER	RESTRS	ACROSS	ACFT

GA4 Set 26	GA4 Set 27	GA4 Set 28	GA4 Set 29	GA4 Set 30
IFR	CAPT	TFC	VOR	BALLOON
PLAN	WE	ACFT	DIRECT	BASKET
VFR	HE	O'CLOCK	RADIAL	ENVELOPE
CLRNC	FO	PASSED	DEG	LINES
FSS	OUR	OTHER	INTXN	LINE
FLT	PF	EVASIVE	INTERCEPT	BALLOONS
FILED	CAPT'S	COLLISION	HDG	FIELD
VOID	CREW	CLBING	FMS	PWR
FILE	WERE	FT	NAV	POLE
CANCEL	HIS	SAW	DEGS	TREE
CONDITIONS	COMPANY	TCASII	DEP	HOT
RELEASE	TRIP	US	COURSE	SITE
WX	PNF	OUR	FILED	LAUNCH
PHONE	US	#	CLRNC	TARGET
CUSTOMS	FMS	MISS	AIRWAY	WIRE
PICK	WT	ACTION	FIX	LNDG
DEP	COCKPIT	SEPARATION	RTE	WINDS
SVFR	DUTIES	TWIN	RNAV	DAMAGE
SQUAWK	DSCNT	R	ROUTING	HOUSE
DENVER	ASSIGNED	CESSNA	OBS	WIND

GA4 Set 31	GA4 Set 32	GA4 Set 33	GA4 Set 34	GA4 Set 35
CLOSED	TCA	WIND	HELI	CLOUDS
NOTAMS	SAN	NOSE	POLICE	VFR
NOTAM	ARSA	THE	HELIS	WX
RWY	CHART	KTS	HOVER	VISIBILITY
CLOSURE	BRIDGE	RWY	ROTOR	CONDITIONS
ARPT	BAY	PWR	WRITER	CLOUD
FSS	DME	LNDG	FIXED	SCATTERED
BRIEFING	NY	RUDDER	HOSPITAL	LAYER
UNICOM	SQUAWK	DAMAGE	HELIPORT	IFR
NOTAMED	CONTACT	XWIND	ENSTROM	CEILING
WALLA	FLOOR	L	EMS	BROKEN
MGR	CHARTS	APPLIED	BANNER	I
MARKINGS	CLEAR	WHEEL	HELI'S	FT
X'S	SALT	BRAKING	PATTERNS	FOG
WX	BOSTON	PROP	PAD	IMC
DISPLACED	TCA'S	BRAKES	HELIPAD	CEILINGS
TRENCH	SFO	PLANE	APPROX	OVCST
ACTIVITY	REMAIN	GRASS	GAT	MI
BRIEFER	TRANSPONDER	FULL	HIS	FORECAST
LANDED	DPA	ACFT	AREA	HOLE

Appendix G. Division of General Aviation Document Sets

Division of the general aviation sets of documents into each of the 33 categories

Categories	GA1	GA2	GA3	GA4
Weather	26	35	33	35
Wind	30	19	15	33
Ice	19	8		11
Fuel / Weight	15	33	23	8
Altitude	22	29	3	7
	27	28		
Autopilot Control	33		29	17
Instrument Flight	4	13	26	26
ILS Approach	7	30	21	22
	18			
Break Issues			20	
Landing Gear	28	7	18	19
Propeller Issues	6		34	2
Student / Instructor	16	9	8	20
NOTAMs / TFRs	34	20	1	9
		25	4	31
Communication / Radio		5	12	4
			17	
Restricted Airspace	2	15	6	6
	12	22	7	12
	17	23	10	21
				23
Navigation	25	4	27	29
	21	17	5	
	31			
TCAs	29	6	9	32
	13	31		
Air Collision	24	18	25	1
	5	10	31	28
Ramp / Parking		26		
Taxi	10	3	16	13
	20	34	32	24
Take-off	32	27	35	3
Departure	23	2	24	16
Night Flying		12		
Arrival / Scheduling	9	24	22	10
Landing	35	11	30	15
Helicopter	14	21	11	34
Aerobatic	1		28	14
Parachuting				18
Hot Air Balloons	8	16	13	30

Categories	GA1	GA2	GA3	GA4
Gliders		1		
Team			19	27
FAA Inspection	11	14	2	5
Reporter Callback	3	32	14	25

Many of the categories within this taxonomy are composed of multiple document clusters. A ratio of the number of repeated keywords to the total number of keywords represented within that selection of documents was calculated as a measure of the equivalence between the sets collapsed within the category. The following table presents the ratio for each of the categories in which multiple clusters were a part of its construction.

Categories	General Aviation (GA) Sets			
	GA1	GA2	GA3	GA4
Altitude	0.15	0.25		
ILS Approach	0.25			
NOTAMs / TFRs			0.15	
Communication / Radio			0.15	
Restricted Airspace	0.17	0.10	0.17	0.16
Navigation	0.23	0.30	0.05	
TCAs	0.20	0.05		
Air Collision	0.15	0.20	0.15	0.20
Taxi	0.55	0.55	0.55	0.50

Division of the general aviation sets of documents into each of the 12 categories

Categories	GA1	GA2	GA3	GA4
Weather	19	8	15	11
	26	19	33	33
	30	35		35
Calculation / Weight	15	33	23	8
Use of Instruments	4	13	3	7
	7	28	21	17
	18	29	26	22
	22	30	29	26
	27			
	33			
Mechanical Issues	6	7	18	2
	28		20	19
			34	
Teaching	16	9	8	20
Monitoring	34	20	1	9
		25	4	31
Communication		5	12	4
			17	
SA	2	4	5	1
	5	6	6	6
	12	10	7	12
	13	15	9	21
	17	17	10	23
	21	18	25	28
	24	22	27	29
	25	23	31	32
	29	31		
	31			
Context	9	2	16	3
	10	3	22	10
	20	11	24	13
	23	12	30	15
	32	24	32	16
	35	26	35	24
		27		
		34		
Types of Aircraft	1	1	11	14
	8	16	13	18
	14	21	28	30
				34
Interpersonal			19	27
Inspection	11	14	2	5

Similar to the calculations done to explore the similarity of keywords present in the document sets combined to create the 33-catgory solution, a ratio of the number of repeated keywords to the number of total terms was calculated for the 12-category solution. The following table presents these ratios.

Categories	General Aviation (GA) Sets			
	GA1	GA2	GA3	GA4
Weather	0.17	0.07	0.00	0.10
Use of Instruments	0.32	0.20	0.20	0.26
Mechanical Issues	0.10		0.20	0.15
Monitoring		0.20	0.15	0.25
Communication			0.15	
Situational Awareness	0.31	0.26	0.24	0.23
Context	0.48	0.50	0.48	0.44
Type of Aircraft	0.03	0.03	0.03	0.08

References

About ASRS Data. Retrieved November 8, 2006, from http://akama.arc.nasa.gov/ASRSDBOnline/about_data.htm.

ASRS general reporting form. (1994, January). Retrieved September 9, 2008, from http://asrs.arc.nasa.gov/report/electronic.html.

ASRS Program Overview. Retrieved November 8, 2006, from http://asrs.arc.nasa.gov/overview.htm.

Baker, D. P. & Krokos, K. J. (2007). Development and validation of Aviation Causal Contributors for Error Reporting Systems (ACCERS). *Human Factors, 49*, 185-199.

Beaubien, J. M. & Baker, D. P. (2002). A review of selected aviation human factors taxonomies, accident/incident reporting systems, and data reporting tools. *International Journal of Applied Aviation Studies, 2*(2), 11-36.

Berber-Sardinha, T. (1999). Using key words in text analysis: Practical aspects. *Direct Papers 42*, LAEL, Catholic University of Sao Paulo. Retrieved January 8, 2009 from http://www.lexically.net/wordsmith/copus_linguistics_links/papers_using_wordsmith.htm.

Boeing (2006). Statistical summary of commercial jet airplane accidents: Worldwide operations 1959-2005. Retrieved June 22, 2007, from http://www.boeing.com/news/techissues.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391-407.

Dekker, S. (2006). *The field guide to understanding human error.* Burlington, VT: Ashgate.

Diehl, A. (2001). Does CRM really work? In T. Kern (Ed.), *Controlling pilot error: Culture, environment, and CRM* (pp. 33-51). New York, NY: McGraw-Hill.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, and Computers, 23*, 229-236.

Dumais, S. T. (2003). Data-driven approaches to information access. *Cognitive Science, 27*, 491-524.

Elsas, J. L. (2005). An evaluation of projection techniques for document clustering: Latent semantic analysis and independent component analysis. Master's thesis, available for download at: http://etd.ils.unc.edu/dspace/bitstream/1901/208/1/MastersPaperFinal.pdf.

Federal Aviation Administration (2004). *Line operational simulations: Line oriented flight training, special purpose operational training, line operational evaluation.* FAA Advisory Circular 120-35C

Flin, R. & Martin, L. (2001). Behavioral markers for crew resource management: A review of current practice. *The International Journal of Aviation Psychology, 11*, 95-118.

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering theory, algorithms, and applications*. Philadelphia, PA: SIAM.

Glista, T. (2004, July/August). FAA/Industry training standards – an improved general aviation training paradigm. *FAA Aviation News*, pp. 6-8.

Hair, J. F. Jr. & Black, W. C. (2004). Cluster analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 147-205). Washington, D.C.: American Psychological Assoc.

Helmreich, R. L. (1998). Error management as organizational strategy. In *Proceedings of the IATA Human Factors Seminar* (pg. 1-7). Bangkok, Thailand, April 20-22.

Helmreich, R. L., Wilhelm, J. A., Klinect, J. R., & Merritt, A. C. (2001). Culture, error, and crew resource management. In E. Salas, C. A. Bowers, & E. Edens (Eds.), *Improving teamwork in organizations: Applications of resource management training* (pg. 305-331). Mahwah, NJ: Lawrence Erlbaum Assoc.

Johnson, R. A. & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Upper Saddle River, NJ: Prentice-Hall.

Kachigan, S. K. (1986). *Statistical analysis: An interdisciplinary introduction to univariate and multivariate methods*. New York, NY: Radius Press.

Kauffman, L. & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: John Wiley & Sons.

Kayten, P. (1993). The accident investigator's perspective. In E. L. Weiner, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 283-314). San Diego, CA: Academic Press.

Kern, T. (2001). *Controlling pilot error: Culture, environment, and CRM*. New York, NY: McGraw-Hill.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

Kolda, T. G. & O'Leary, D. P. (1998). A semidiscrete matrix decomposition for latent semantic indexing in information retrieval. *ACM Transactions on Information Systems, 16*, 322-346.

Kontostathis, A., Holzman, L. E., & Pottenger, W. M. (2004). Use of term clusters for emerging trend detection. Technical report, available for download at http://webpages.ursinus.edu/akontostathis.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Krokos, K. J. & Baker, D. P. (2005, June). *Development of a Taxonomy of Causal Contributors for Use with ASAP Reporting Systems*. Technical Report, FAA Grant #99-G-048, Washington, DC: American Institute for Research.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Pyscholgoical Review, 104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education, 10* (3), 295-308.

Lauber, J. K. (1993). Foreward. In E. L. Weiner, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. xv-xviii). San Diego, CA: Academic Press.

Leon, S. J. (1998). *Linear algebra with applications, 5th ed.* Upper Saddle River, NJ: Prentice Hall.

Lerman, K. (1999). Document clustering in reduced dimension vector space. Retrieved June 3, 2008, from http://www.isi.edu/~lerman/papers/Lerman99.pdf.

Letsche, T. A. & Berry, M. W. (1997). Large-scale information retrieval with latent semantic indexing. *Information Sciences, 100*, 105-137.

Magliano, J. P. & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction, 21*, 251-283.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge, MA: Cambridge University Press.

Martin, D. I. & Berry, M. W. (2007). Mathematical foundations behind latent semantic analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 35-55). Mahwah, NJ: Lawrence Erlbaum Assoc.

Mathworks (2007). *MATLAB*. Natick, MA: The MathWorks.

Medlock, R (2001). Count *MATLAB* Code. http://webscripts.softpedia.com/script/Scientific-Engineering-Ruby/Mathematics/Matlab-Count-37955.html

Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88*, 1-15.

NTSB (National Transportation Safety Board), 2007. U.S. Air Carrier Operations Calendar Year 2004. Annual review of aircraft accident data, NTSB/ARC-08/01, PB2008-108720.

Parsons, L., Haque, E., & Liu, H. (2004, June), Subspace clustering for high dimensional data: A review. *SIGKDD Explorations, 6*(1), 90-105.

Popping, R. (2000). *Computer-assisted text analysis*. Thousand Oaks, CA: Sage.

Quesada, J. (2007). Creating your own LSA spaces. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 71-85). Mahwah, NJ: Lawrence Erlbaum Assoc.

Reason, J. (1990). *Human error*. New York, NY: Cambridge University Press.

Romesburg, H. C. (2004). *Cluster analysis for researchers.* Logan, UT: Lulu Press.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Salas, E., Wilson, K. A., Burke, C. S., Wightman, D. C., & Howse, W. R. (2006). A checklist for crew resource management training. *Ergonomics in Design, 14* (2), 6-15.

Scott, M. (1997). PC analysis of key words – and key key words. *System, 25*, 233-245.

Scott, M. (2008). *WordSmith Tools* (Version 5.0). [Computer Software]. Liverpool: Lexical Analysis Software.

Shappell, S. A. & Wiegmann, D. A. (1997). A human error approach to accident investigation: The taxonomy of unsafe operations. *International Journal of Aviation Psychology, 7*, 269-291.

Shappell, S. A. & Wiegmann, D. A. (1998, April). *Failure analysis classification system: A human factors approach to accident investigation.* Presented at SAE: Advanced in Aviation Safety Conference and Exposition, Daytona Beach, FL.

Shappell, S. A. & Wiegmann, D. A. (1999). Human factors analysis of aviation accident data: Developing a needs-based, data-driven, safety program. In *Proceedings of the 3rd International Workshop on Human Error, Safety, and Systems Development* [CD-ROM]. Vienna, Austria: International Federation for Information Processing.

Shappell, S. A. & Wiegmann, D. A. (2000, February). *The Human Factors Analysis and Classification System – HFACS*. Final Report, DOT/FAA/AM-00/7. Washington, D.C.: Office of Aviation Medicine.

Sinka, M. P. & Corne, D. W. (2002). A large benchmark dataset for web document clustering. Paper presented at the 2nd Hybrid Intelligence Conference, Santiago, Chile.

Skillicorn, D. (2007). Understanding complex datasets: Data mining with matrix decompositions. Boca Raton, FL: Chapman & Hall/CRC.

Summers, M. M., Ayers, F., Connolly, T., & Robertson, C. (2007). Managing risk through scenario based training, single pilot resource management, and learner centered grading. Retrieved September 4, 2008, from http://www.faa.gov/education_research/training/fits/guidance/media/RM_thorugh_SBT.pdf. [sic]

Willett, P. (1988). Recent trends in hierarchical document clustering: A critical review. *Information Processing & Management, 24*, 577-597.

Wright, R. A. (2004, July/August). General aviation safety in the second century: The flight training challenge. *FAA Aviation News*, pp. 4-5.

Zeimpekis, D. & Gallopoulos, E. (2007). Text to matrix generator (Version 4) [Computer Toolbox]. Available from http://scgroup6.ceid.upatras.gr:8000/wiki/index.php/Main_Page.