Application of Provalis Research Corp.'s Statistical Content Analysis Text Mining to Airline Safety Reports

A Technology Demonstration of the SimStat/WordStat Software by Provalis Research Corporation at JetBlue Airways

In Partnership with the Federal Aviation Administration and the Global Aviation Information Network (GAIN)

Report Prepared by:

Dr. Normand Péladeau
President
Provalis Research

Mr. Craig Stovall
Safety Data Analyst
JetBlue Airways

February 2005

jetBlue AIRWAYS

Provalis Research

Disclaimers; Non-Endorsement

All data and information in this document are provided "as is," without any expressed or implied warranty of any kind, including as to the accuracy, completeness, currentness, non-infringement, merchantability, or fitness for any purpose.

The views and opinions expressed in this document do not necessarily reflect those of the Global Aviation Information Network or any of its participants, except as expressly indicated.

Reference in this document to any commercial product, process, or service by trade name, trademark, servicemark, manufacturer, or otherwise, does not constitute or imply any endorsement or recommendation by the Global Aviation Information Network or any of its participants of the product, process, or service.

Notice of Right to Copy

For Further Information on this Technology Demonstration

Dr. Normand Péladeau
President
Provalis Research
2414 Bennett Avenue
Montreal, QC H1V 3S4 CANADA
+1 514-899-1672
peladeau@simstat.com
www.simstat.com

Mr. Andy Muir
GAIN Program Office
Federal Aviation Administration / FSAIC
800 Independence Avenue, SW
Washington, DC 20591 USA
+1-202-267-9180
andy.muir@faa.gov
www.gainweb.org

Mr. Craig Stovall
Safety Data Analyst
JetBlue Airways
118-29 Queens Boulevard
Forest Hills, NY 11375 USA
+1-718-709-3040
craig.stovall@jetblue.com
www.jetblue.com

Table of Contents

Acknowledgements

This project was funded by the US Federal Aviation Administration, Office of System Safety (known as "ASY" within the FAA) to facilitate the application of advanced methods and tools in the analysis of aviation safety data with the goal of improving aviation safety industry-wide. The project also involved the support and guidance of the Global Aviation Information Network (GAIN), an industry-led international coalition of airlines, manufacturers, employee groups, governments (including FAA) and other aviation organizations formed to promote and facilitate the voluntary collection and sharing of safety information by and among users in the international aviation community to improve aviation safety. Specifically, the project was guided by principles developed by GAIN's Working Group B, "Analytical Methods and Tools" (WG B), which has been tasked with fostering the use of existing analytical methods and tools and the development of new tools that elicit safety information out of aviation data.

The report authors acknowledge that this project could not have been completed without the valuable support of: a number of JetBlue flight safety crew members including Vince Zaccardi, Randy King and Craig Hoskins; associates of Provalis Research Corporation including Jean-Philippe Raymond, Jean-François Allie, and Daniel Charland; Dominic Forest, from Yago Technologies; and Andy Muir, from the FAA's Office of System Safety (ASY) and a member of GAIN WG B; as well as all the individual members of WG B. Without the active support, advice, guidance and direction of these individuals this project could not have been accomplished.

Executive Summary

This technology demonstration applied an advanced analysis technique called "statistical content analysis" by Provalis Research Corporation to airline safety data at JetBlue Airways. The Provalis software products used for this project are known as WordStat and SimStat and involve a variety of text and data mining routines. "Statistical content analysis" involves the combination of custom categorization dictionaries to group words and phrases under content categories and statistical as well as visualization tools like cluster analysis, multidimensional scaling, correspondence analysis and heatmaps.

Input data for this project consisted of sample data from two JetBlue safety databases: Traffic Collision Avoidance System (TCAS) Reports (about 200 reports) and a selected portion of Flight Crew Irregularity Reports (FCIR--about 580 reports). Both of these reports involve a narrative event description and numerous fields of coded or categorical information regarding the event.

Based on the limited analysis available during this technology demonstration, flight safety crew members at JetBlue Airways believe the statistical content analysis routines to be useful in enhancing the current analysis of airline safety reports. This report summarizes the findings as follows:

1. Current safety report analysis methods provide useful analysis information that has improved management of flight safety, operations, and training programs. Those methods typically involve analyzing events put into categories or coded by occurrence type. Depending on the database structure analyst may be limited to the number of categories or types that events may be coded. Text mining techniques show promise in greatly improving efficiencies in the current methods by automating the query and output process that may be missed or not categorized or coded due to database limitations.

2. Text analysis software has often relied on two different techniques called stemming and lemmatization to aggregate inflected forms of the same noun or verb into a common lemma or root, and thus reduces the total number of linguistic units to process. The current project allowed the identification of potential benefits and inconveniences of these two techniques when applied to aviation safety data and yielded to specific recommendations as to possible ways to reduce those inconveniences.

3. The development of dictionaries is instrumental to aid the analysis of the information in documents to specific topics of interest, factors that are known to be related directly or indirectly to safety issues. This can be an intensive task that often requires many months of heedful work, but once developed such a taxonomy of abbreviations, idioms and phrases may be applied to any text mining project related to the same domain and be very useful for document indexing and categorization as well as document retrieval. A detailed description of the proper steps required to develop such a dictionary and to validate it has been presented. A preliminary version of an aviation dictionary has also been created. This dictionary includes over 3100 technical terms, abbreviations, and common phrases organized into a hierarchical taxonomy of nearly 400 categories and subcategories.

4. WordStat offers two broad strategies or categories of text mining and knowledge discovery tools. The first one is based on the analysis of co-occurrences of words or keywords within reports and includes statistical and visualization techniques such as hierarchical cluster analysis, proximity plots and multidimensional scaling. The first two techniques have been successfully applied in this project. Dendrograms produced by cluster analysis were found to be direct and easy to interpret and likely to support discovery of unexpected relationship. Proximity plots were especially useful to establish profiles of airports and identify differences as to the nature of TCAS events occurring at those airports.

5. The second broad type of strategies looks at potential relationship between categories of words and values of numerical and categorical variables. It includes crosstabulation, heatmaps with dual clustering, and two-dimensional and three-dimensional correspondence plots. While the wealth of numerical information provided by crosstabulations was found to be overwhelming, heatmaps proved to be a visualization technique easily understood and useful to extract relevant information from those large tables.

6. This technology demonstration has led both Provalis Research and JetBlue flight safety team to identify limits of some tools and to make specific recommendations to further improve their usefulness. Examples of those suggested improvements include the development and testing of better ways to measure co-occurrences, the implementation of text retrieval features closely integrated with visualization tools and the integration of automatic document classification techniques.

It is likely, based on this project, that similar text mining analysis techniques can be effectively applied to information derived from other safety reports in an aviation organization, such as operational delay reports, maintenance logbook entries and maintenance cancellation and delay reports, quality assurance reports, system safety assessments, and internal evaluation program reports, to deliver a more comprehensive picture of overall safety issues.

1.0 Introduction

Airline flight safety offices (FSOs) and other aviation organizations typically collect, integrate, and analyze safety reports using a combination of manual and automated methods. Pilots, cabin crew, ground crew, maintenance personnel, and others involved in aircraft operations are encouraged to report safety incidents, events or hazardous situations or other safety concerns to the airline. Such information may have various report forms and various databases in which the information is stored, depending on the type of issue being documented and the type of airline employee reporting the information. These safety reports can be called air safety reports/incident reports, human factors reports, technical reports or other names at various airlines. In the United States, through a program termed the Aviation Safety Action Program (ASAP), occurrence reports that meet certain requirements are shared with the Federal Aviation Administration, which in turn has agreed not to impose regulatory penalties on either the airline or the personnel filing the report, if the report is accepted into the program.

The collected safety information is often stored in a database, with many fields of coded or "structured" information, such as aircraft data, phase-of-flight, location, type of event, and so on. Most such reports also have a narrative describing the safety event or concern. These narratives may consist of a short paragraph or a page or more of text, depending on the nature and complexity of the issue. The narratives are often called "free form" information and can vary significantly within the database as the various reporters may include varying levels of detail and their own terms and phrases.

Other capabilities that are provided by some systems include functions to support the classification of events into predefined categories, to assign risk levels to each event, and to filter the information in the event report database to identify subsets of previous reports that have common characteristics and extract relevant information. These capabilities are fundamental to effective safety management, since they allow flight safety personnel to identify areas of significant risk and track the long-term effectiveness of corrective actions. Having an effective event classification system is essential to be able to perform meaningful trend analysis and information filtering. Risk assessment of each event allows flight safety management personnel to identify those incidents that pose the most serious threat to operational safety and to focus appropriate attention on high-risk events.

The extent to which these systems have built-in capabilities to perform trend analysis, generate charts and graphs, or perform other statistical analysis varies. However, most such systems have limited analytical capabilities beyond fairly simple trend analysis. The analytical process often relies heavily on the FSO or other analyst's expertise and memory to identify anomalies and trends, rather than the use of automated tools.

1.1 GAIN Interest in Text Mining Algorithms

The Global Aviation Information Network (GAIN) is an industry-led international coalition of airlines, manufacturers, employee groups, governments and other aviation organizations formed to promote and facilitate the voluntary collection and sharing of safety information by and

among users in the international aviation community to improve aviation safety. GAIN's Working Group B, "Analytical Methods and Tools" (WG B) has been tasked with fostering the use of existing analytical methods and tools and the development of new tools that elicit safety information out of aviation data. WG B's mission includes a search for analysis tools developed to support other industries that could be applied to aviation data to increase the value of data currently collected, as well as support to the development of tools specifically targeted toward aviation safety data.

One of the most interesting developments in information management is the proliferation of "data mining" and "text mining" analysis tools. One definition for data mining is "the process of discovering hidden patterns and relationships in data." Text mining involves the application of data mining techniques to narrative or textual information. Emphasis placed on automated learning from data, as the mining tools find interest patterns without humans asking the initial queries. However, subject matter expertise from a human is always needed to review the initial results from a text or data mining project, to determine which of the patterns and relationships are of real value, and which are "commonly understood" (e.g., increased stopping distance in wet conditions) or otherwise not valuable.

Text and data mining includes various types of analysis, such as the following:

Classification	*Predicting a category for an example*
Regression	*Predicting a numeric value for an example*
Cluster Analysis	*Grouping examples with similar characteristics*
Association/Sequence Discovery	*Finding examples that occur together frequently*
Anomaly Detection	*Identifying unusual example*
Exemplar-Based Learning	*Classifying based on similarity to previous*

WG B believes that airlines can learn significantly more about their operations by expanding the types of analyses provided by their current safety data analysis programs. If text mining algorithms could be applied to narrative data, safety analysts might identify patterns and trends in the operation of their aircraft that were not previously known. In addition, text mining tools are designed to increase the level of automation in an analysis project, and therefore decrease the amount of time an analyst must invest in a particular project.

Traditional analysis, based on utilizing coded data and basic searches for specific text strings, has known limitations and complications for the safety analyst. For example, event coding often focuses on event outcomes or a primary occurrence. If a hard landing results in an injury to a flight attendant, the event code might be "FA Injury." A search for all hard landing events using the event code would miss this event. An additional problem is that various terms can be used in

the narratives submitted by reporters of the same type of events. A report that includes the sentence "We smacked down onto the runway" would not be found in a text search query for "hard landing" events. Both of these examples are false negatives, when events of interest are not returned by particular searches. A false positive (a result that does not meet the intended criteria) can arise from basic text searching if a report includes the search terms, but outside the intended meaning ("The pilot recovered quickly enough to avoid a hard landing."). WG B hopes that advanced text mining tools will overcome those types of limitations in current analysis.

1.2 JetBlue Airways Corporation

JetBlue Airways is a low-fare, low-cost passenger airline that operates a fleet of 73 new Airbus A320 aircraft and plans to add 11 additional A320s and 7 Embraer E190s to its fleet in 2005. Based at New York City's John F. Kennedy International Airport, JetBlue currently operates 295 flights a day and serves 29 destinations in 12 states, Puerto Rico, the Dominican Republic and The Bahamas.

1.3 Provalis Research Corporation

Provalis Research Corporation is a software company founded in 1989 and devoted to the design and development of data analysis software tools for numerical and textual data. One of the distinctive features of company's products is the integration of quantitative and qualitative research methods as well as the application of data analysis tools and visualization techniques in various domains including social sciences, statistics, linguistics and biomedical sciences. Provalis Research provides tools and services to customers worldwide, including companies working in the areas of surveys and market research, pharmaceutics, computer equipment, and other manufacturing companies as well as to various governmental services.

1.4 WordStat and Simstat

WordStat is a text analysis module specifically designed to study textual information such as responses to open-ended questions, interviews, titles, journal articles, public speeches, electronic communications, technical reports, etc. WordStat may be used for statistical content analysis of text using a dictionary based approach and for text mining. It also provides numerous tools for the iterative development and validation of general or specialized categorization dictionaries or taxonomies. WordStat includes numerous exploratory data analysis and graphical tools that may be used to explore the relationship between the content of documents and information stored in categorical or numeric variables, such as product types, dates, age, etc. Relationships among words or categories, as well as document similarity may be identified using hierarchical clustering and multidimensional scaling analysis. Other analysis techniques, such as correspondence analysis and heatmap plots, may be used to explore relationship between keywords and different subgroups of documents or individuals. WordStat is not a standalone application but a module that must be run from Simstat, a statistical data analysis software, or QDA Miner, a text management and qualitative coding software. All three programs share the

same data file format allowing one to combine in a single database numerical, categorical, date and logical fields as well as full documents stored either as plain text or in Rich Text format.

Simstat provides a wide range of statistical procedures for the analysis of quantitative data. It can import data stored in various database and spreadsheet formats (MS Access, Paradox, Excel, Lotus, SPSS, etc.) as well as any other major commercial database through an ODBC (Open Database Connectivity) or OLE DB connection. A special document conversion wizard also allows one to import text from numerous standard formats (ASCII, RTF, MS Word, Word Perfect, HTML, etc.). It offers advanced data file management tools such as the ability to merge data files, aggregate cases, perform complex computation and transformation.

When used with Simstat, WordStat can analyze textual information stored in any alphanumeric, plain text and rich text memo field (or variable). It includes various tools to explore the relationship between any numeric field of a data file, and the content of alphanumeric ones. Its close integration with SimStat facilitates further quantitative analysis on numerical results obtained from the content analysis (e.g., factor analysis, multiple regression, etc.).

1.5 Project Objectives

The main goal of this technology demonstration is to help the airline safety community, by the means of text mining techniques, move closer to a future goal of wide-spread application of advanced analysis tools to aviation safety data.

The **first objective** is to examine various statistical and graphical content analysis tools found in WordStat and assess their usefulness in their application to text mining of aviation safety data.

This assessment will focus on two broad classes of statistical and graphic techniques: 1) tools based on the analysis of word co-occurrences such as hierarchical cluster analysis and proximity plots, and 2) tools analyzing the relationship between words or themes and a categorical or numerical variable, especially correspondence analysis plots and heatmaps.

The report identifies for each techniques strengths and limitations, describes required conditions for their successful use, as well as any implied assumption. Some potential applications for aviation safety issues are also described.

The **second project objective** is to help the text mining community improve and adapt their tools for specific application to the airline safety environment. This project documents JetBlue's perceptions about the tool and list suggestions for how it could be improved, that is, what needs to be fixed or speed up or made more understandable or automated for the use by the airline safety industry. JetBlue will have the opportunity to recommend features to be added that might improve its usefulness to airline safety.

The **third project objective** is to support future application of text mining to aviation safety reports by building an initial thesaurus of aviation safety terms that will be made available to the industry (i.e., other tool vendors and other aviation groups). The thesaurus will spell out

abbreviations and acronyms, document synonyms for particular terms, and present a hierarchy of related terms. The thesaurus will be based largely on terms retrieved from text pulled from JetBlue Airways' data used in this project and will not be comprehensive across all aviation terms (e.g., all aircraft parts) or all aviation safety concepts (e.g., human factors issues for piloting aircraft).

2.0 Data Preparation

In most data mining projects, a large portion of the time is spent preparing the data for analysis, checking its consistency and validity and distributional properties of numerical and categorical variables. When the mining effort involves the analysis of textual data, additional times is usually spent on the development and testing of a text processing strategy which may involve the spell-checking of the original documents, lemmatization and stemming of words as well as text categorization. This section provides a description of the various preparation tasks performed during this project. It also includes a critical appraisal of common text pre-processing methods often used when performing content analysis or text mining tasks with regards to the analysis of aviation safety data.

2.1 Description of Input Data

Input data for this project consisted of sample data -- from late 2001 through July 2004 -- from two JetBlue safety databases: Flight Crew Irregularity Reports (FCIR--about 580 reports) and Traffic Collision Avoidance System (TCAS) Reports (about 200 reports).

Flight Crew Irregularity Report (FCIR) – This type of report is used anytime a pilot experiences an incident or abnormal operational event. Specific types of incidents and events are outlined in the Flight Operations Manual. This type of report is not confidential and information is shared with appropriate departments for review and follow-up actions.

Traffic Collision Avoidance System report (TCAS) – This type of report is used anytime a pilot experiences a TCAS occurrence. Flight Crew reporting of this information allows the Safety Department and Manager Air Traffic Programs to coordinate with appropriate Air Traffic Control (ATC) facilities to resolve airspace issues.

Each data file includes a text field containing the full event description as well as numerous additional fields, the most common ones being the author of the report, the date and time of the event, the departure and destination airport, the airplane and flight number and a short text field containing codes assigned by a JetBlue flight safety crew member to categorize the event that took place.

While it would have been possible for Simstat to directly import the original SQL data files, the data were exported instead to MS Excel allowing a JetBlue flight safety crew member to easily remove any sensitive data and perform basic data transformation. Excel files were then directly imported from within Simstat.

After further discussion with JetBlue flight safety crew members, it was decided to limit the data to be analyzed for the purpose of this project to about 200 Traffic Collision Avoidance System Reports (TCASR) and 580 Flight Crew Irregularity Reports (FCIR).

2.2 Data Cleansing and Transformation

In any data mining task, the first step, and often the most time consuming one, is to prepare the data for the analysis. Thus, in this project, several transformations and recoding processes were necessary.

2.2.1 Correction of typographical errors and data entry inconsistencies

Typographic errors are common in large databases. In aviation safety databases, this can take the form of misspelled airport codes or aircraft numbers, invalid dates, or aberrant values such as an altitude of 100,000 feet in a data field containing the altitude of the aircraft. Logical inconsistencies in information from different fields may also be found, such as a reported altitude of zero for an event that occurred during the cruise phase or a report mentioning a problem during pushback and specifying an altitude of 28,000 feet. Also, data entry conventions may not be shared by all people doing data entry or may vary over time. For example, in a field indicating the time of the critical event, the entered value can be expressed using either local or GMT (universal) time. All those errors and inconsistencies need to be fixed prior to any analysis involving those structured fields.

Blank cells in databases are especially problematic since it is not always easy to decide whether they are used to indicate an absence, a lack of information or the fact that such information was not relevant.

For the current project, data cleansing was restricted to the fields that were actually used in analyses, such as airport codes, dates and times. Time of occurrence of TCAS incidents were also transformed into the local time of the airport where the event occurred. Because of time constraints, all fields with a significant proportion of missing values were ignored.

2.2.2 Numeric recoding of alphanumeric variables

WordStat was designed to analyze textual information stored in alphanumeric fields and identify potential relationship between those text fields and any other numeric variable in the data table. For this reason, short alphanumeric strings containing categorical information, such as the code for the departure or destination airports, the aircraft number or the categories of events assigned by JetBlue flight safety crew members to describe the events, were all transformed into numerical variables.

2.2.3 Recoding of categorical values

No matter whether data mining is performed using machine learning heuristics or statistical algorithms, when one wants to perform comparison among several groups, it is essential that the

number of examples for each group (or class of a categorical variable) be large enough to insure that the information obtained for this subgroup is reliable and representative. If only a few reports are available for a specific subgroup, then any descriptive or inferential statistic computed on such a class is at risk of being unreliable. For example, if only a few TCAS events occurred at a specific airport, then it would be hazardous to rely on the description of this limited number of reports to characterize the nature of the collision risks at this airport. For this reason, frequency distributions of the numeric variables were obtained in order to identify which variables (or values of those variables) could be used in further analysis. Classes of categorical variables with low frequency were either recoded as missing or grouped under broader categories. For example, because of the small quantity of TCAS reports occurring at some airports, the number of classes for this variable had to be reduced from the original 19 airports to only four. All remaining airports were lumped together in the category "OTHERS". For the same reason, it was decided to group TCAS events occurring during the takeoff and the climb phases into a single category, as well as those occurring during approach and landing. Again, such grouping is justified by the low number of TCAS events occurring during either takeoff or landing. Another solution would have been to simply remove those infrequent events from any analysis involving comparison between phases of flight at the risk of loosing potentially valuable information.

2.2.4 Spell-checking of reports

No spell-checking was made on the original reports. While such a preliminary cleaning is often performed prior to text mining, it was decided to identify instead the most common spelling errors and consider those errors in the development of the aviation safety thesaurus. Developing a thesaurus that would include the most common misspelling would allow one to relax the condition of its application and allow its use on original reports, proof-read or not.

2.3 Lemmatization and Stemming

English is an inflectional language where a single word (or lemma) may be written in several inflected forms. For example, the verb "to talk" may appear in reports as "talk", "talks", "talked" or "talking". While a native speaker has no difficulty establishing the correspondence between plural and singular forms of the same noun, or between various inflected forms of a single verb, computers will typically treat all those word forms as distinct entities. To alleviate the problems that may result from such a situation, various techniques have been used to aggregate inflected forms into a common lemma or root, and thus reduce the total number of linguistic units to process. Text analysis software has often relied on two different techniques: stemming and lemmatization.

2.3.1 Stemming

Stemming is a well known technique of form reduction by which common suffix and sometimes prefix are stripped from the original word form. For example, a stemming algorithm will remove the final "s" from the word form "areas". It will also successfully treat "believe", "believing",

"believe" and "believed" as a single linguistic unit by transforming all those words into the root word "believe".

While this technique has been found to be useful for many types of natural language processing applications, such as indexing, document retrieval or automatic categorization, we decided not to us stemming in this project and are also reluctant to recommend its use for statistical content analysis of aviation safety data for many reasons. First, stemming is an aggressive technique. Stemmed forms are often not valid words but word roots, making it more difficult - if not impossible - for aviation safety experts to relate them to any known word, and thus unnecessarily complexity the interpretation of results. For example, it may not be obvious that "PAG" and "IC" are stemmed forms of "page" and "paging" and "ice" or "iced" respectively.

Also, stemming will sometimes reduce words with different meanings such as "negligible" and "negligent", "ignore" and "ignorant", to the same root. Other examples of improper stemming especially troublesome for the analysis of aviation safety data are the reduction of words such as "terminal" and "terminated" to "termin", of "arrival", "arrived" and "arriving" to "arrive", and of "indicator", "indication", "indicating" to "indic".

2.3.2 Lemmatization

Lemmatization is another form reduction process by which inflected forms are reduced to their canonical form (lemma or dictionary entry form). Usually, lemmatization attempts to preserve the syntactic categorization of each word. For example, verb forms will be reduced to their infinitive; inflected forms of nouns will be transformed into their singular form. One benefit of lemmatization over stemming is that it relies on a lexicon and thus always returns valid words.

Several methods of lemmatization have been explored. The lemmatization performed in WordStat 4.0 is based on a set of suffix substitution rules (Krovetz, 1993), similar to those used for stemming, but in this case, the substitution process is moderated by a dictionary to insure valid words are returned. Another common method of lemmatization consists in a comprehensive list of inflected words and their corresponding canonical form. Such a list can be applied with or without a prior part-of-speech tagging. While we estimate the accuracy of the suffix substitution algorithm to be quite high, two problems may result from such an approach. Like all lemmatization routines, this algorithm is not 100% accurate. Invalid lemmatization may still occur, some of which could be problematic for the analysis of aviation safety data. For example, we found that the lemmatization algorithm mistakenly changed the word "gash" to "gas". It also reduced some valid abbreviations like "MSL" (i.e. "mean sea level") and "MSG" (i.e. "message") to "ms".

The second problem with this approach is the possibility that a lemmatization, while potentially valid from a linguistic point of view, may be semantically incorrect. A very good example is the substitution of the word "ground" (used as a noun) commonly found in aviation safety reports with the infinite verb "grind". From a linguistic perspective, the word "ground" can be used as the passive form of the verb "to grind", yet in the context of safety reports, it is seldom used this way. While a lemmatization algorithm based on a preliminary part-of-speech tagging may

prevent such kind of error, we believe the computing requirements to obtain such a greater accuracy would considerably slow down the text mining process to an unacceptable level. Our own speed tests suggest that part-of-speech tagging would slow down the text mining process by a factor of maybe 1000 times. As a result, an analysis that would normally take 5 or 10 seconds to complete could take several hours. While part-of-speech tagging may be useful, even essential, for some tasks, its slow processing time represent a major inconvenience when doing text mining in real time on large collections of documents.

An additional problem that may occur from lemmatization is that the resulting root form may be more ambiguous that the original inflected form. For example, lemmatizing "training" and "trained" down to "train" introduce a potential confusion between the learning activity, the wheelwork, and the public transportation mode. While such a confusion already exists for the verb form "train", it is absent for "training" or "trained". In the context of aviation safety, we believe the reduction of word forms may result in some important loss of information. Here are a few examples of possible confusions that may result from lemmatization:

"braking" changed to "brake" (as in "parking brake")
"controlling" changed to "control" (as in "ground control" or "fuel control")
"approaching" changed to "approach" (flight phase)
"fueling" changed to "fuel"
"smoking" changed to "smoke"

This last example is especially instructive. In the limited context of the JetBlue aviation safety reports, it was found that the word "smoking" was always used to refer to the act of smoking tobacco or other substances. Lemmatizing this verb form to "smoke" would make the differentiation of "smoke" used as a verb and smoke to refer to fumes more difficult.

When applied to the JetBlue safety reports, lemmatization reduced the total number of word forms to process by a percentage between 20% and 25%. Such a reduction can greatly facilitate the text analysis and the development of content analysis dictionaries by automatically taking care of inflected forms of verbs and nouns.

It was decided for this project not to rely on this pre-processing method because of potential errors and enhanced ambiguity that may result from lemmatization. Instead, the project team developed a categorization dictionary to take into account the potential presence of those inflected forms. This approach has the additional benefit of allowing such a dictionary to be used by any text analysis software, some of which may not have lemmatization routines. Would someone choose to perform lemmatization, Provalis Research would strongly recommend making sure the software is able to provide to the user a list of all the substitutions made during the lemmatization process and offer him the possibility to override some of those changes in order to correct inaccuracies or prevent lemmatization that would result in greater ambiguity. While such a feature is not available in the current version of WordStat, it will be added to the next major revision.

2.4 Development of Dictionaries

While WordStat text mining capabilities may be used to identify patterns among all words found in safety reports as well as trends and relationships with structured fields in the database, it is often preferable to restrict the analysis of the information in documents to specific topics of interest, factors that are known to be related directly or indirectly to safety issues. Such a focus can be achieved by 1) the removal of words that provide no relevant information and 2) the selection and categorization of specific terms, phrases or abbreviations that are associated with specific concerns and questions expressed by aviation safety experts. The development and assessment of categorization dictionaries or thesaurus are intensive tasks that often requires many months of heedful work. However, once developed such a taxonomy of abbreviations, idioms and phrases may be applied to any text mining project related to the same domain and be very useful for document indexing and categorization as well as document retrieval. This section introduces some considerations in the use and development of both an exclusion list and categorization dictionaries, based on our experience with the analysis of JetBlue safety reports.

2.4.1 Exclusion list

In content analysis or text mining project, lists of exclusion words are often used to filter the initial lexicon found in the databases to be analyzed. The main objectives of this process are: 1) to reduce significantly the size of the lexicon; and, therefore reduce the processing time and 2) to retain only the most relevant or informative words. Such an exclusion list contains common words that may modify other words but carry no inherent meaning themselves, such as adverbs, conjunctions, propositions or forms of "be" or "have". These words are commonly referred as "stop words", "trivial words" or "function words" and typically include items like "and", "before", "after", "maybe", "very", "too", etc.

WordStat, like most other text mining or content analysis software, comes with such a predefined list of about 550 words. However, very often, the aim of the analysis will require that some modifications be made to such a list. While working on JetBlue's data it quickly became clear that the analysis of aviation safety data would require some adjustments to the exclusion lists. Technical documents, such as aviation safety documents, are very often composed of technical terms and acronyms that can be found in typical English exclusion lists. For example, the preposition "TO" is a known acronym for "take-off", while the preposition "AT" may also be used to refer to "Air Traffic". Another good example encountered in several JetBlue databases is the acronym "FAR" which is used to refer to Federal Aviation Regulation.

According to the objective and nature of the analysis, it can also be essential to manually remove specific words typically found in exclusion lists. For example, in one attempt to measure the urgency of a situation, then words like "quickly", "suddenly", or "immediate" may become very relevant. If one tries to assess the role of human factors in safety incidents including typical dimensions such as the lack of training, the presence of confusion, indecision or misunderstanding, then analyzing the usage of words like "maybe", "seems", "probably", "perhaps", "whether" may become appropriate.

To prevent the loss of any information caused by exclusion of significant words, it was thus decided to drastically reduced the number of excluded words and restrict such a list to about 40 words.

2.4.2 Categorization dictionary - An Introduction

Creating a comprehensive and valid categorization dictionary for a specific domain such as aviation safety is a time-intensive activity that require several months if not some years to achieve. One not only has to select relevant vocabularies including words, phrases and abbreviations, but also group them into specific concepts as well as broader categories. Part of the difficulty lies in the extensive variability of human language that allows people to express the same idea in many different ways. If some mechanical devices are usually mentioned in reports using a limited number of terms, expressions or abbreviations, other concepts such as specific actions performed by flying pilots may be described in numerous ways. For example, in a small database of 179 TCAS reports, 44 of those reports mentioned the autopilot, most of them referring to the specific action of turning it off. No less than 26 different ways of referring to this action were found, including phrases like:

```
disconnected the autopilot
disconnecting A/P
turned off autopilot
turned the autopilot off
turning off AP
disengaged the auto pilot
kicked off the autopilot
autopilot was disconnected
```

The problem became even more important when attempting to measures more abstract concepts such as fatigue, confusion, stress or any other issues related to known human factors in aviation safety problems.

While the development of a comprehensive dictionary that could readily be used for text mining aviation safety data goes well beyond the limited time available in this project, one of the objectives was to create a preliminary version of such a dictionary and to document the building process. WordStat provides several tools to assist the development of such taxonomy. The remaining part of this section will describe some of those features and how they had been used to build such a dictionary.

Step #1 - Identification of abbreviations and technical terms

The first step performed was the identification of technical terms and abbreviations used in aviation safety reports. This was performed by comparing the list of all word forms in the TCAS and Flight Operation Safety Reports against a list of common English words and retrieving all those that were not part of this latter list as well as common words with irregular capitalizations[1]. While some of the retrieved words were spelling errors, most of them were either airport codes (e.g., `LGB, JFK`), intersection names (e.g., `DAWNA, DEKAL, CAMRN`), abbreviations or acronyms

[1] This feature will be part of the next version of WordStat 5.0 expected to be released in the second quarter of 2005.

(e.g., `TCAS`, `RA`, `RWY`, `IVSI`, `Capt`) or technical terms (e.g., `glidepath`, `overspeed`, `jetbridge`, `pushback`, `towbar`, `airstairs`). All those items where collected for further categorization.

Step #2 - Identification of common phrases

The second step was to use WordStat phrases finder feature to identify the most frequent word sequences in the safety reports and select from those the ones that could be considered phases specific to the aviation vocabulary and included in a categorization dictionary. Two reasons justify looking at the most frequent phrases rather than starting with the individual words. First, many frequent word sequences are phrases or idioms that are specific to a domain and whose meanings cannot be inferred from the meanings of the words that make them up. They should preferably be treated as single concepts. A good example is "ground control". A second related reason is that while single words may have many different meanings, phrases and idioms are much more specific and, most of the time, refer unequivocally to a single concept. For example, while the single word "landing" may be used either as a noun or a verb and may convey many different meanings even in the limited context of aviation safety reports, its inclusion in phrases like "landing gear", "landing rollout", "approach and landing" or "overweight landing" allows one to better differentiate its different meanings and measure each of those concepts more precisely.

It is also important to remember that when WordStat processes words it encountered in documents, it treats phrases and idioms specified in the categorization dictionary as whole entities, and do not process individual words that are part of it. In the above example, if the single word "landing" is part of the dictionary as well as all the above phrases, it will collect frequency information on "landing" only when it is not part of one of those phrases. This processing rule is especially useful when one wants to perform disambiguation and restrict the measure of a specific word to a limited number of meanings.

Applying the phrases finder feature to the 573 Flight Operation Safety Reports returned a list of more than 830 phrases used at least on three occasions. The most frequent of those phrases were:

```
FLIGHT ATTENDANT
GROUND CREW
BIRD STRIKE
RETURN TO THE GATE
MINUTES LATE
INFLIGHT CREW
RETURN TO JFK
MAINTENANCE CONTROL
MX CONTROL
ECAM ACTION
OVERWEIGHT LANDING
GATE AGENT
APPROACH AND LANDING
GROUND CONTROL
```

Step #3 - Initial categorization

This list of common phrases as well as extracted terms and abbreviations were used to develop a first taxonomy of aviation safety terms. Phrases and abbreviations that meant the exact same thing or were closely related were first grouped together in narrow categories. It then became apparent that all those items could be classified into broad logical categories referring to either specific events or outcomes or (e.g., "bird strike", "overweight landing"), actors (e.g., "flight attendant", "maintenance control"), actions (e.g., "return to the gate") or devices and equipment (e.g., "parking brake", "auto pilot").

One of the objectives sustaining the development of the current categorization dictionary was an attempt to measure not only the occurrence of specific events but how the cockpit crew reacted to the events. An initial effort was made to measure not only specific actions but also the contribution of specific human factors associated with aviation safety problems. It was thus decided to create specific categories of words and phrases that could indicate the presence of those factors, including items like confusion or misunderstanding, level of awareness or any other cognitive or emotional state. An initial attempt was also made to measure the nature of interpersonal relations and contextual factors, hoping that such personal or situational factors could be potentially linked to specific events, actions or devices. The development of such kind of content categories is much more difficult and subjective than the creation of taxonomies of technical terms and also requires a careful assessment of their validity in a variety of context. While such a task could not be realistically accomplished in limited time we had, it could be nevertheless provide an illustration of another useful application of categorization dictionaries.

Step #4 - Use of Integrated Thesaurus

While the development of a categorization dictionary should initially be based on the categorization of words and phrases encountered in a specific text corpus such as the JetBlue safety reports, relying on a single source of data may result in a dictionary that cannot adequately be used outside this context. Since the content of the dictionary is limited to what has been encountered in the initial corpus, applying it to either safety reports from other airlines companies or even to new reports from JetBlue crew members may result in an underestimation of the categories that the dictionary attempt to measure. Such a situation occurs because people will likely express similar ideas in different ways with different words.

One way to insure the generalization of the categorization system would be to apply it to another collection of safety reports and identify uncategorized words, abbreviations, and phrases that should be added to the existing dictionary. The more numerous and diverse would be the source of those safety reports, the more likely the dictionary will have a generic value and could be used for the analysis of similar data in a wide variety of context. However, the availability of textual data from numerous sources is often difficult, and in the context of the actual project, inconceivable. Another way to increase the comprehensiveness of a categorization dictionary is to attempt to imagine various alternate ways of expressing the same ideas. This is a difficult, time-consuming and subjective task. However, WordStat can assist in finding words that may be related to existing categories by the use of three lexical tools:

- An English language dictionary is used to propose inflected forms of existing words already in the dictionary.

- A thesaurus is also used to propose synonyms of words already in the dictionary. The thesaurus data file contains over 8,700 context topics (indexed words) and offers a database of over 75,000 synonyms.

- A WordNet based lexical database is used to find synonyms, antonyms as well as hypernyms, hyponyms, coordinate terms, homonyms, metonyms, etc. This database contains over 120,000 root words and offers over 100,000 synonym sets. The availability of word sense definitions allows for manual as well as automatic identification and filtering of proper word senses.

While the usefulness of such semantic tools is rather limited when one categorization dictionary consists of technical terms from the aviation vocabulary (aircraft parts and devices, pilot actions, etc.), they can be quite useful when dealing with more common words like those used to measure human factors. For example, one of the initial category attempted to measure a cognitive states of uncertainty or confusion with words like `AMBIGUOUS`, `CONFUSED`, `DOUBTFUL` and `UNCERTAIN`. Using the integrated thesaurus, WordStat suggested items like `BLURRED`, `HESITANT`, `UNCERTAIN`, `MIXED UP`, `PUZZLED`, etc. Looking at inflected forms also returned `AMBIGUOUSLY` and `AMBIGUOUSNESS`, `CONFUSING`, `CONFUSEDLY`, `CONFUSION` and many other words of the same family as those already in this category.

As an example of how those tools can be useful to develop comprehensive dictionaries, starting with only the above initial four words in the confusion category, WordStat basic dictionary building feature suggested 26 synonyms and 13 inflected forms for possible inclusion in this dictionary, most of them relevant. The use of the more advanced lexical database led to a total of 143 suggestions. After only a few iterations of adding relevant words and obtaining additional suggestions, it is possible to construct a content category that is quite comprehensive.

Step #5 - Validation of dictionary entries

An important aspect in the construction of dictionaries that should not be neglected is its validation. Validating entries in a dictionary is usually performed by looking at all instances of a specific word in the collection of documents and deciding whether the word usage corresponds to what we were attempting to measure. Such a task is essential because of the polysemous nature of most words in common language use. A very good example is the word STRESS which is considered by many to be an important human factor in aviation safety. Yet, this single word, when used either as a noun or a verb, has many different meanings. For example, here are three different uses of this word and related inflected forms in JetBlue's aviation safety reports:

1. Stress as a state of mental or emotional strain or suspense

   "The whole sequence of events placed the additional stress of distraction on the crew"
   "He was under a lot of stress"

2. Stress as a force that produces strain on a physical body

   "No excessive stress was placed on the A.C"
   "I had a stress fracture in the 3rd metatarsal of the right foot"

3. Stress as a special emphasis attached to something

   "I would like to stress the outstanding job of everyone on the ground"
   "They further stressed that it was a good decision".

The most useful tool for this purpose is what is commonly known as a keyword-in-context list or KWIC list. This technique allows one to display in a single table, the occurrences of either a specific word or of all words related to category in their textual environment. The text is aligned so that all keywords appear aligned in the middle of the table. Figure 1 present an example of a KWIC list of some the words associated to a category STRESS.

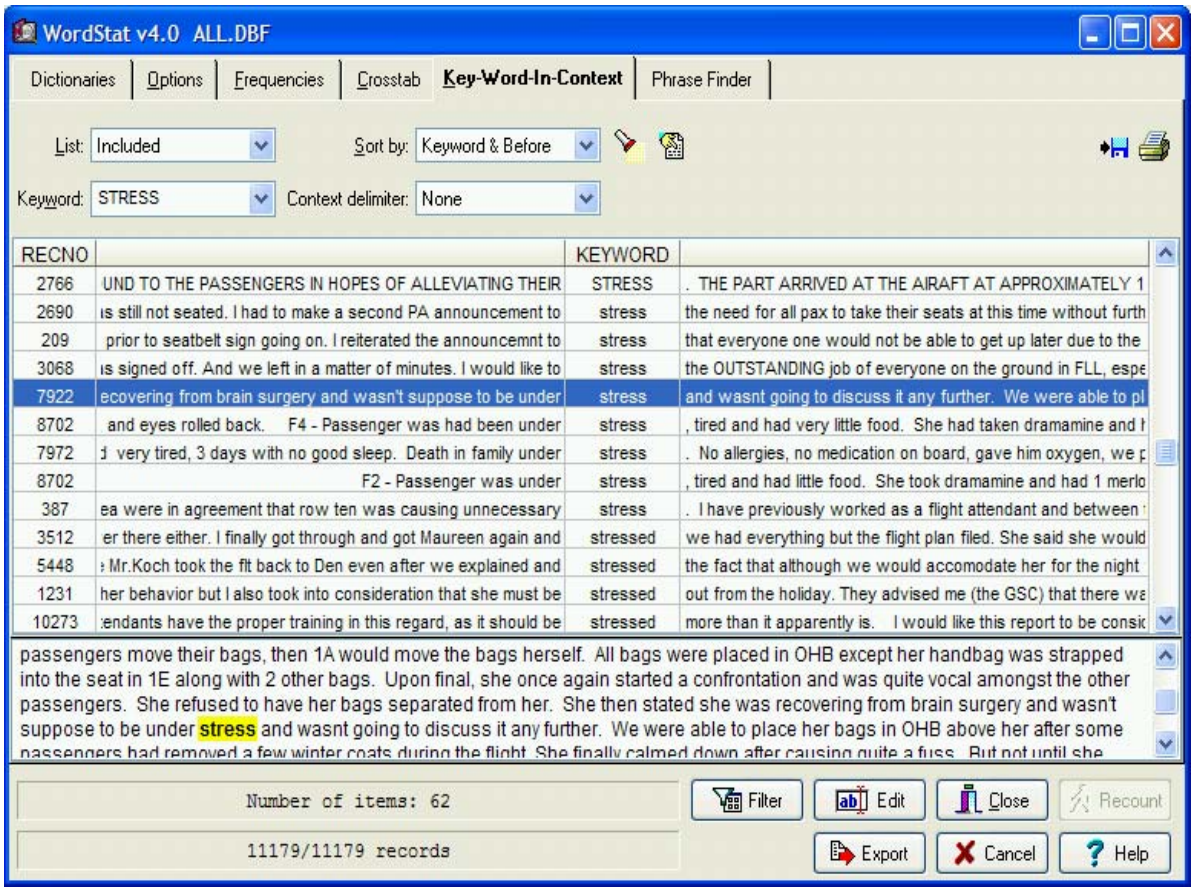


Figure 1. Keyword in context list of words in category "stress"

KWIC list is useful to assess the consistency (or lack of consistency) of meanings associated with a word, word pattern or content category. The list can be sorted on the keyword as well as the words appearing immediately before or immediately after it, allowing one to easily identify common phrases. Once an inconsistency has been detected one may decided to simply remove

this word from the content category or identify more specific expressions by looking at a few words appearing just before or after the target keyword. In the above example, it was found that phrases like "a lot of stress", "under stress" and "stressed out" seemed always associated with the emotional state and could very well be included as indicators of the emotional state. On the other hand, when "stress" or "stressed" was followed by "the", "to" or "that" then it was always referring to the action of singling out the importance of something. When such a situation occurs one may decide to only include phrases that allows one to catch the proper meaning or get rid of those phrases that are associated with improper ones. One may also decided not to include the original word because of the difficulty of performing disambiguation using the surrounding context. It is important to keep in mind that no matter how careful one will be in the selection of words and phrases to measure a specific dimension, it is likely that the inclusion of some items will result in categorization errors or what we may call false positives. The final decision of whether or not those items should be kept or removed ultimately depends on the relative importance of the two basic types of errors: 1) The exclusion of an item may result in an inability to identify the presence of a factor or the inability to retrieve relevant information (false negatives). 2) Its inclusion may, on the other hand, result in the improper identification of a factor or the retrieval of irrelevant information (false positives). As a rule of thumb, some authors have suggested using a threshold of 80% for the selection of words and phrases within a category. In other words, an item will be kept in a category if at least 80% of the returned hits are true positives. However, one may also consider that in some situations, missing a critical case has far worse consequences than generating false alarms or retrieving irrelevant reports. One could then reduce this threshold to a much lower value of 50% or even 20%. One may also wish to keep only the most reliable indicators by applying a more stringent criteria of 90% or even 100%.

3.0 Analysis Techniques Applied in this Project

WordStat offers two broad strategies or categories of statistical content analysis tools. The first one is based on the analysis of co-occurrences of words or keywords within reports, while the second one look at potential relationship between those keywords and numerical and categorical variables.

3.1 Analysis Based on Co-Occurrences of Keywords

Many contemporary text mining techniques, such as cluster analysis, latent semantic indexing, and many others are based on the analysis of how words co-occur within either sentences, paragraphs or entire documents. One of the assumptions behind such kind of analysis is that the close proximity of words reflects the actual or perceived structure of a specific domain, the nature of the relationship between its various elements or how attributes are tied to those elements.

WordStat provides three statistical and visualization techniques for the analysis of co-occurrences: hierarchical cluster analysis, multidimensional scaling and proximity plots. All those techniques require the computation of a matrix of similarity that measures the proximity among all selected words or keywords. The more often two words co-occur, the more similar (or related) they will be considered. Traditionally metrics such as the cosine metric, Dice or Jaccard coefficients have been frequently used to measure word co-occurrences. Many other metrics have also been proposed in the literature and there is still a lot of research going on to identify the best ones. While there are few consensuses today as to which measures are the most appropriates, we strongly believe that the answer such a question ultimately depends on the type of application one is interested in but also on the specific conditions under which those measures are being applied. In other words, if one attempt to perform data reduction for a document classification task, generate a taxonomy, discover hidden patterns or relationships, or improve text retrieval, then different similarity measures may prove to be useful. The size and homogeneity of the text corpus, the length of the documents, and the nature of the word categorization process may also influence the performance of specific metrics and should also be considered when selecting a metric.

In this project, different similarity coefficients were evaluated with regard to a knowledge discovery task in order to identify measures that could increase the likelihood of detecting unknown patterns or relationships. To our knowledge, few studies have assessed the values of similarity measures for such a task and the hypothesis was made that the required characteristics of similarity coefficients may well differ in nature from those needed for other tasks like dimension reduction, automatic taxonomy generation or document retrieval tasks. WordStat provides four indices of similarity frequently used by other text analysis and text mining tools. Those indices are the Jaccard, the Sorensen (or Dice), the Ochiai and the Cosine coefficients. It also includes a fifth measure seldom used by other text mining software: the inclusion index. This coefficient, developed in library sciences, was found to be very good at detecting asymmetric relations of dependence, inclusion or subordination among words, while other coefficients are more able to assess mutual occurrences. We believe such a characteristic could

be valuable for knowledge discovery precisely because of its greater sensitivity to asymmetric relationships so we decided to further test this impression during this project.

Another parameter that one may vary when computing co-occurrences and that will ultimately affect the nature of the results is the extent of the context on which co-occurrence will be defined. One can choose to define co-occurrence for two words when they appear anywhere in the same report or restrict the definition of co-occurrence to situations where those two words appear in the same paragraph, the same sentence or in a smaller window of 2 or 3 words. While all those four options may lead to different results, we believe that it is possible to reduce the number of options when analyzing aviation safety reports to two: co-occurrence within documents and within sentence. Analyzing co-occurrences within documents should provide information as to which events, actions or persons were mentioned in the same incidents (topical similarity). The analysis of co-occurrence within sentence should then allow one to more precisely associate those items (grammatical or semantic similarity). For example, one may find that the content category CONFUSION is frequently associated with a specific critical event. To assess whether such confusion is related to the event itself, a specific procedure, a device or someone in particular, then one could restrict the criteria of co-occurrences to sentences allowing one to locate more precisely the source of this confusion. Since the vast majority of safety reports consist of several sentences grouped under a single paragraph, we considered the analysis of co-occurrence within paragraphs to be redundant and found that it yields almost identical results as the analysis based on whole reports.

3.1.1 Cluster Analysis

Cluster analysis is a data reduction method by which a large number of items are grouped in a smaller number of clusters of similar items. While some clustering methods classify items based on a predefined number of clusters and present only a single partition solution, hierarchical cluster analysis progressively build clusters of items by successively grouping similar words, up until a single large cluster is formed. By providing detailed information on the entire agglomeration sequence, this technique allows one to select the optimal number of clusters that best suit one's needs.

The result of a hierarchical cluster analysis is typically presented using a dendrogram (see Figure 2) also know as a tree graph. In such a graph, the vertical axis is made up of the items (words or keywords) and the branches on the horizontal axis represent the clusters formed at each step of the clustering procedure. Words or categories that often appear together are combined into a cluster at an early stage while those that are independent from one another or don't appear together tend to be combined at the end of the agglomeration process. Figure 2 presents a dendrogram obtained by clustering categories of words in TCAS reports. A solution with 17 clusters was selected. Each cluster is represented using a different color.

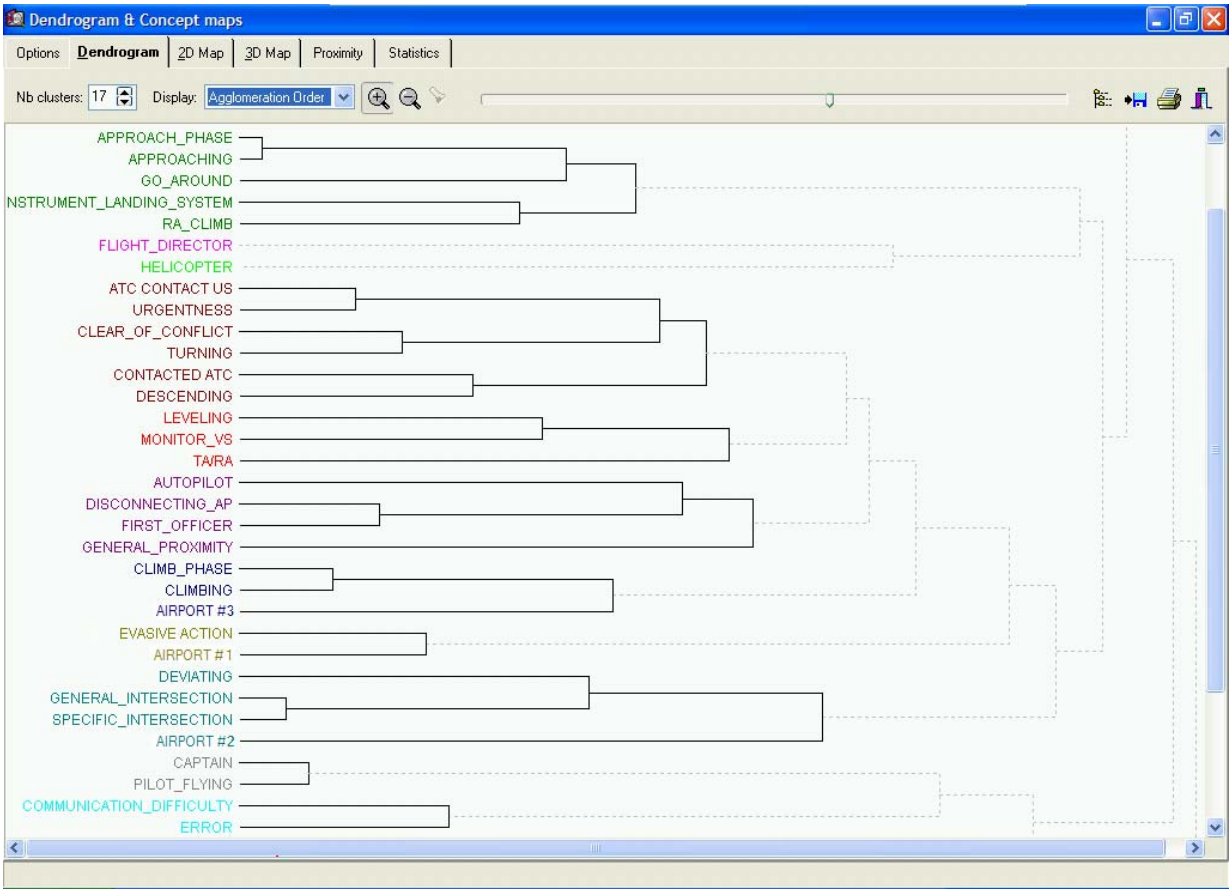


Figure 2. Dendrogram of keywords in TCAS reports

The top most cluster contains five items: APPROACH_PHASE, APPROACHING, GO-AROUND, INSTRUMENT_LANDING_SYSTEM, and RA_CLIMB. The next two clusters are isolated keywords that have not been grouped yet. The fourth cluster has six items, and so forth.

To illustrate the hierarchical clustering process, we will focus on the first cluster of five items and look at the sequence in which they had been aggregated. We can see in Figure 3 that the first two items that had been grouped were APPROACHING and APPROACH_PHASE. The short length of the leaves connecting these two items suggests that they co-occurred frequently in the same reports.

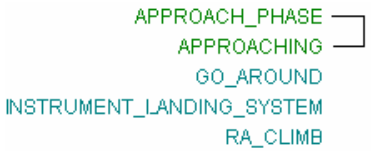


Figure 3. Four clusters solution

Categories INSTRUMENT_LANDING_SYSTEM and RA_CLIMB were then lumped together creating a solution with three clusters (Figure 4).

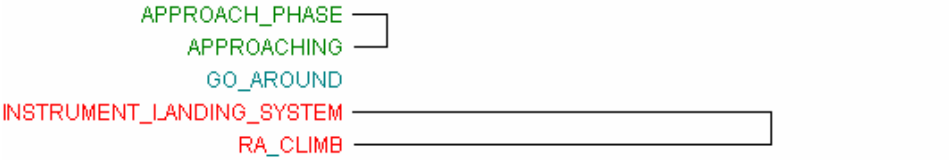


Figure 4. Three clusters solution

Then, the content category measuring the presence of go-around procedures was connected to the first two items, resulting in two distinct clusters (Figure 5).

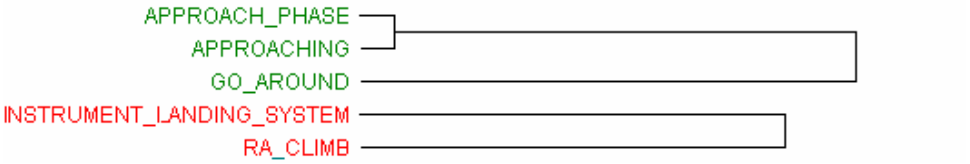


Figure 5. Two clusters solution

Those two clusters were then joined together to form a single cluster of five items (Figure 6).

Figure 6. Single cluster solution

When applied to textual data, cluster analysis is often used to identify themes or broad classes of concepts. It has also been used as a data reduction technique for document classification purposes, as well as a method to automatically generate taxonomies. These types of applications follow from the property of cluster analysis to form coherent groups of items that are logically or semantically related. The way items come together in the dendrogram of Figure 2 provides a clear illustration of this property. If we look at the items in the first cluster, it is representative of situations of TCAS alerts caused by the presence of a target below the airplane while in its final approach. In such a situation RA messages to climb are more likely to occur and will often result in a missed approach and thus require the pilot to initiate a go-around procedure. Also, the fact that items like CAPTAIN and PILOT_FLYING are grouped together in another cluster or that a TCAS message to monitor vertical speed (e.g., MONITOR_VS) is grouped with the content category measuring the action of leveling the airplane also illustrated this ability of hierarchical clustering to group's items that are logically related.

While cluster analysis seems suitable to replicate a known reality and identify logical relationships that are obvious to the subject matter expert, one should not conclude that such a tool could not be used to uncover new facts. Even if cluster analysis was not developed primarily as a tool for discovering unsuspected patterns or relationships, its ability to group words based on their co-occurrences and present those grouping graphically make it a potentially useful tool for such knowledge discovery tasks. In order to uncover new facts one should however be more attentive to the identification of unexpected associations than on the identification of coherent patterns. Those odd relationships could manifest themselves in at least two forms:

1) An aggregation of two logically unrelated items at an early stage of the clustering process.

2) A peculiar item in an otherwise coherent cluster.

The necessity to focus on clusters at an early stage or on coherent clusters comes from the fact that two items may be part of a same cluster yet never co-occur in the reports. Their common presence in this cluster can be explained by the fact that they both co-occur frequently with a third item. To assess more accurately the level of co-occurrence of two items in a large cluster, one should preferably consult the original similarity matrix or use other visualization methods such as the proximity plot presented in section 3.1.2.

The likelihood of discovering new knowledge may be increased by examining the co-occurrence of items belonging to different conceptual categories or items that one would not necessarily analyze together. For example, while analyzing co-occurrences of mechanical problems in aircrafts may yield interesting discoveries, one could also attempt to mix those types of problems with aircraft types, types of emergency procedures or content categories measuring different kinds of human factors and see how those separate dimensions are related. Good examples of such unexpected relationships may be found in the dendrogram presented in Figure 2. In this example, content categories for airports had been clustered along with items describing flight phases, TCAS messages, pilot actions and some situational variables. By focusing on airports codes, one will notice that the mention of the airport #1 seems to be more often associated with the content category measuring the presence of evasive actions. Also, the airport code for airport #3 is appended to a cluster of 2 items indicating the presence of a climbing action or a direct reference to the climb phase. This suggests that TCAS events observed at this airport more frequently occurs during this phase of flight. One can also observe that TCAS events occurring at the airport #2 are characterized by more frequent mentions of intersections, either by their names, contained in the SPECIFIC INTERSECTION category, or by the generic words used to designate them (e.g., "fix" or "intersection"). Whether those associations represent known facts, trivial new knowledge or potentially important discoveries is an issue that goes beyond the application of this statistical and graphical tool. The originality and the importance of such discoveries ultimately depend in part on what the subject domain expert already knows about the studied topic. Only such an expert can decide whether this information is new and interesting enough to warrant further investigation.

Technical comments

WordStat offers a choice of five measures of co-occurrences, each of them producing a different clustering solution. One may legitimately wonder whether the choice of such a metric could affect the likelihood that one would uncover unknown facts. While providing a definitive answer to such a question would require much further research, two factors justify the need for a closer examination of this question. First, from a logical point of view, most studies on text clustering have compared the effect of similarity measures and agglomeration methods on their ability to create coherent taxonomies, perform useful dimension reductions and other type of tasks only distantly related to knowledge discovery. Since the usefulness of different metrics

may very well depend on what one tries to achieve, then some metrics currently considered non optimal for the creation of coherent taxonomies could nevertheless be useful for discovering new facts. A practical example should allow us to illustrate such a possibility. Figure 7 presents a portion of a dendrogram obtained from the clustering of aircraft parts mentioned in In-flight Crew Irregularity Reports. A special metric called the inclusion index was used to measure the co-occurrence of those parts. This index remains virtually unknown to most experts in the text mining area, yet it appears to have a special ability to uncover unsuspected relationships.

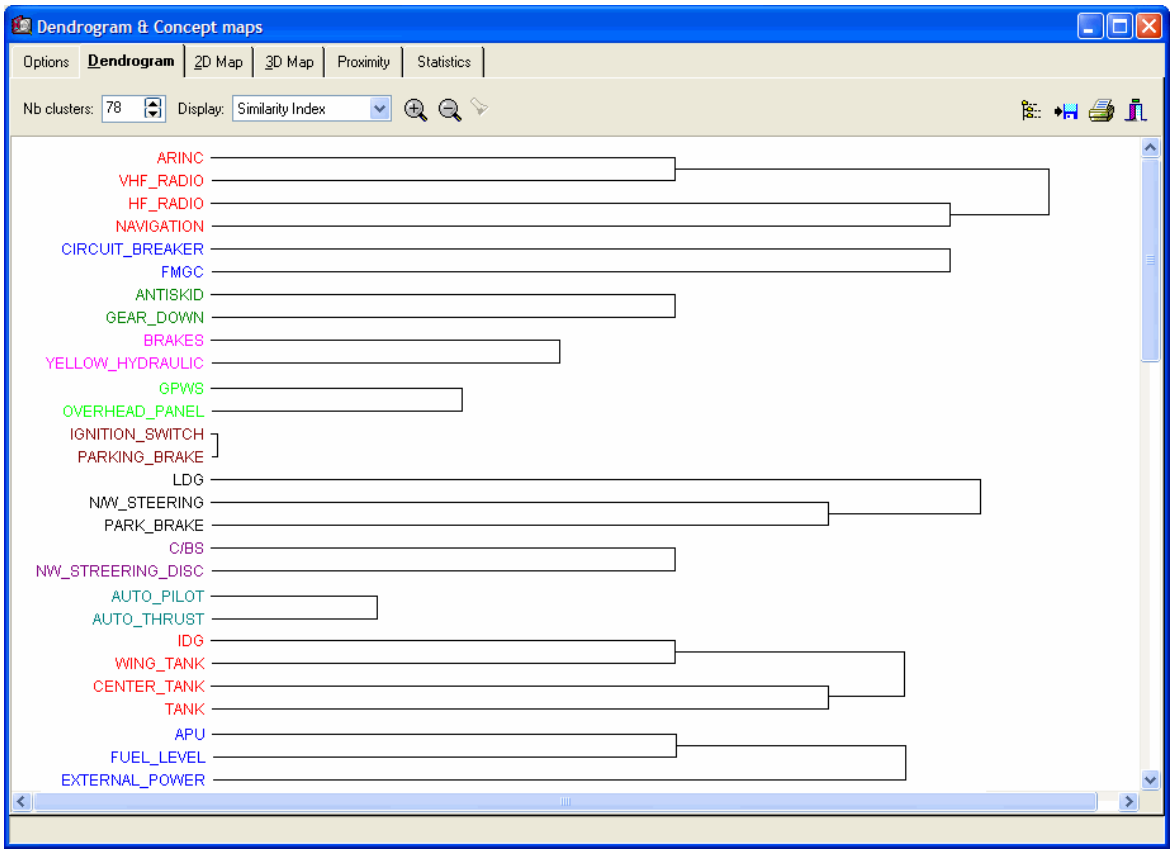


Figure 7. Dendrogram of aircraft parts

One will notice that many of the clustered of items represent logical associations (like AUTO-PILOT and AUTO-THRUST or GPWS and OVERHEAD PANEL). However, the early grouping of IGNITION SWITCH and PARK BRAKE appear to defy logic since these two controls are functionally unrelated, the first one being used for starting engines when departing, while the second one is used to immobilize the plane. Yet, the length of the leaf that joins these two items in the dendrogram clearly indicates the presence of a strong relationship between these two controls. This result was not entirely surprising to the JetBlue flight safety team, since it confirmed a recent discovery that pilots sometimes confused these two controls, both of which had to be turned clockwise and were located near each other, one above the other (see Figure 8). Examination of associated safety reports containing both items revealed that on some occasions during the taxi-out procedure, the park brake had been activated by mistake.

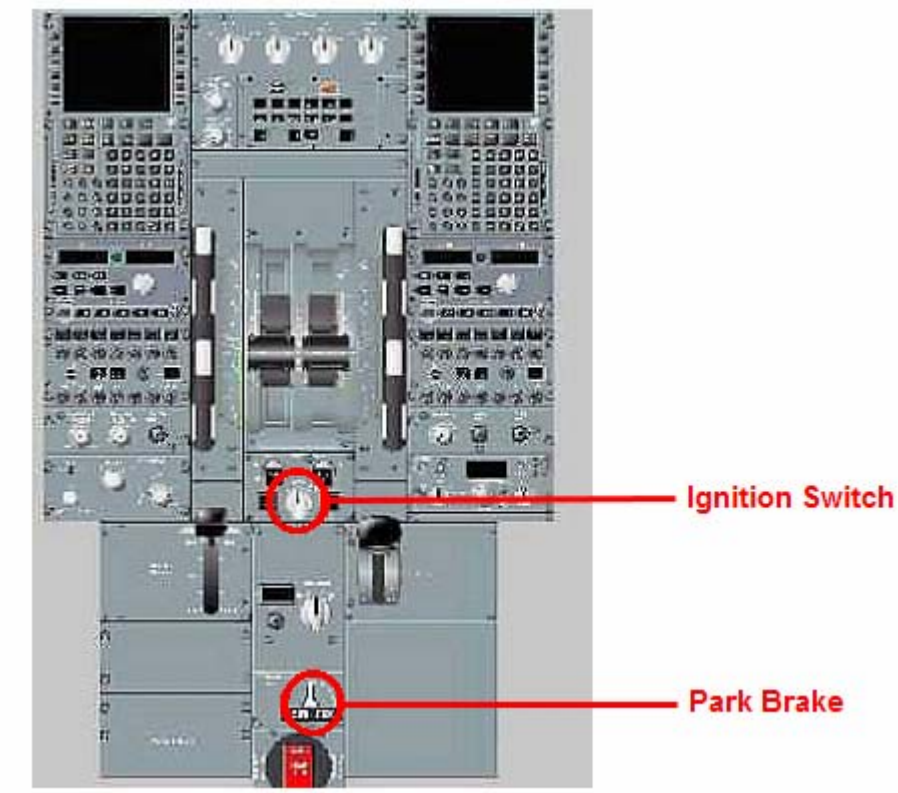


Figure 8. Cockpit panel

While the identification of such a relationship was not a discovery per se for the JetBlue flight safety team, it confirmed the value of the clustering method for uncovering unexpected but real facts. Yet, such a relationship would have remained undetected if a more conventional similarity index like the Jaccard or the Dice coefficient had been used. The higher sensitivity of the inclusion index is due to the fact that it will reach a high value when an item is often associated with another one, even when the reverse is not true. In the above example, mentions of parking brakes in safety reports were much more numerous than references to the ignition switch. However every report where a mention to ignition switch was made also included a reference to park brakes, causing the inclusion index to reach its maximum value of 1.00. In comparison, the Jaccard similarly index between those two items was relatively low (i.e. only 0.11). While the value of this inclusion index still need further confirmation, we believe that its usefulness for knowledge discovery may very well stands whether or not this metric is found to be useful for other types of clustering applications such as automatic taxonomy generation or data reduction.

The application of conventional metrics to aviation safety data also led us to the identification of another limit of some co-occurrence indices. We noticed that clustering solutions obtained using those metrics could be strongly biased toward the formation of clusters of high frequency items. It was even sometimes possible to predict the formation of specific clusters simply by looking at the initial frequency distributions of keywords: the more frequent a word or keyword was, the more likely it was to be included in a cluster with other high frequency words. Such a situation happens because traditional co-occurrence measures do not take into account the possibility that two words will sometimes co-occur by chance. While the problem may remain undetected when clustering low frequency words and when analyzing co-occurrence within a limited context (such

as within a sentence of within a window of a few words) - the problem become much more apparent when clustering broad content categories or highly frequent words. This problem is also exacerbated when examining co-occurrences within large contexts such as whole safety reports. Such a finding led us to develop and start testing probabilistic versions of those co-occurrence metrics that would take into account co-occurrences due to chance. While the testing of such measures will likely take several months, the initial results are encouraging and we believe such measures may prove to be more useful for knowledge discovery and text mining of aviation safety data than traditional co-occurrence metrics.

Findings -- Cluster Analysis

Our work on the JetBlue safety data also required us to examine the potential value of dendrograms for knowledge discovery tasks and search for ways to improve it for this type of application. This led us to the implementation and testing of two important changes: 1) the addition of an integrated text retrieval feature and 2) the optional removal of single item clusters.

While attempting to generate new knowledge using cluster analysis, the very first thing one should normally do when an unexpected relationship is identified is to return to the original reports that led such a discovery. The examination of those reports will either confirm the importance of the finding and justify further investigations or yield the user to discard this information. However, the software initially offered no easy way to select a cluster and retrieve associated safety reports. It quickly became evident to the software vendor that such a feature was not only useful or convenient but essential for this type of knowledge discovery process. Such a feature was already available elsewhere in the program allowing a user to select a block of cells in a heatmap chart and retrieve all associated reports. It was thus decided to develop a similar feature for dendrograms. The implementation that was made consisted of allowing the user to click anywhere on a cluster to highlight it and then click on a search button to retrieve all reports that had contributed to its formation (all reports with at least two items from the selected cluster). It became clear that this ability to move from any graphic representation to the original reports behind a specific segment of it was extremely useful and should be generalized as much as possible to other graphic representations.

As mentioned previously, one way to extract potentially interesting knowledge from dendrograms is to identify an aggregation of two logically unrelated items at an early stage of the clustering process. However, when clustering hundreds or thousands of items, locating items that have been grouped early may be difficult and require the user to scroll through a very long list of items in order to identify those that had been grouped together. Since at such an early stage, a majority of clusters are composed of isolated items and that those items provide no useful information, we considered that it would make things easier for the user to hide all those single items clusters and display items only when they become aggregated with other ones. Such a small change has the effect of simplifying the dendrogram display and allows the user to concentrate on the most relevant information.

3.1.2 Proximity Plots

Hierarchical cluster analysis is typically used as a data reduction technique. A dendrogram displays the sequence by which items had been grouped together into clusters. While it attempt to form homogeneous groups of related items, it does not always accurately represent the true proximity of keywords to each other. Two keywords may be grouped under the same cluster, not because they co-occur frequently, but because they are both highly related to a third keyword. Only the first few items added to a specific cluster can readily be interpreted as a direct indication of proximity. While multidimensional scaling, also available in WordStat, provides a more accurate representation of the real proximity of objects, the fact that it attempts to represent the various points into a two- or three-dimensional space usually results in some distortion. As a consequence, some items that tend to appear together or be very similar may still be plotted far from each other, especially when a large number of words or categories are plotted. Therefore, multidimensional scaling was not used in this project.

A third and more accurate way to graphically represent the distance between objects is the proximity plot. This type of plot presents a comprehensive list of all words or categories of words associated to a specific item and displays them in descending order of proximity. In this plot, all measured distances are represented by the distance from the left edge of the plot. The closer an object is to the selected one, the closer it will be to the left and the higher it will be in the ordered list. This graphic representation is ideal to get a profile of words or keywords associated with a specific item. Figure 9 below shows all categories of words in TCAS safety reports associated with a mention of the airport #2.

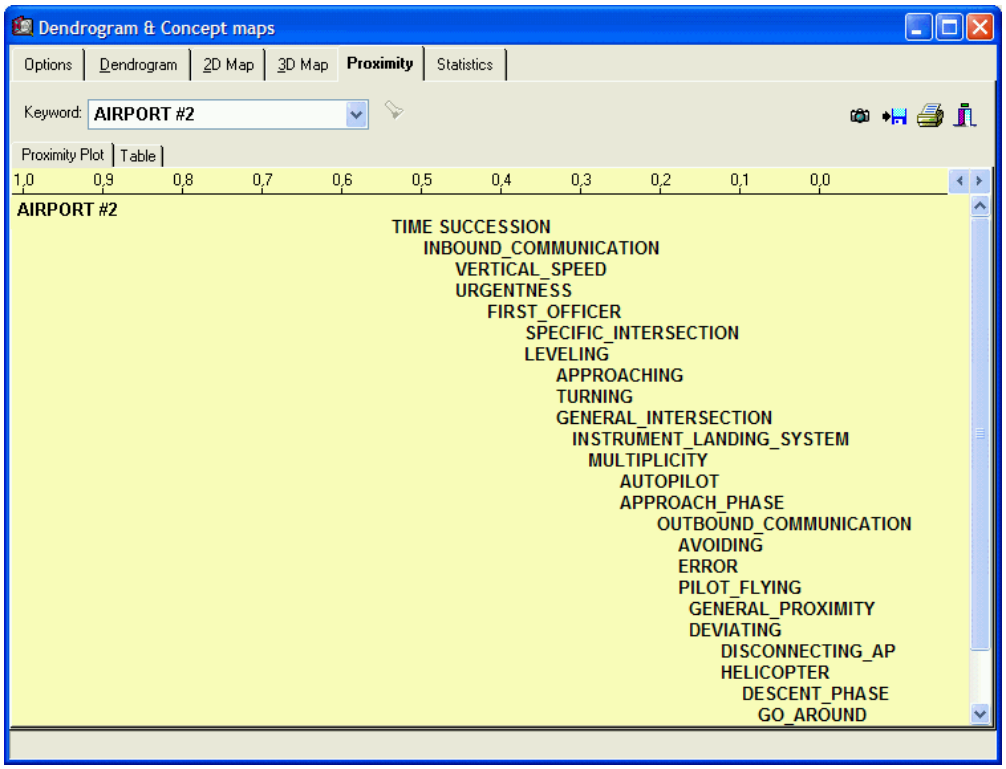


Figure 9. Proximity plots of keywords related to Airport #2

This profile may then be compared with others profiles obtained for other airports like the one below (Figure 10) describing reports for TCAS events occurring at the airport #3.

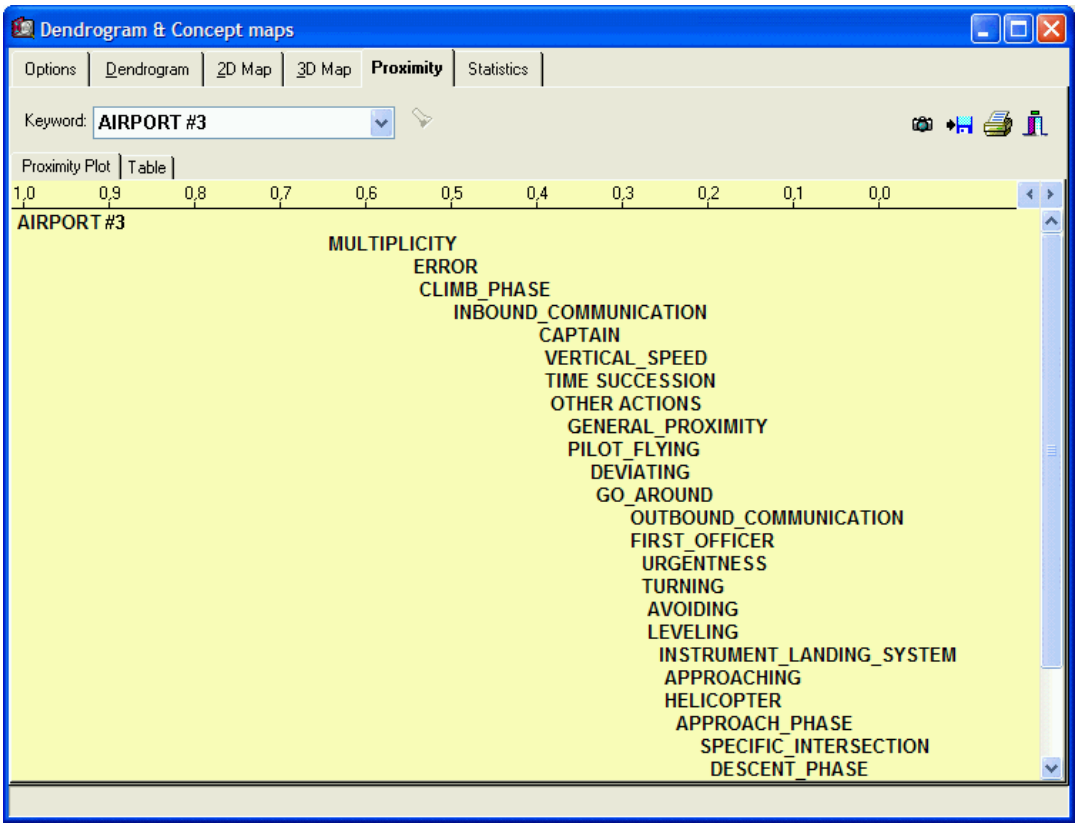


Figure 10. Proximity plots of keywords related to Airport #3

A close examination of the two profiles led us to make some hypothesis as to possible differences in the nature of TCAS events occurring at these airports. Numerically speaking, collision risks are known to be more frequent at the airport #3. The proximity profile indicates that those collision risks occur more frequently during the climb phase. It also suggests that those events often involve several conflicts occurring either at the same time or in succession. This is evident in the proximity plot above where one can see that the content category most closely related to this airport is MULTIPLICITY, which includes items such as "2 aircraft", "second RA", "multiple", "several" as well as plural forms of nouns such as "conflicts" and "targets". While numerically less frequent, TCAS events occurring at the airport #2 seemed to be characterized as more urgent, occurring quickly and requiring quick reactions. This is expressed by the relatively high positions of the TIME SUCCESSION and URGENTNESS content categories. The first one suggests a quick appearance or succession of events with items such as "all of the sudden", "almost immediately", "quickly followed by", "seconds later", etc. The second category tries to tackle the rapidity of responses with phrases like "immediately initiated", "stopped climb immediately", "quickly disconnected" or the abruptness of some moves with words like "aggressively", "sharp", "radical", "abrupt", "speedily", etc. While one could argue that the two categories could be logically merged into a single category, the fact that both appear to be characteristic of TCAS events occurring at airport #2 seems to be worthy of interest. From the analysis of the other keywords characterizing events occurring at this airport, we can assume

that those events occur more often during the approach phase (categories APPROACHING and APPROACH PHASE). While those two categories are also associated with the airport #3, they are nevertheless relatively more predominant for the airport #2. Another characteristic feature of TCAS events at this airport is the more frequent mention of intersections, a finding also revealed by hierarchical cluster analysis (see section 3.1.1). This last finding is expressed by the frequent mention of either specific intersection names included in the SPECIFIC INTERSECTION category or by the predominant use of generic words used to refer to intersections such as "fix" and "intersection". Great caution should however be taken when interpreting those results. Because of the low number of TCAS reports analyzed, there are still some risks that some of those patterns may not be reliable. It was nevertheless decided to further investigate some of those findings and try to substantiate them with other sources of evidence.

Findings -- Proximity Plot Analysis

The ability to focus on specific items and establish a profile of co-occurrences with the selected item is one of the main benefits of this type of chart compared to dendrograms or multidimensional scaling plots. In the above example, we attempted to compare profiles of TCAS events taking place at different airports. One could also use proximity plots to identify relationship between specific categories of human errors and flight events or aircraft devices, between mechanical problems and aircraft types, or look at the range of actions associated with a critical event.

Just like it has been done with dendrograms, a "select and search" feature had been implemented to automatically retrieve reports underlying the measured co-occurrence between two items. Such a feature has proven to be very useful to assess the validity of the findings and to provide relevant information for the interpretation of such co-occurrences.

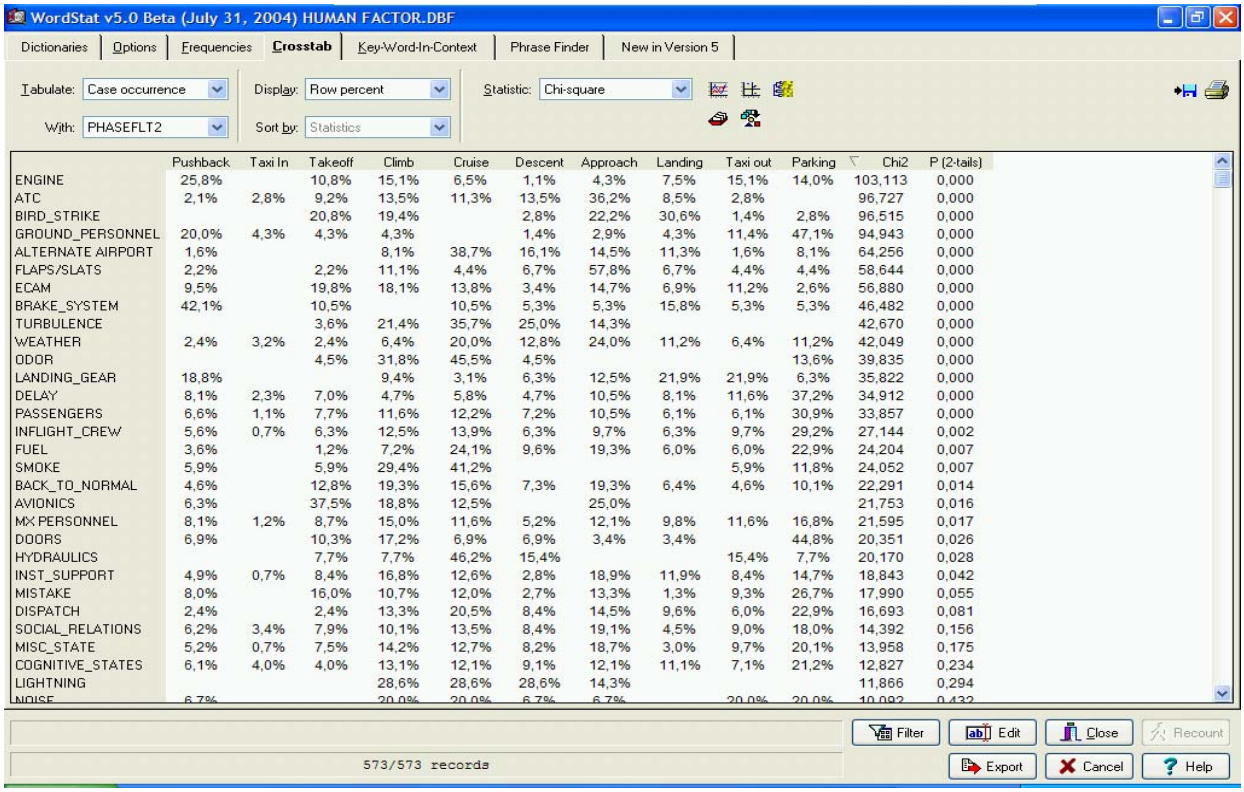
Various suggestions for further improvements of proximity plot could also be made. Comparisons of profiles require the user to move back and fourth between the items he wishes to compare. Such a task would be greatly facilitated by the simultaneous display of several profiles side-by-side or by the charting of the differences in the co-occurrence measures. Also, information about the statistical significance or the sampling variability of each measure could also provide useful information and prevent investigating patterns that are likely caused by chance.

3.2 Keywords by Numerical or Categorical Variables

The second broad type of analysis provided by WordStat is the ability to compare keyword occurrence or frequency in documents to numerical or categorical attributed of those documents. In the context of text mining aviation safety reports, it consists of comparing words or categories of words in safety reports to any structured information in the database such as the date, the time of the event, the flight number, the altitude, the author of the report, or any other information that has been previously categorized (e.g., phase of flight, type of event, etc.).

This type of analysis imposes the supplemental requirement over plain analysis of co-occurrences of ensuring that numerical or categorical information are devoid of data entry error and that they are reliable and exhaustive. The distributional property of those variables should also be verified to make sure that reports are properly spread across the categories or across the range of values of numerical variables. For example it was found that for the categorical variable "Phase of Flight" in the FCIR reports, there were too few events related to the Taxi-in phase to provide reliable description of events associated with this phase. The number of TCAS events occurring at many airports were also too low to be used for describing the nature of collision risks at those airports so comparisons had to be restricted to the four most frequent airports. For other variables, the number of missing values was quite high and it was not possible to figure out whether they were missing because this information was not available or simply not required.

The most common way of displaying relationship between words or categories of words and those categorical or numerical variables is through a crosstabulation table, like the one showed in Figure 11 below.



	Pushback	Taxi In	Takeoff	Climb	Cruise	Descent	Approach	Landing	Taxi out	Parking	Chi2	P (2-tails)
ENGINE	25,8%		10,8%	15,1%	6,5%	1,1%	4,3%	7,5%	15,1%	14,0%	103,113	0,000
ATC	2,1%	2,8%	9,2%	13,5%	11,3%	13,5%	36,2%	8,5%	2,8%		96,727	0,000
BIRD_STRIKE			20,8%	19,4%		2,8%	22,2%	30,6%	1,4%	2,8%	96,515	0,000
GROUND_PERSONNEL	20,0%	4,3%	4,3%	4,3%		1,4%	2,9%	4,3%	11,4%	47,1%	94,943	0,000
ALTERNATE AIRPORT	1,6%			8,1%	38,7%	16,1%	14,5%	11,3%	1,6%	8,1%	64,256	0,000
FLAPS/SLATS	2,2%		2,2%	11,1%	4,4%	6,7%	57,8%	6,7%	4,4%	4,4%	58,644	0,000
ECAM	9,5%		19,8%	18,1%	13,8%	3,4%	14,7%	6,9%	11,2%	2,6%	56,880	0,000
BRAKE_SYSTEM	42,1%		10,5%		10,5%	5,3%	5,3%	15,8%	5,3%	5,3%	46,482	0,000
TURBULENCE			3,6%	21,4%	35,7%	25,0%	14,3%				42,670	0,000
WEATHER	2,4%	3,2%	2,4%	6,4%	20,0%	12,8%	24,0%	11,2%	6,4%	11,2%	42,049	0,000
ODOR			4,5%	31,8%	45,5%	4,5%				13,6%	39,835	0,000
LANDING_GEAR	18,8%			9,4%	3,1%	6,3%	12,5%	21,9%	21,9%	6,3%	35,822	0,000
DELAY	8,1%	2,3%	7,0%	4,7%	5,8%	4,7%	10,5%	8,1%	11,6%	37,2%	34,912	0,000
PASSENGERS	6,6%	1,1%	7,7%	11,6%	12,2%	7,2%	10,5%	6,1%	6,1%	30,9%	33,857	0,000
INFLIGHT_CREW	5,6%	0,7%	6,3%	12,5%	13,9%	6,3%	9,7%	6,3%	9,7%	29,2%	27,144	0,002
FUEL	3,6%		1,2%	7,2%	24,1%	9,6%	19,3%	6,0%	6,0%	22,9%	24,204	0,007
SMOKE	5,9%		5,9%	29,4%	41,2%				5,9%	11,8%	24,052	0,007
BACK_TO_NORMAL	4,6%		12,8%	19,3%	15,6%	7,3%	19,3%	6,4%	4,6%	10,1%	22,291	0,014
AVIONICS	6,3%		37,5%	18,8%	12,5%		25,0%				21,753	0,016
MX PERSONNEL	8,1%	1,2%	8,7%	15,0%	11,6%	5,2%	12,1%	9,8%	11,6%	16,8%	21,595	0,017
DOORS	6,9%		10,3%	17,2%	6,9%	6,9%	3,4%	3,4%		44,8%	20,351	0,026
HYDRAULICS			7,7%	7,7%	46,2%	15,4%			15,4%	7,7%	20,170	0,028
INST_SUPPORT	4,9%	0,7%	8,4%	16,8%	12,6%	2,8%	18,9%	11,9%	8,4%	14,7%	18,843	0,042
MISTAKE	8,0%		16,0%	10,7%	12,0%	2,7%	13,3%	1,3%	9,3%	26,7%	17,990	0,055
DISPATCH	2,4%		2,4%	13,3%	20,5%	8,4%	14,5%	9,6%	6,0%	22,9%	16,693	0,081
SOCIAL_RELATIONS	6,2%	3,4%	7,9%	10,1%	13,5%	8,4%	19,1%	4,5%	9,0%	18,0%	14,392	0,156
MISC_STATE	5,2%	0,7%	7,5%	14,2%	12,7%	8,2%	18,7%	3,0%	9,7%	20,1%	13,958	0,175
COGNITIVE_STATES	6,1%	4,0%	4,0%	13,1%	12,1%	9,1%	12,1%	11,1%	7,1%	21,2%	12,827	0,234
LIGHTNING				28,6%	28,6%	28,6%	14,3%				11,866	0,294
NOISE	6,7%			20,0%	20,0%	6,7%	6,7%		20,0%	20,0%	10,092	0,432

Figure 11. Crosstabulation of content categories by phases of flight

WordStat can display statistics based on the total number of times a specific word or category of words occurs (word frequencies) or on the total number of reports that include those words (case occurrences). To facilitate the comparison across classes of categorical variables, one will typically transform absolute frequency into percentages. In the above example, the percentage per row is used and can be interpreted as the distribution of events across phases of flights. For

example, we can see that bird strikes are more likely to occur during landing (30.6%) followed by the approach (22.2%) and the take-off phase (20.8%). Percentages on the seventh line indicates that ECAM alerts are more likely to occur during take off and the climbing phase that any other phases of flights, with percentages of 19.8 and 18.1% respectively.

One could also identify the most prevalent events occurring at specific phases of flight by setting the percentage to "category percent" rather than row percent. When this option is selected, percentages are computed on the total number of reports occurring in this phase of flight.

While the information displayed in such table is comprehensive, the quantity of numerical information it contains may be overwhelming for most users, especially when dealing with very large tables. WordStat provides several tools to facilitate the extraction of relevant information from those tables. For example, one may sort, rows in ascending or descending order of percentages. One may also compute an association statistic to assess whether there are any substantial or statistically significant differences. For example, in the above table, rows have been sorted on the computed chi square value allowing one to focus on words that are more strongly associated with phases of flights. WordStat provides eight statistics to assess the strength and probability of associations. The most commonly used measures are the chi-square of the F-test for categorical variables (e.g., phases of flight, aircraft types) and the Pearson correlation for numerical variable (e.g., date, altitude, level of visibility).

From this, table, one may also select one or several rows and display differences across phases of flight using either bar chart or line charts. Such types of charts are appropriate to visualize the relationship between the select variable and specific words or keywords. However, bar charts and line charts are inherently limited in the number of dimensions they can display and quickly become cluttered as the number of classes or keywords increases. The next two sections will present different types of charts that are especially useful to visualize data from large crosstabulation tables: the heatmap and the correspondence plot.

3.2.1 Heatmaps

Heatmaps are especially useful for text mining when one wishes to examine the relationships between a large number of words and categories of words against a large number of classes of categorical variables. Such a visualization tool requires that the number of cases (or reports) for each row and column be large enough to insure the stability of the estimated percentages. For this reason, heatmaps are generally applied on databases with tens of thousands of cases. Unfortunately, such a requirement could not be met during the current project because of the limited sizes of the JetBlue safety reports databases. We were thus unable to fully assess the value of heatmaps for the analysis of aviation safety data. We nevertheless present this tool using a specific example obtained from the analysis of JetBlue database and discuss its potential usefulness for aviation safety data. Heatmaps have proven to be very helpful in biomedical research, especially in micro-arrays research, helping researchers to deal with huge amount of published results. Hopefully such a tool will also prove to be helpful for the analysis of large databases of airline safety data.

While some graphical tools like correspondence plot (see section 3.2.2) attempt to reduce the quantity of information stored in huge crosstabulation tables to a smaller number of dimensions, a heatmap preserves the integrity of the data in such tables. It simply organize and present this information in a way that maximize the likelihood that one would identify relationships between words or categories of words and specific classes of a categorical variable. This is achieved in two different ways. First, relative frequencies are represented using color brightness or tones rather than numerical values, facilitating the identification of high or low frequency cells, which corresponds respectively to bright and dark cells. Rows and columns are also reordered based on their similarity of patterns. For example, let suppose we examine the relationship between mechanical problems, each problem being represented on a separate row, and different aircraft types, plotted in separate columns. The dual clustering of rows and columns will result in the grouping of aircrafts sharing the same mechanical problems as well as the grouping of mechanical problems that tend to appear in the same aircrafts.

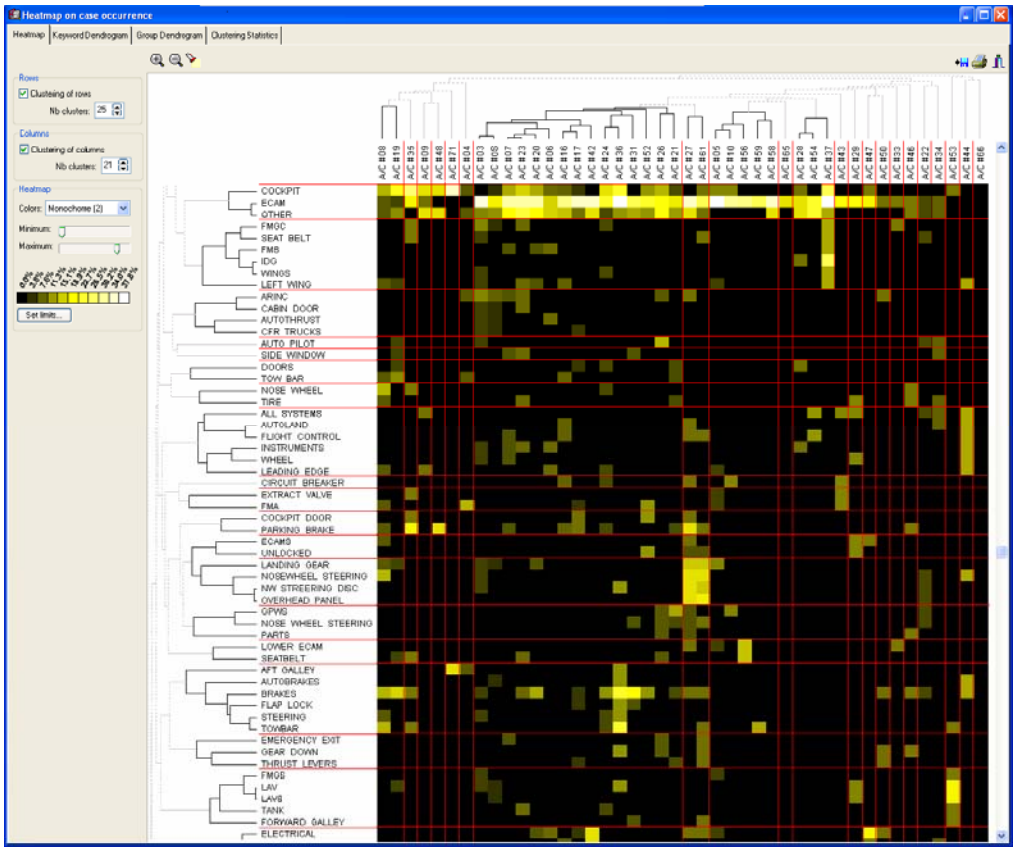


Figure 12. Heatmaps of relative frequencies of aircraft parts by aircraft numbers

In the above example (Figure 12), a crosstabulation of aircraft parts mentioned in safety reports in relation with aircraft numbers is represented graphically. While identifying cells with high frequency in a large crosstabulation table would normally be a tedious task, one can immediately spot on the corresponding heatmap regions of relatively high frequency. For example, at the top of the map, a large area of brighter cells illustrates the high frequency of ECAM alerts in safety reports. Isolated bright cells may also be spotted at various locations.

On the left side and at the top of the chart, one can see dendrograms indicating how rows and columns have been clustered together. Figure 12 represents two segments of those dendrograms.

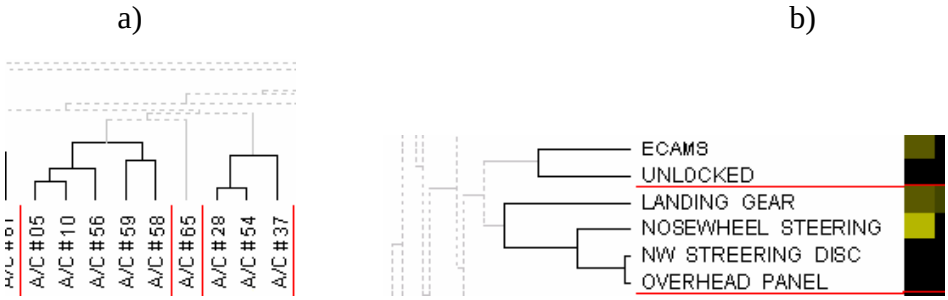


Figure 13. Two segments of heatmaps

We can notice in Figure 13a that aircrafts #05, #10, #56, #59, and #58 have been grouped together and that #28, #54 and #37 were also joined together to form a second cluster. These aircrafts were grouped because safety reports associated with them were similar in regard to the aircraft parts mentioned. Also, if we look at a portion of the dendrogram of rows (Figure 13b) we can also notice a cluster formed by 4 categories: LANDING_GEAR, NOSEWEEL_STEERING, NS_STEERING_DISC and OVERHEAD_PANEL. This cluster suggests that mention of these parts tend to be associated with the same aircrafts.

One of the purposes of clustering both rows and columns in heatmaps is the identification of functional relationships between rows and columns. For example, if a heatmap is used to represent frequencies of mechanical problems across a wide variety of aircraft types, then it becomes possible to identify whether one or several mechanical problems are characteristics of an aircraft of a group of aircrafts. Such a functional relationship would be expressed by the presence of cells clumps of relatively high or low frequencies at the intersection of rows and columns clusters (the limits of those clusters are indicated by horizontal and vertical red lines). Figure 14 below illustrates three examples of potentially functional relationships that could be extracted from the heatmap presented in Figure 12.

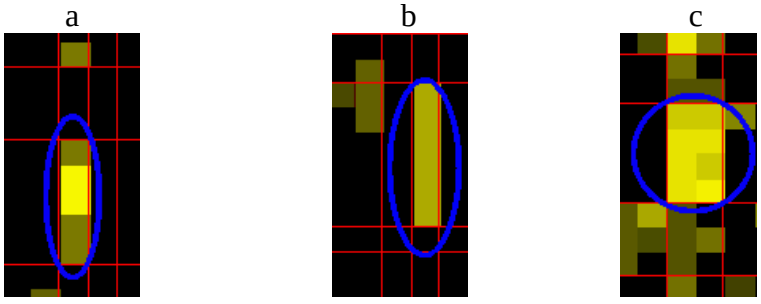


Figure 14. Examples cell clumps of relatively high frequencies

These groups of cells are characterized by relatively coherent patterns of brighter cells. For example, Figure 14c suggests *that two aircrafts could be considered to have a similar patterns of safety issues because of the* relatively high number of reports mentioning the same four parts (e.g., LANDING GEAR, NOSEWHEEL_STEERING, NS_STEERING_DISC and OVERHEAD_PANEL). To

assess whether such pattern is an indication of a genuine functional relationship, the user can select with the mouse the region of eight cells corresponding to the intersection of the two clusters and click on the Search button to retrieve the associated reports. Examination of those reports can "bring" the airline safety expert to either discard this information or confirm the validity of such a relationship.

Findings -- Heatmaps

It is important to remind that the above example is based on the analysis of a limited quantity of safety reports. The number of reports available per aircraft was sometimes very low and consequently, the observed patterns "are very likely to be the caused by chance". We nevertheless presented this charting method to illustrate how one could interpret results presented in heatmaps when applied on a large database. Yet, this example provides a clear illustration of the major limitation of such a visualization tool: in order to support the identification of reliable patterns, it should be applied on database with a large number of cases. One could however restrict the analysis to those categories with the highest frequencies. However, it would alleviate one of the major strengths of this kind of graph, the ability to graphically represent high-dimension data.

3.2.2 Correspondence Plots

Correspondence analysis is a descriptive and exploratory technique designed to analyze relationships among entries in large crosstabulation tables. Its objective is to represent the relationship among all entries in the table using a low-dimensional Euclidean space such that the locations of the row and column points are consistent with their associations in the table. The correspondence analysis procedure implemented in WordStat allows one to graphically examine the relationship between words or content categories and classes of an independent variable. The results are presented using a 2 or 3-dimensional map. Correspondence analysis statistics are also provided to assess the quality of the solution. WordStat currently restricts the extraction to the first three axes.

Contrary to heatmaps (section 3.2.1), correspondence analysis can be considered a data reduction technique, since it allows one to extract from large crosstabulation tables the most important dimensions and to locate the most significant differences. Figure 15 illustrated the patterns of relationships between four airports and categories of words in TCAS reports. Those patterns illustrate the first two factors extracted using correspondence analysis. The single most important dimension is represented by the horizontal axis with the airport #2 on the left side and the airport #1 on the opposite side. The vertical red line separating the left from the right half is designated as the origin of this dimension. Since factors are extracted in descending order of importance and since the distance from the origin can be interpreted as a measure of the singularity of items, we can conclude from the examination of this first dimension that the most singular airport in this small group is airport #1, airport #2 being the second most singular one. Airport #3 and airport #4 airports are located near the origin of this axis, suggesting that they do not differentiate themselves from the others, at least on this first factor.

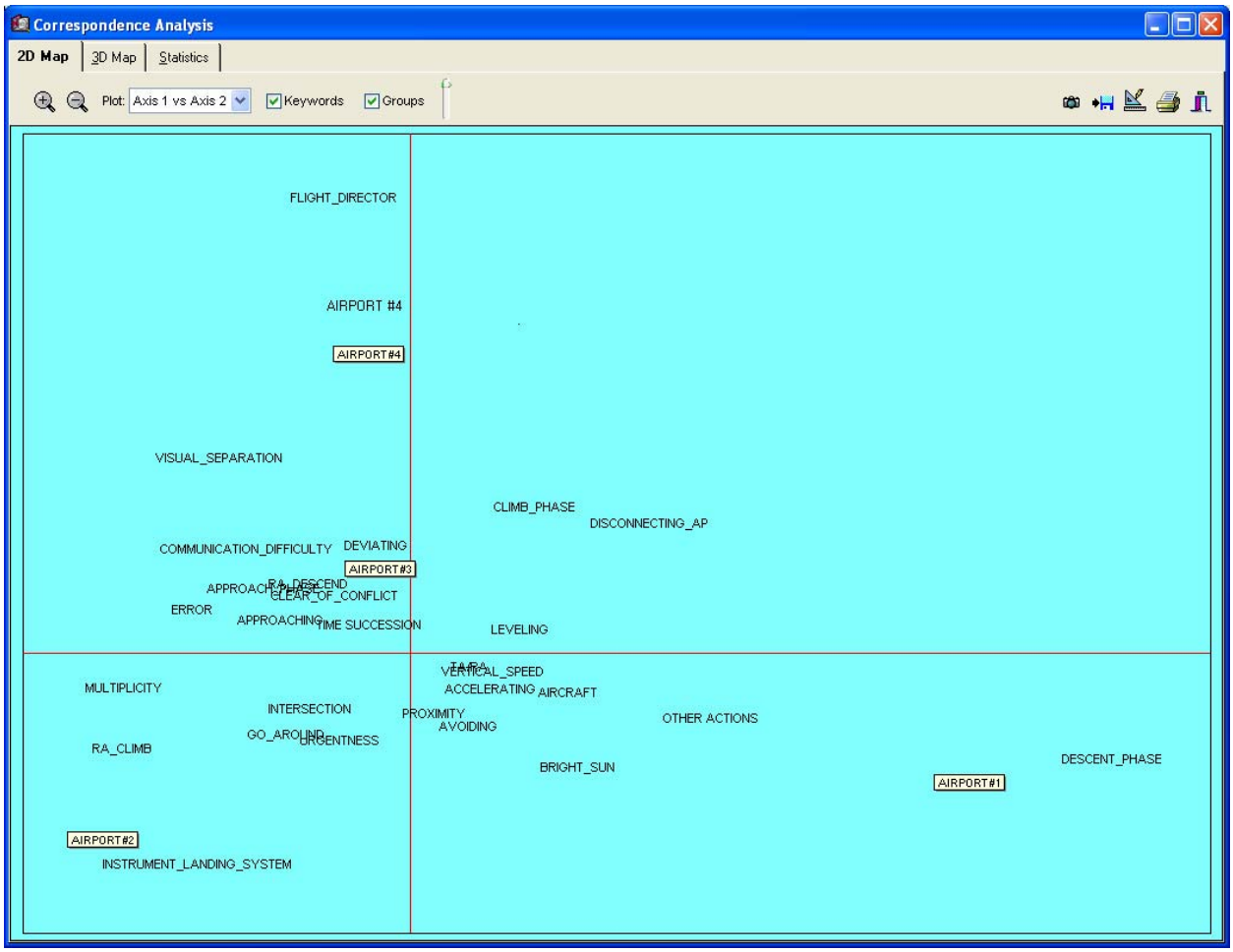


Figure 15. Correspondence plot of airports and keywords in TCAS reports

The second extracted factor, represented by the distance on the vertical axis suggests that airport #4 (top) differentiates itself from the others, especially from the #2 and #1 airports. Airport #3 remains near the origin of both axes. It is important to keep in mind that correspondence analysis does not depict the numerical importance of the TCAS events occurring at different airports. We already knew that TCAS events were much more likely to occur at this airport than at any other airports serviced by JetBlue. Correspondence analysis illustrates differences not in the number but in the nature of those events, suggesting that TCAS events at airport #3, while more frequent, are generally not different from those occurring at these other airports.

Correspondence analysis not only identifies the most singular classes of a categorical variable (in the above example, airports), but also provides some information as to why those items are different from the others. Such information is provided by the position of the content categories on the same axis as well as to their location relative to the airports codes. Again, the more singular a word or content category is, the farther it will be from the origin of the axis. In the above example, we can see that the most singular content category for the first axis is DESCENT_PHASE since it is the farthest from the origin of the horizontal axis. Other things being

equal, we can thus make the assumption from this single dimension that the singularity of the airport #1 may in part be explained by a higher proportion of TCAS events occurring during descent phase. If we look at the vertical dispersion of keywords, we can observe that the content category FLIGHT_DIRECTOR seems to be characteristic of the second extracted dimension, allowing us to assume that reports of TCAS events at the airport #4 are characterized by more frequent mentions of the flight director.

While the above interpretations were performed by considering a single dimension at a time, it is also possible and often recommended to take into account multiple dimensions when assessing the relationships between keywords and classes. One way to do this is by visualizing the angle of both the class (airport code) and the content category from the origin of all dimensions represented in the graph. An acute angle indicates that the two characteristics are closely associated or correlated while an obtuse angle indicates that the two items are negatively related. The very acute angle between airport #1 and DESCENT_PHASE (see Figure 16) suggests that the two items are closely associated. On the other hand, since items like FLIGHT_DIRECTOR and VISUAL_SEPARATION are located in the opposite quadrant and are rather far from the origin, we can assume that they are less likely to occur at this airport. In fact, a close examination of TCAS reports confirm that words associated with those two content categories were never found in reports of TCAS events occurring at this airport.

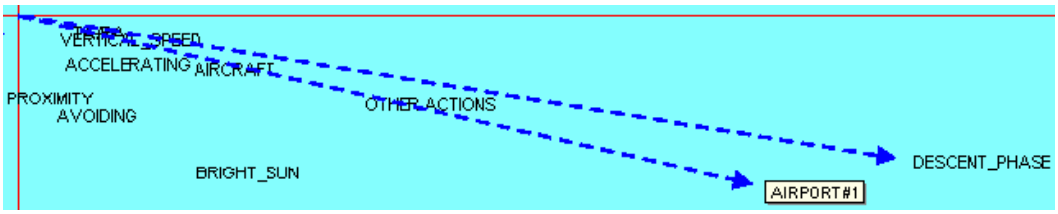


Figure 16. Angle from the origin between airport #1 and DESCEND PHASE

Similar characterizations of TCAS events occurring at different airports can also be made using the same technique. For example, we may conclude from Figure 15 that TCAS events occurring at the airport #4 are less likely to occur during the DESCENT_PHASE (lower-right quadrant) and contain a mention of the INSTRUMENT_LANDING_SYSTEM (lower-left quadrant) but be characterized by more frequent mentions of the FLIGHT_DIRECTOR and VISUAL_SEPARATION (upper right quadrant).

When interpreting a correspondence analysis plot, one should always keep in mind that some distortions may result from the representation of all the observed relationships into a limited number of dimensions. As a consequence, two items that may be plotted close each other or in a closed angle from the origin may in fact be far apart when considering a third or fourth dimension For example, the examination of how items were distributed on a third dimension allowed us to conclude that while CLIMB_PHASE seems somewhat related to the airport #4, it is in fact negatively related to it and is highly characteristic instead of events occurring at the airport #3. WordStat allows one to display the first three dimensions using either 2D or 3D plots.

Findings -- Correspondence Plots

In the opinion of Provalis Research, one of the most interesting features of correspondence plots for knowledge discovery is its unique ability to identify in very large crosstabulation tables the most singular categories. In the previous example, the airport #1 was readily identified as the airport with the most singular TCAS reports. The Figure 17 below provides another clear example of this property of correspondence plot. This plot represents the relationships between specific aircraft and mentions of parts and devices in FCIR reports. While most aircrafts are grouped together on the right side, one can readily identify an outlier on the left (aircraft #27). We can interpret the distant location of this aircraft from the others by the content category APU, which contains various expressions used to refer to the auxiliary power unit. Although the small size of the database from which this graph was computed prevent us from making a definitive conclusion about the representativeness of observed patterns, it nevertheless indicated quite clearly that reports of events that took place in this aircraft more often contains references to such a device.

This example illustrates another characteristic of correspondence plot: it allows one not only to identify outliers but also the reason why those items differ from the others.

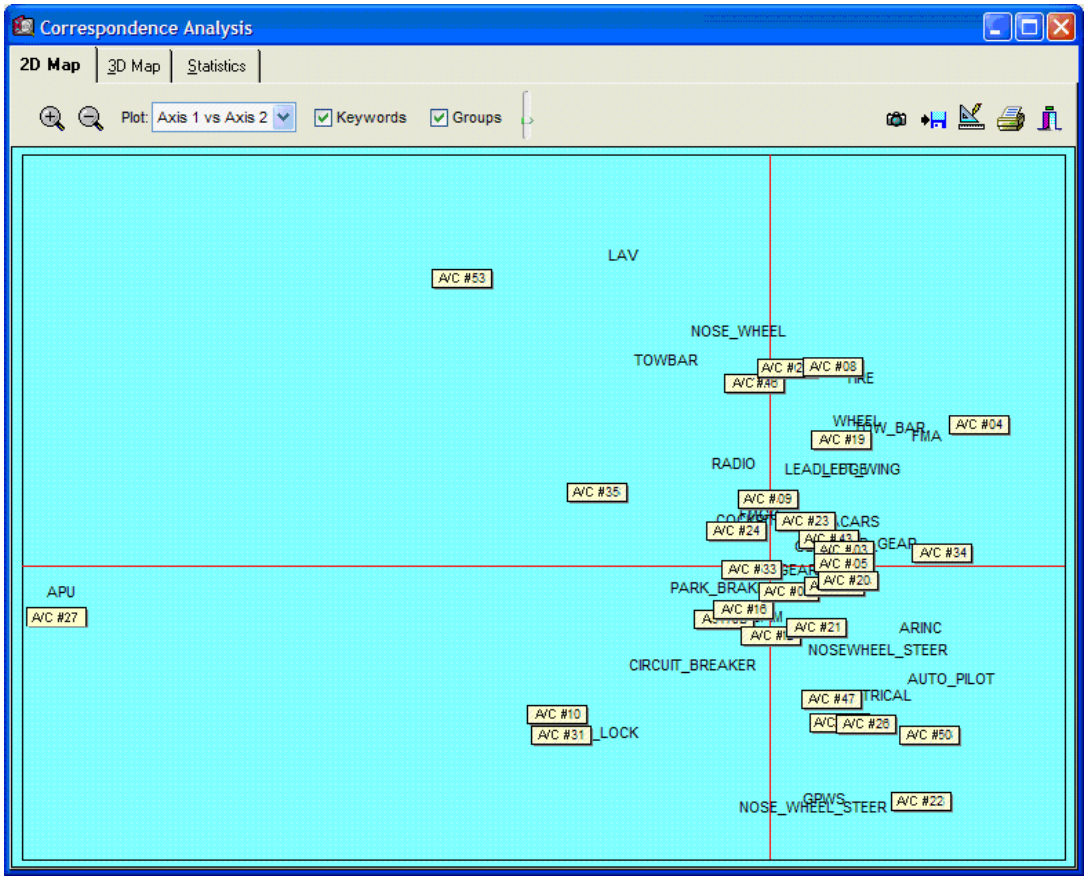


Figure 17. Correspondence plot of aircrafts by airplane parts.

The major drawback of correspondence plot is quite likely the fact that learning how to extract all the relevant information and how to interpret this information takes times. Its interpretation

can sometimes be tricky and should be made with great care, especially when looking at the relationship between classes of a categorical variable and content categories. Also, one should always keep in mind that any representation of a complex reality into a limited number of dimensions always results in some distortion. To prevent any misinterpretation, one should preferably take into account higher dimensions or return to the original cross-frequency table in order to validate his findings. Also, like any statistical tool, the reliability of the representation obtain by correspondence plot remain highly dependent on the size and the representativeness of the data from which it come from. The smaller the size of the database or the less representative the data it contains, the more likely the observed patterns could be caused by chance.

4.0 Assessment of Results by JetBlue

This technology demonstration represented the first time that Provalis Research has performed text mining with airline narratives. The JetBlue flight safety team believes the perspectives offered by Provalis Research and the questions they asked were refreshing and will help JetBlue with future safety analysis. The analysis conducted during the technology demonstration delivered useful and intriguing results. JetBlue found these tools to be useful. The detailed analysis report provided by Provalis Research both validated and expanded on previous JetBlue safety analysis results and identified new safety areas for further study. In addition to the specific analyses provided, it was valuable to have outside experts review and comment on JetBlue's current analytical methods and reporting process. Although the project scope was quite limited, the text mining methods show great potential for providing airline managers with clear, relevant insights that lead to improvements in flight safety, operations and training programs.

4.1 JetBlue Assessment of the Value of Statistical Content Analysis to Flight Safety Analysis

The JetBlue flight safety department had not used any form of text mining analysis prior to this project. The statistical content analysis text mining methods applied in this project show promise as methods that would be valuable supplements to the current safety data analytical methods, tools and techniques at JetBlue. The advanced text mining methods go beyond current capability and represented new ways to analyze information. Based on the (somewhat limited) experience in this project, the JetBlue flight safety department believes the following techniques would be most helpful to the types of safety analysis they perform routinely:

- **Dendrogram cluster analysis.** The attempt to discover new knowledge through unexpected relationships allowed the analysis to confirm the relationship existence. The improvement to the software enabling the ability to click on the cluster and open the associated reports speed up the analyst time reviewing that relationship. The cluster analysis view is direct and easy to interpret.

- **Proximity plots**. These graphical presentations of keyword co-occurrences, those words directly associated with one another were most helpful in identifying the specific issue of TCAS events at airport #2 often involving multiple aircraft.

- **Heat map with dual clustering.** This technique was used on a small number of reports but its application was easily understood and provided useful information. The visualization parameters could be adjusted to graphically display relationships and confirm their association.

JetBlue used some features of the WordStat and SimStat software on other projects outside the scope of this technology demonstration and found two very helpful.

- The **Key Word In Context (KWIC)** technique was very useful in researching de-icing procedures and incidents as the preparation for winter operations began. This technique was helpful in going through a large number of reports outside of those being reviewed and determining which ones were valid in giving us the supporting data needed. As simple as this technique seems, the ability to highlight and identify the key word being researched was valuable in finding relevant reports quickly.

- The **Document Conversion Wizard** was used to convert JetBlue's Internal Evaluation Program assessment and surveillance reports from MS Word. The ability to import this text using the optional variable extraction from structured documents allowed JetBlue's Office of System Safety to view and analyze the data gathered from the evaluations and audits in a more structured and results-driven environment.

The various analysis modules explored in this project identified a number of issues of potential interest to the JetBlue flight safety team. The following issues are examples of information not previously known by the team and therefore represent new knowledge identified by the statistical content analysis techniques:

- Analysis of TCAS events at airport #2 identified terms related to time succession, multiplicity and urgentness. These terms indicate that TCAS events around this airport often involved multiple targets requiring immediate action by the flight crews. The supporting evidence was forwarded to the Flight Operations Quality Assurance (FOQA) department for further study and analysis.

- Analysis of TCAS reports by year showed a decrease in unfavorable words, that is, words associated with "PROXIMITY," "CLEAR OF CONFLICT," and "AVOIDING." This result implies improving resolution of TCAS events through pilot and controller collaboration. (Note: Analysis of the total rate of TCAS reports received per 10,000 flights showed a relatively constant rate over the 24 months of data analyzed.)

This technology demonstration also confirmed some safety issues previously identified by the JetBlue flight safety team. The fact that the statistical content analysis text mining process identified the same issues confirms the validity of the software. Two examples of these confirmations are as follows:

- TCAS events at airport #3 are more common in the take-off and descent phases of flight than at other airports serviced by JetBlue.

- There was a relationship within JetBlue safety reports between the ignition switch and the parking brake. The dendrogram example in Figure 8 pointed out a clear and strong correlation between the two text phrases. A novice aviation analyst might be surprised by this relationship, however, the JetBlue flight safety team had prior knowledge through reviewing incoming reports from the flight crews that the parking brake was inadvertently being set when initiating the engine start sequence with the ignition switch.

4.2 JetBlue Observations on the WordStat and SimStat Mining Tools and Suggestions for Improvement

In applying the WordStat and SimStat tools to aviation safety reports during this project, JetBlue identified a number of areas in which the tools might be improved to facilitate their use in this application. JetBlue's suggestions include:

- Combining WordStat and SimStat to eliminate the need to switch from one tool to the other for various analyses.

- Simplifying the crosstabulation table. Although it does provide numerous statistics to assess the strength of relationships between keywords and variables, it is not easy for a novice user to decipher.

- Enhancing the capability of the correspondence plots. While it does take a grasp to understand and determine what the tool is trying to tell you, it has been identified that it does take some understanding two-dimensional and three-dimensional relationships. Enhancing the ability to click on that textual relationship and drill down into the keyword associations much like the functionality enabled for the dendrogram cluster would be a definite improvement.

- Expanding the capability of clicking on a relationship shown graphically to retrieve the reports underlying the relationship. This capability is currently present in some WordStat and SimStat modules, but is so valuable that is should be added to all of the modules.

- At the time of the study, tools for automatic document classification in WordStat were still under development. We believe such kind of tools could potentially be valuable for the JetBlue aviation safety team to support the manual categorization of events, to allow retrospective categorization of uncoded reports.

5.0 Summary

The purpose of this technology demonstration was to evaluate practical usefulness of statistical text mining tools applied to the analysis of narrative-based airline safety reports. The text mining methods showed promise in improving efficiency by automating portions of the query and output process.

The results proved to be useful to the JetBlue flight safety team for identifying potential safety issues across multiple attributes, such as checklist procedures, flight conflict and resolution procedures, and operational procedures. The project demonstrated that WordStat and SimStat are easy to implement and learn. The array of tools offered was found effective in making accurate and useful knowledge discoveries.

The project clearly demonstrated the value in partnering between airlines and specialist consulting firms. Text mining tools are the next generation in safety data analysis and having an effective tool in the analyst's toolbox will enhance the results in providing for better guidance and direction in improved decision making.

The project demonstrated that significant value could be generated through:

- Informative associated data visualization approaches
- An additional tool to the safety analyst in their analytical methods and techniques toolbox
- Identifying patterns and trends in the operation not previously known
- Automation of repetitive processes
- Efficient use of analyst's time
- Increased level of automation in an analysis project.

It is likely, based on this project, that similar text mining analysis techniques can be effectively applied to information derived from other safety reports in an aviation organization, such as operational delay reports, maintenance logbook entries and maintenance cancellation and delay reports, quality assurance reports, system safety assessments, and internal evaluation program reports, to deliver a more comprehensive picture of overall safety issues.