2005

In search of patterns in incident reports : a syntactic approach

Gaston R. Cangiano
San Jose State University

IN SEARCH OF PATTERNS IN INCIDENT REPORTS:

A SYNTACTIC APPROACH

A Thesis

Presented to

The Faculty of the Department of Graduate Studies and Research

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Interdisciplinary Master of Science

by

Gaston R. Cangiano

May 2005

APPROVED FOR THE DEPARTMENT OF GRADUATE STUDIES & RESEARCH

Dr. Kevin Corker – Industrial and Systems Engineering

Dr. Soteria Svorou – Linguistics and Language Development

Christian Posse, PhD - Statistical and Mathematical Sciences

Battelle Pacific Northwest National Laboratory

APPROVED FOR THE UNIVERSITY

Rhea L. Williamson

ABSTRACT

IN SEARCH OF PATTERNS IN INCIDENT REPORTS: A SYNTACTIC APPROACH

by Gaston R. Cangiano

This research presents a novel methodology for automated analysis of text narratives. Current approaches hinge on statistical analysis of keywords and phrases and to a minor extent syntax. Due to their over-reliance on domain-specific knowledge and their lack of underlying behavioral models, these approaches have significant limitations particularly for problems dealing with human performance.

What is presented here is an approach based on text segmentation utilizing syntactic templates. This segmentation is inspired by linguistic theory and the idea of "mental event" types in narratives. The goal is to contribute with a complementary approach to the current methods, facilitating the detection of behavioral patterns in text. This work contains the method's conceptual formalization together with its computational implementation. Finally the data distribution is examined, raising the hypothesis about the method's potential usefulness. The results are applied to a specific problem in Aviation Safety dealing with Human Performance Errors.

DEDICATION

I would like to dedicate the present work to the memory of my mother, Norma B. Albano, who went to great extents to ensure that her only son would obtain the best education possible. This work is the fruit of that endeavor.

Also, I am forever indebted to my two mentors and committee members, Dr. Kevin Corker and Dr. Soteria Svorou. Dr. Corker not only provided the intellectual support and guidance for this work, but also the logistics and the real-world context for its application. I am forever grateful to him for having guided me through the labyrinth of scientific inquiry and for teaching me how to look at science from a larger perspective. He has been a role model for me in terms of professional intellect and drive, and has been gracious enough to share with me some of his vast experience in the field of Cognitive Engineering. This work is largely inspired by him.

Dr. Svorou has been my mentor as well and, also, my good friend at San Jose State University. I am very grateful to her for listening to my endless complaints during my studies in San Jose and for keeping me good company for the last two years. Her demanding criticism has always guided me toward the highest quality of research. I thank her for being as critical toward my work as she is toward her own. This work is also inspired by her and her willingness to listen to my creative rambling.

ACKNOWLEDGEMENTS

First and foremost, I want to thank Christian Posse of Battelle National Laboratories for his participation and contribution in this work. His perspective and expertise as a statistician was an invaluable asset for critical input. Also, his participation enabled the intellectual exchange with Battelle which is crucial to a large portion of the ideas in this work.

The team at the Human Automation Integration Laboratory (HAIL) at San Jose State University has been a tremendous support for completing this work. I want to thank all of them but specially Samm Starr, the manager at the laboratory for always being available and willing to help out. Her good disposition and no-nonsense attitude are a refreshing oasis in the Silicon Valley's arid milieu. I also want to thank the rest of the crew, Amit Jadhav, Eromi Guneratne, Erik Weber, Martin Howard and Koji Muraoka for their valuable intellectual input and their friendly company. They provided a light and humorous spirit at the laboratory, which was very important throughout my stay and work there.

Finally I want to thank my good old friend David Shields, who provided his invaluable editing skills to help me add the last touches to this work. His well-read mind and keen analytical eye allowed me to catch conceptual and syntactic subtleties otherwise overlooked.

Table of Contents

Table Index

Illustration Index

CHAPTER 1

INTRODUCTION

As the information age begins to mature, we delve into the 21st century with some major challenges. One of these challenges is to harness the explosion of data which technology such as the World Wide Web has made available. In particular, written language has become one of the most important sources of data for the development of new information technology and the theoretical frameworks for its use. Harnessing the power of language through technology would bring substantial benefits to our modern society in the areas of security, productivity and services.

The main difficulty in dealing with natural language is the lack of a consistent theoretical framework for human cognition, which is inseparable from language itself. Linguistic theories for the most part are concerned with capturing the *form* and *structure* of language within a systematic framework, but fail to answer the question about the underlying model of the *user* of the language. That is, the question of a *model* of human goals, processes, resources and constraints. This trend has begun to shift in the last years, due to the advent of *embodied* theories of cognition and language. Embodied theories are "ecological" theories which take into account the inseparability of language, cognition, behavior and the environment. In other words, "ecological" stands for the fact that they conceive meaning and function as a result of the interaction of a cognitive entity with its environment, and the inherent limitations imposed by the nature of these cognitive entities and the environment itself. This is a significant shift away from theories of language which only study language as an isolated phenomenon. Within linguistics, *Cognitive Linguistics* views language as arising from environmental and perceptual primitives, and as sharing a large portion of the same processes and resources as cognition.

It is from this perspective that this work was inspired. The question is how to use this knowledge to design new methodologies that will aid in detecting patterns in text reflecting behavior. The specific problem-at-hand in the work to be presented here is to assist the development of technologies for this purpose. The specific goal is to apply the results of this work to the analysis of Human Errors for Aviation Risk Management. The context of this work is an effort initiated by the National Aeronautics and Space Administration's (NASA) Aviation Safety Program (AvSP) in the year 2000, to enable technologies to reduce the accident rate in the U.S. National Aviation System (NAS). Within this larger effort, one of the projects, Aviation Systems Monitoring and Modeling (ASMM), " *addresses the need to provide decision makers with the tools for safety improvement by identifying and correcting predisposing conditions that could lead to accidents*" (Maille et al., 2004). The work to be described in this thesis is part of an effort to automate the process of data mining and data analysis from a large incident report database: the Aviation Safety Reporting System (ASRS). This database was established in 1975 under a Memorandum of Agreement between the Federal Aviation Administration (FAA) and NASA. The ASRS collects voluntarily submitted incident reports by pilots, air traffic controllers, flight attendants, mechanics, ground personnel, and others participating in aviation operations when they are involved in, or observe, an incident or situation in which aviation safety was compromised. All submissions are voluntary. More than 300,000 reports have been submitted to date, and no reporter's identity has ever been breached by the ASRS. These reports include a "free text" section in the form of *narratives* submitted by the reporters. These narratives provide an exceptionally rich source of information for human factors research. The focus of this project then is to come up with a new methodology to extract information from the free text in the ASRS forms as it relates to human performance errors. This methodology needs to be not only useful in the context of the problem-at-hand, but also complementary to the current efforts underway and technologies available. The goal of the ASMM program is to enable technologies that will be able to identify *precursor events* which act as "signals" of unsafe operational conditions in the NAS. It is part of a larger enterprise in preemptive risk management. One of its components is the free text narratives in the ASRS database. Here is where our effort applies.

Consider the following artificial example, a single story, told in two slightly different ways:

1. "I WAS DRIVING TO L.A. LAST SUMMER IN MY OLD 1972 BMW. PRIOR TO DEPARTURE, I HAD THE ENGINE CHECKED AND THE OIL CHANGED. ON MY WAY THERE THE OIL INDICATOR WENT OFF. NONETHELESS I CONTINUED ON MY WAY. ABOUT 40 MI FURTHER MY ENGINE SEIZED. LATER THE MECHANIC INFORMED ME THAT THE SERVICEMEN AT JIFFY LUBE HAD LEFT THE OIL CAP OFF! **I JUST CAN'T BELIEVE THIS PEOPLE CAN BE SO INCOMPETENT! I FILED A COMPLAINT WITH THE BETTER BUSSINESS BUREAU. THEY SHOULD PAY FOR THE DAMAGE...**"

2. "I WAS DRIVING TO L.A. LAST SUMMER IN MY OLD 1972 BMW. PRIOR TO DEPARTURE I HAD THE ENGINE CHECKED AND THE OIL CHANGED. ON MY WAY THERE THE OIL INDICATOR WENT OFF. I **HAD JUST CHANGED THE OIL, AND PLUS THIS KIND OF THING ALWAYS HAPPENS TO MY OLD CAR, SO I CONTINUED MY TRIP**. ABOUT 40 MI FURTHER MY ENGINE SEIZED. AS I FOUND OUT LATER, THE SERVICEMEN AT JIFFY LUBE HAD LEFT THE OIL CAP OFF! MORAL: I'LL NEVER GO BACK TO THAT SHOP AGAIN."

The differences in the two stories are shown in bold. In the first one, the narrative is concluded by a single continuous "subjective" block, where the writer expresses his personal opinion on the incident, and makes suggestions as to what actions should be taken. It is highly charged emotionally, and also offers prescriptive solutions. This stands in contrast to the rest of the narrative which is more factual, that is, strictly narrating the development of events. Therefore it can be considered *extraneous* to the incident per se, and it is important to discriminate it. The second version only has a single sentence at the end with such "subjective" level. On the other hand it presents a sentence in the middle of the narrative development (in bold) which reveals the presence of knowledge extrinsic to the development of the incident. The difference here is that this knowledge is *interleaved* within the narrative development. It portrays knowledge in the mind of the speaker which could have been present at the time of the incident, and thus reveals an aspect of the situation awareness of the operator at the time. In this case it turns out to be incorrect information. From the point of view of error analysis both narratives have the same source of error, namely, the servicemen at the oil exchange shop. Nonetheless an argument can be made for the case that it was the driver's responsibility to check the oil even though he knew that it could have been a false alarm. Therefore this incident could be considered a *representational* error on the part of the operator, as opposed to an *external* cause, since it should have been detected by him. Representational here means that the operator had the wrong situation awareness of his environment at the time. This concept will be further developed in the coming chapters. Syntactically, this one piece of information stands out due to a shift in the temporal anchoring of that sentence with respect to the previous sentences, namely, the shift to present tense. It is a very salient grammatical signal considering the overall temporal structure of the narrative (mainly past). This is the type of "signals" that we will addressing in this work. Our approach will capture these temporal deviations in a quantitative and systematic manner. Our hypothesis will be that this information can be used to better analyze and classify errors from the ASRS incident report database.

In summary, our effort is to tackle a specific data mining problem in the aviation domain, by providing a new analytical and computational tool. Our the hope is that it will help experts gain further understanding about the current behavioral models employed, and to improve the existing error taxonomies. The ultimate goal is in contributing to make the National Airspace safer.

Problem Statement

The main problem to be addressed in this thesis consists of providing a new approach to aid in the detection of "signals" in the narrative text of the ASRS database. These signals are indicators of precursor events to incidents in aviation. In order to do so, we need to evaluate the current technologies and methods for extracting information from text. Also we need to look at the current models of human performance that are being applied to this problem. Finally, we need to come up with a novel approach that could serve both as a complement and augmentation to the current effort underway in this area, both technologically and conceptually. Therefore, our desiderata is

- an new approach to text mining geared for the behavioral model under study, in this case *Situation Awareness Loss;*
- a tool that will complement the current technologies available;
- a hypothesis about how to use the tool to better understand the behavioral model being used, as described in Maille et al. (Maille et al., 2004).

Limitations

The main limitation is with regards to the statistical validity of source data. This has been discussed widely in the literature, since it is a very well known problem with data of this type. The following excerpt was obtained from a compilation of selected narratives that NASA makes available for research purposes in the ASRS website. It clearly explains the limitations that the database and ASRS form structure impose on studies of this data.

> *Certain caveats apply to the use of ASRS statistical data. All ASRS reports are voluntarily submitted, and thus cannot be considered a measured random sample of the full population of like events. For example, we receive several thousand altitude deviation reports each year. This number may comprise over half of all the altitude deviations that occur, or it may be just a small fraction of total occurrences. We have no way of knowing which. Moreover, not all pilots, controllers, air carriers, or other participants in the aviation system, are equally aware of the ASRS or equally willing to report to us. Thus, the data reflect reporting biases. These biases, which are not fully known or measurable, distort ASRS statistics. A safety problem such as near midair collisions (NMACs) may appear to be more highly concentrated in area "A" than area "B" simply because the airmen who operate in area "A" are more supportive of the ASRS program and more inclined to report to us should an NMAC occur.*
>
> *Only one thing can be known for sure from ASRS statistics –they represent the lower measure of the true number of such events that are occurring. For example, if ASRS receives 300 reports of track deviations in 1993 (this number is purely hypothetical), then it can be known with certainty that at least 300 such events have occurred in 1993.*
>
> *Because of these statistical limitations, we believe that the real power of ASRS lies in the report narratives. Here pilots, controllers, and others, tell us about aviation safety incidents and situations in detail. They explain what happened, and more*

> *importantly, why it happened. Using report narratives effectively requires an extra measure of study, the knowledge derived is well worth the added effort.(NASA-FAA, 2003)*

Delimitations

The study will be delimited to a set of narratives obtained from Battelle Pacific Northwest Division (Battelle-PNWD). Battelle is a national research contractor working for NASA under the AvSP/ASMM project. The narrative set will be a subset of the sample pool of ASRS records employed by Battelle in their portion of the research. This decision was made in order to make the validation of this work possible. The results originating from our work will be compared to Battelle's in order to draw inferences and discussion.

Therefore our work will be delimited to the sampling criterion imposed by Battelle and NASA, which is to limit the selection of records to specific anomalies as classified by their experts.

Assumptions

The primary assumptions for this work are theoretical in nature. We are assuming that the authors of the narratives were making a serious effort to be as accurate as possible, given memory limitations, in describing the events that occurred during the incident. This applies to the order, nature and details of events described. It also applies to the degree of "objectivity" used in the narration process. The first assumption is fairly safe, in that it has been employed substantially in the past for experiments in Clinical Psychology, Content Analysis, Sociology and Mass Communication Studies. The collected results from the usage of human language as data can be seen in the literature (Erisson & Simon, 1993; Johnson-Laird, 1983; Genter & Stevens, 1983; Popping, 2000; Kintsch, 1992; Givon, 1995). In particular, the most important review of this type of work was done by Ericsson and Simon (1993). Their work on *Protocol Analysis* states the sufficient validity of this same assumption.

The next theoretical assumptions stems from the use of linguistic theory. The particular theory that we will apply here has been labeled situation types, and has been established as a foundation for language semantics. It is an integral part of academic curricula nowadays and is mentioned as an important linguistic element in the Natural Language Processing (NLP) literature (Jurafsky & Martin, 2000). We will make the assumption then, that the linguistic phenomena to be employed for this analysis overlaps significantly with the equivalent mental processes in cognition. Specifically, we will assume that situation types have a counterpart in the *quality* of the mental representations they correspond to. We will modify these situation types to create a correlation with the type of *events* and *states* that are relevant to the error analysis model – called the "Scenario model" (Maille et al., 2004) to be described in Chapter 3. We will assume that the process of *temporal integration* of mental events in memory, corresponds to a similar process occurring during narrative production and is manifested via grammar. This is in agreement with the current theoretical position in Cognitive Linguistics. We will describe this idea in more detail in Chapter 3.

Definition of Terms

Ecological. This is a term that emerged from the "cognitive revolution" in contemporary American psychology. It is a functional approach to cognition that sees it as an active process of interaction between a cognitive agent, its environment and the resulting constraints from the inherent characteristics of both, rather than a static process in the mind of the agent (cognitive psychology) or as a result of its actions (behaviorism). Its origins trace back to work in biology by David Marr (Marr, 1975), Maturana & Varela (Maturana, 1970), and in psychology by Gibson (Gibson, 1977). It is very prevalent today in the field of Cognitive Engineering (Flach, 1995).

Situation Awareness. A term used in the Aviation Human Factors community to describe a descriptive model of human performance, which aims at explaining errors produced by cognitive agents in a complex system. Situation Awareness is " *the perception of the elements in the environment within a volume of time and space,* *the comprehension of their meaning and the projection of their status in the near future"* (Endsley, 1995).

Mental Model. A mental representation on knowledge about the world in the mind of a cognitive agent. The exact characteristics and extent of these models is far from known to date, but it is commonly agreed that they exist (as a useful construct for study) and that they play a key role in comprehension, prediction, and production of cognition. They originate from work by psychologist Kenneth Craik (1943, see Johnson-Laird, 1983). They contain the *functional* and *representational* information necessary for operating in the world. It is not known to what extent they are isomorphic to external phenomena, or if they are instead "cognitively efficient" transformations of it. More on this in the next chapter.

Informational content in language. Language, being an extremely rich source of data, can be analyzed at several levels of abstraction, specificity and granularity (or resolution). In particular, language can be studied as a *vehicle* or as a *container* of information. As a container, it simply carries the meaning of a message to be decoded by a listener, with all of its information residing in the lexical meaning of the words. As a vehicle, language provides a set of functional devices and parameters *to be set* by the speaker which alter the intent, impact and success of the communication process. Our method will study language as a vehicle. We will refer to this as *latent* content, as cited by psychologist Charles Osgood in Pool Sola's*Trends in content analysis (Pool, 1959).* The term latent refers to seemingly hidden information (largely phenomenological) that resides in the structure and form distribution of language. An example of this is the detection of genre and authorship in text. The reader can recognize the genre of a given text, its author and talk about the storyline, but he is not able to describe the linguistic characteristics in the text that make it different from others. In this sense, the grammatical style and the distribution of syntactic features play the role of conveying the message in a different manner than other genres or authors, and thus acts as the vehicle for the communication act. More on this in Chapter 3.

Segment. A section of text (a syntactic clause) which has been identified and tagged as corresponding to one of our (to be defined) variables stemming from Situation Type theory. A segment corresponds to an Extensible Markup Language (XML) tag, containing the clause contents as produced by our implementation program and written out to a plain text file on a personal computer.

Style. A given manner of portraying an incident, in terms of the *temporal distribution* of the events presented, and the *quality* of each of these events as they are presented to the reader. Temporal distribution will be operationalized in this document as the distribution of segment types across narratives. Quality will be formalized in terms of situation types (Chapter 3).

Syntactic Template. A feature structure, or aggregate of attributes corresponding to lexical items (words) and/or syntactic constructions (phrases, clauses or sentences). A template could also be a composition of many smaller templates; in this case, it represents a given situation type, and it is the construct that gets matched against each syntactic clause to produce segments. In our computational implementation, they correspond to software data structures, which are populated dynamically by the program from a lexical database called WordNet (Miller et al., 1990) and statically from a set of preconceived situation types. An example of a lexical template for a noun follows

```
[ N            String ]
[ Volit          +/-  ]
[ Person         +/-  ]
[ Proper         +/-  ]
```

A noun template contains a string with the actual word, and a set of binary attributes indicating the quality of the noun in terms of it being a volitional agent (e.g., a pilot is a volitional agent versus a control instrument which is not).

Narrative timelines. We will treat each narrative as possessing a rich set of time structures. *Speech Time* (Sp) is the time anchored at the moment of production of the text. *Reference Time* (Rt) is the internal (linear) timeline of the narratives. *Situation Time* (St) is internal to an event, anchored with respect to Rt. These distinctions were obtained from current work in Discourse Theory (Smith, 2003). These distinctions will allow us to detect deviations from the temporal structure of narratives via grammar.

Importance of the Study

The importance of this study stems from the limited set of available tools and methods for investigating behavioral patterns emerging from language. In the context of the AvSP project, such limitation is evident by the reliance on a single approach to the problem, which is based on a generalized statistical analysis of keywords and word patterns. This is the predominant approach in the field of Information Extraction (IE), known widely as the "bag of words" approach. Even though it is a very powerful approach, it is best suited for tasks where a significant portion of *a priori* knowledge about the problem is available. Therefore, it has been very successful in obtaining patterns of data from language where those patterns are derived from domain-specific knowledge. Its limitations also lie in its lack of explanatory power for cases of human behavior analysis. We claim that it is feasible to devise robust complementary methods for information extraction geared specifically to the problem-at-hand: methods based on general models of human memory and not focused on keywords (i.e., domain-independent). There is a body of knowledge available which will allow us to apply general models of language and cognition together with technology. This work will demonstrate the potential of one such method.

The idea of analyzing *style* in text has been exploited as a means for enhancing current methods in Information Extraction and Retrieval (Karlgren, 1999).

Similarly here, we will introduce an augmentative methodology to the current bag of words approach. Our approach analyzes the distribution of syntactic characteristics in text. We will show how this approach yields categories of syntactic style, and how we can identify *saliency* within these categories and relate them to the behavioral problem-at-hand in studying human error. Therefore the practical importance of our effort is to provide a computationally feasible complement to the existing tools currently applied to the ASMM portion of the AvSP project. This increases the significance of the contribution we are making both to the field and to the project itself.

CHAPTER 2

REVIEW OF LITERATURE

Our review of the literature will necessarily consist of several areas. This is so due to the interdisciplinary nature of the problem we are tackling and the solution we are proposing. It is naturally adequate for an interdisciplinary research thesis. These three areas are:

1. Human Cognitive Models/Situation Awareness/Error taxonomies
2. Cognitive Linguistics/Discourse Theory
3. Information Extraction Technologies and Methods

Choosing a Topic

As described briefly before, the main context of the AvSP is to reduce the accident rate in the U.S. NAS. Within the AvSP, the ASMM focus is on devising tools and technologies for enabling the identification of patterns that could lead to accidents. Therefore the main goal in the ASMM is the design of automated analytical tools and methods for *pattern detection.* Our main intention within this effort was to focus exclusively on one type of data in a larger set, namely the incident reports. Furthermore we are analyzing only the narrative portion of these reports. Effectively this delimitation constrains our problem to one of *Discourse Analysis,* with the advantage that we are trying to capture phenomena related only to a specific domain and to a specific human performance model. Otherwise the task of discourse analysis remains largely an area of basic research in linguistics.

From the analytical standpoint, the portion of the problem we are addressing here refers to what is known in the field of Human Factors as Human Performance Modeling, and within this, to the analysis of *error taxonomies* in the aviation domain. Further narrowing our scope, we are targeting a very specific error type known as *Situation Awareness Loss (SA-loss).* SA-loss refers to human performance errors arising from the degradation of a person's perception of the current state of the environment around him. In other words, at this stage of its development, SA tries to capture with the use of a descriptive model errors arising from an operator's*misperception* or *misrepresentation* of the situation around him. This is significant in light of the high complexity of the NAS and modern aviation technology. SA-loss is a very prevalent type of error in the aviation domain, as we will see from the literature review. It has been claimed to be the primary source of errors.

In sum, our topic is to help with the analysis of the current SA error taxonomies, and in particular to gain more insight into the distinction between misperception and misrepresentation as represented in the narratives that comprise the ASRS. Therefore, our review starts on this subject, and then moves onto other areas which are necessary to devise a methodology for addressing this problem, together with an implementation. This entails looking at linguistic theory (Discourse and Semantics) and current available technologies for text mining. The next section will proceed in this same order.

Selecting Appropriate Articles

The material we will review in the next section consists mainly of technical articles and to a lesser extent of books. This is due to the nature of the knowledge sources for SA, which is a moving target. There is no definite model of SA, but several contending views, so the best choice at the moment is a descriptive model of the phenomenon. A similar assessment can be made for error taxonomies: *"... there appears to be as many taxonomic schemes as there are people interested in the topic"* (Shappell & Wiegmann, 1997). The review of theory on the other hand will make use of books, since we are going to be relying on established work.

Reviewing the Articles

Human Cognitive Models

In order to understand the principled motivation behind the logic of our review, some models are in order. The main conceptual model behind the study of human performance errors of the type to be described here originates from the perception-action cycle proposed by Niesser (Niesser, 1976). In this framework (see Illustration 1 below), cognition is seen as a knowledge-driven, prediction-oriented, cyclical and dynamic phenomenon; this stands opposed to the classic Human Information Processor model (Card et al., 1983), which is data-driven and passive. The performance of experts in real life seems to exceed the limits and capabilities of human performance as described by the conventional static model.

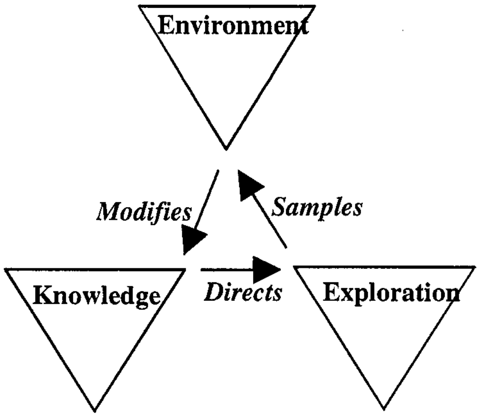


Illustration 1. Active Cognition

This *active* framework, instead, models human performance as a cycle where information processing is guided by expectation and the updating of mental representations about the environment. This type of behavior is akin to what's been called " *recognition-primed decision-making*" (Klein, 1989) used in the context of studying higher decision-making processes. What this means, conceptually, is that operators respond via fast activation of stored procedures triggered by cue pattern recognition, which enhances performance beyond that of non-experts via a *cognitive efficient* mechanism highly tuned by experience.

For the sake of the AvSP study, a simplified and linear version of this model has been adopted. The model consists of five levels of processing, occurring cyclically and in sequence, but not necessarily in strict ordering:

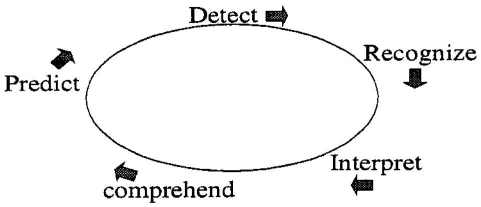


Illustration 2. DRICP Cycle

This has been labeled the *Detection Recognition Interpretation Comprehension and Prediction* (DRICP) cycle (Maille et al., 2004), and it is defined as follows:

> *Detection is the act of discovering, discerning, or capturing attention as this is related to the existence, presence, or fact of an event...*
> *Recognition is the act of relating a detected event* ***e*** *to a class or type of event that has been perceived before...*
> *Interpretation is the act of relating a specific event type to a*

> *network of actual and possible events of various other types...*
> *Comprehension is the act of perceiving the significance of an event....*
> *Prediction is the act of forecasting what will happen in the near future...(Maille et al., 2004)*

The motivation behind the usage of this model is to be able to operationalize the most salient cognitive events that could possibly be detected in an incident. In other words, to at least be able to discriminate some basic levels of human cognitive performance from the data.

Situation Awareness

What is the relation between SA and misrepresentation? Misrepresentation is a term used in the Human Factors community for a type of error similar to that of SA-loss. It has been equated to SA-loss and employed for error taxonomies used in aviation risk management studies (Carmino et al., 1990). This was the question that set off our research in the literature. To try to answer this question it is necessary first to look at the existing models of SA, the related models of memory, and the error taxonomies that derive from them.

Endsley's model of SA (Endsley 1995, 2000), shown inIllustration 3 below, considers SA as a separate component from decision-making and action in the information processing chain described in the previous section, but with important effects on these. There is an explicit distinction made in this model:

that of SA *products* versus SA *processes.* Endsley makes a distinction between SA as a construct, and the processes that create it and update it, namely *situation assessment.* The notion of a clear distinction between behavior and knowledge is questionable for not being psychologically tenable (Adams et al., 1995). At the same time this separation is against the ecological framework that has emerged in the recent years in the aviation and Human Factors community (Hancock & Smith, 1995). According to this framework, behavior and information lie at the junction of the interaction between agent and environment. In other words, it is the *adaptation* of an agent to the constraints present in the environment and the goals of the agent, that behavior and information (as they relate to performance) arise. Therefore the distinction of process versus product does not make sense from that perspective. Others in the literature have defined the concept of *state interpretation* as the process of forming and updating state knowledge about a dynamic system (Baxter, 1999), which also bears the same distinction between process (interpretation) and product (state).

Furthermore Endsley (Endsley, 2000) defines SA as a *state of knowledge* about a system, which implies that this is only a partial representation of all the system's variables and components in the operator's cognitive reach. External factors contributing to SA-related performance according to this model are categorized into four: stress/workload, system interface, preconceptions and objectives and level of expertise/skills. These correspond partially to those in the taxonomy proposed by Rasmussen (Rasmussen, 1982), where he classifies these categories as "*causes of human malfunction*", "*situation factors*", "*personnel task*" and "*performance-shaping factors.*"

Endsley's(Endsley, 1995) definition of situation awareness ties the notions of time, integration of information and prediction. This model implies that there must be a collection – or memory history – of events forming the basis for the predictive outcome of the operator. The notion of space in her model pertains to the establishment of functional relationships among elements in SA, which therefore have a potential effect on modulating attention, scanning patterns and assigning relevance weights to elements in the environment at any given point in time.

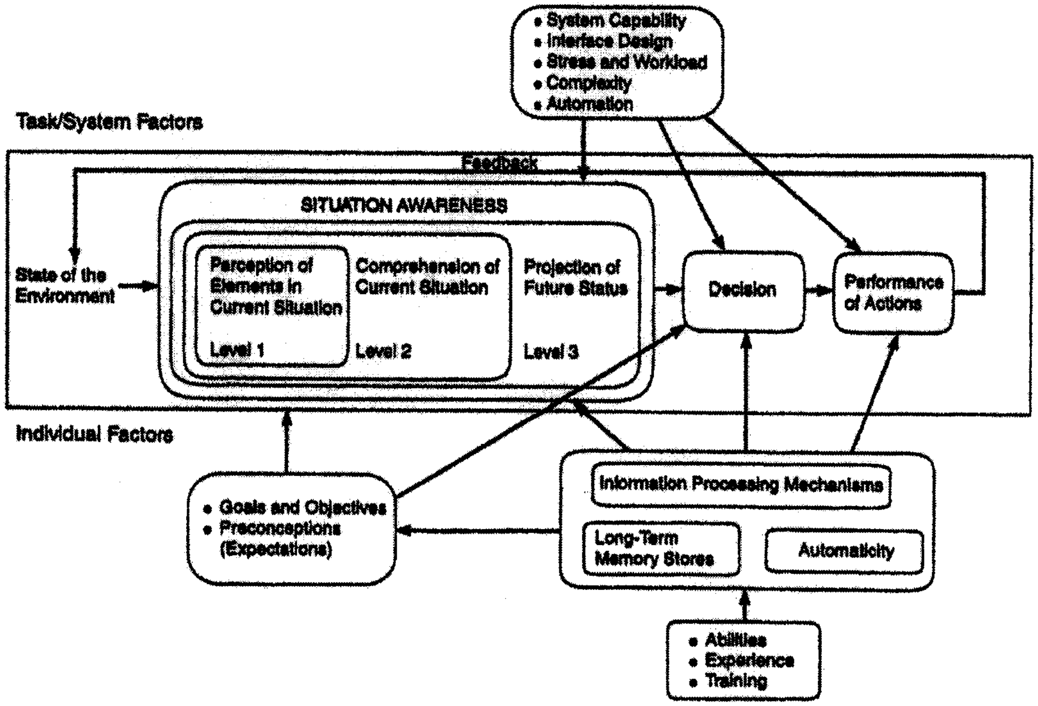


Illustration 3. Endsley Model of SA

Another descriptive model, albeit less elaborated, is that of Adams et al.(1995), which is based on Neisser's(Niesser, 1976) perception-action cycle. It has been suggested (Hancock & Smith, 1995) that this cycle corresponds roughly to the three levels specified by Endsley's model. This model is shown inIllustration 4 below. This model expresses an interesting relationship between the so called *representational schemas* in focus (active) memory and perception. It is probably more relevant for explaining relation between misrepresentation errors and SA errors.

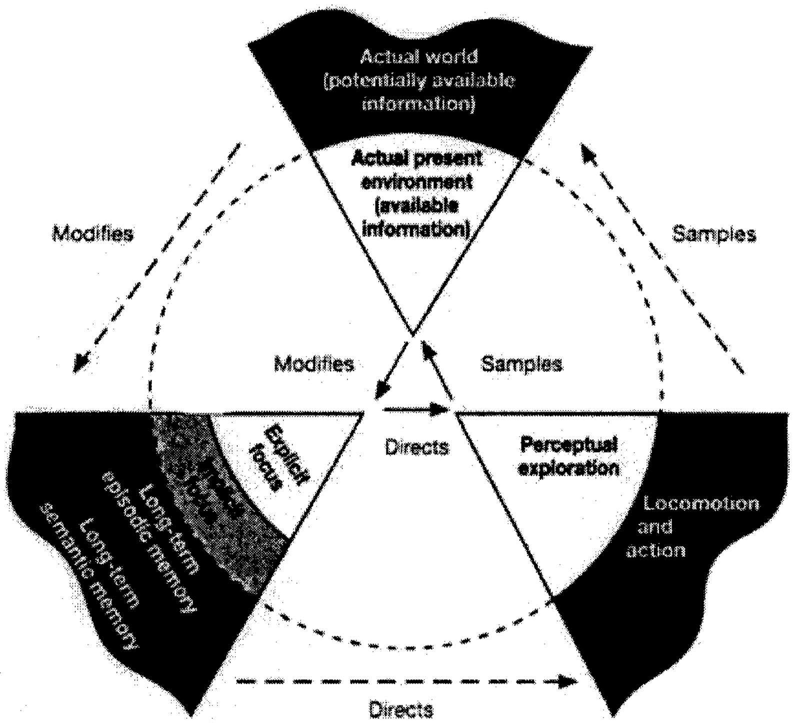


Illustration 4. Extended Neisser Model

An important aspect of this model is that it expresses a distinction between the *types* of memory involved in performance. The authors based their model of memory on work done on language comprehension. What they propose is that the type of memory involved in complex systems control (as with text comprehension) requires an ability to overcome short-term memory and attention limitations by setting up *situation* (or context) structures on-line. These structures are accessible in a significantly shorter time, and are cognitively feasible according to our knowledge of human memory limitations in bandwidth. It is postulated that these situation structures are fundamental in the decision-making process and come to a much lower cost in terms of cognitive resources. They also possess a larger scope in the amount of stored information they can access. In other words, these structures are present in *focus* and *episodic* memory, linked together by topic relevance and processing proximity. A situation structure allows operators to perform efficiently in the face of complex tasks and also have access to larger set of information about their environment. This is in agreement with the theory of *Long-Term Working Memory* (LTWM) proposed by Kintsch and Ericsson (Kintch & Ericsson, 1995), which is also based on studies of text comprehension. According to this theory, short-term memory limitations are overcome by setting up retrieval cues (equivalent to situation structures) for fast access to relevant information stored in long-term memory. Adams et al's (1995) version of memory is shown inIllustration 5 below.

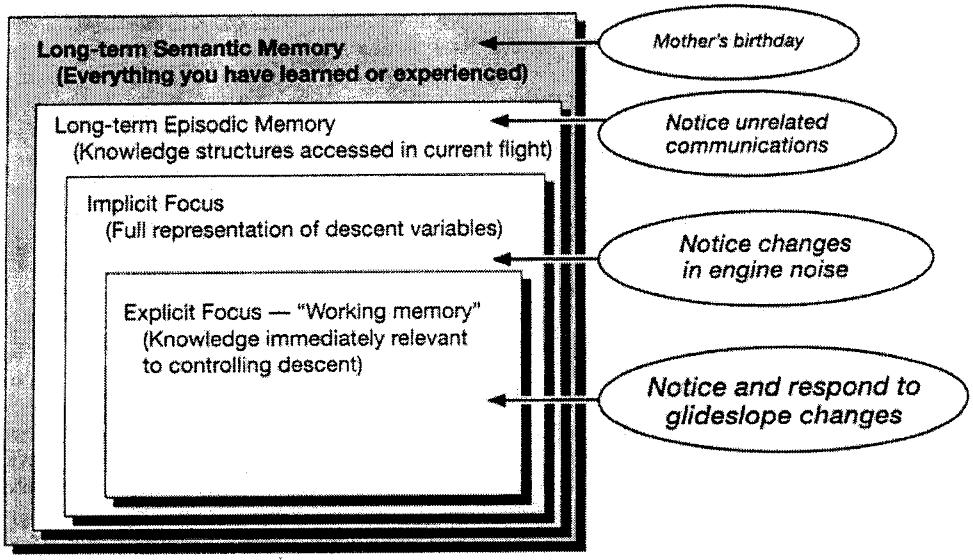

Long-term Semantic Memory
(Everything you have learned or experienced)
Long-term Episodic Memory
(Knowledge structures accessed in current flight)
Implicit Focus
(Full representation of descent variables)
Explicit Focus — "Working memory"
(Knowledge immediately relevant
to controlling descent)
Mother's birthday
Notice unrelated communications
Notice changes in engine noise
Notice and respond to glideslope changes

Illustration 5. Active Memory Model

A *situation model* as defined by Endsley (2000) is a dynamic representation of the operator's knowledge and understanding of the state of a system at any given time. This includes observations of states across time used in order to make predictions. She argues that situation models could be understood as *instantiations* of mental models at a given point in time. She fails to explain two things: one is what kinds of models are these which are accessible from short-term memory, and the other one is how to conceptually explain certain type of SA-loss errors (level 2 primarily) when the product and processes of SA are separated. The relationship of mental models to the situation awareness model is shown in Illustration 6 (from Endsley, 2000).

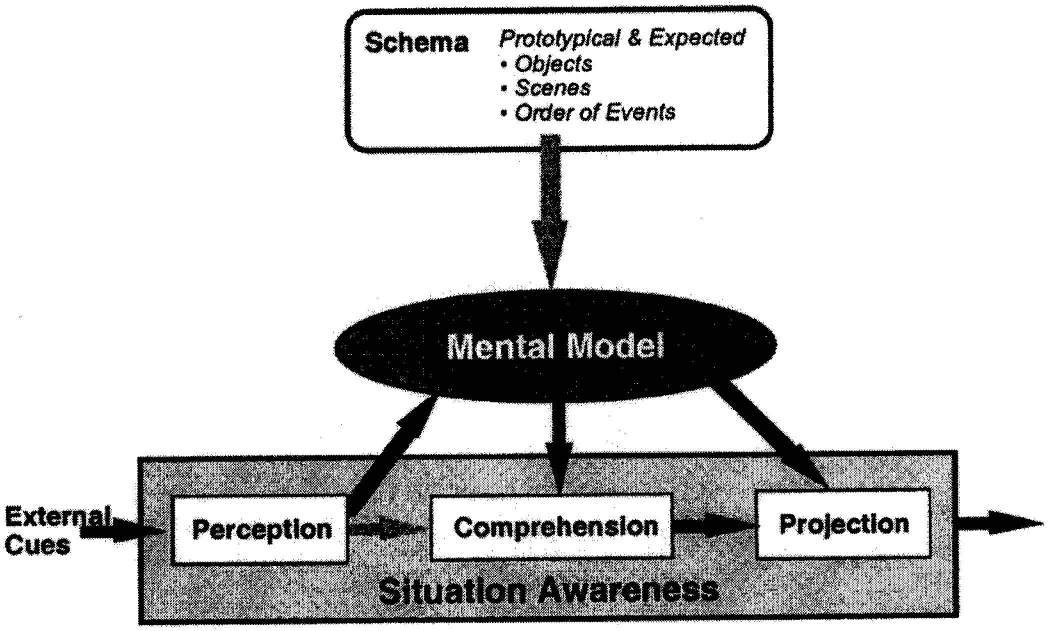


Illustration 6. Situation Model

The main thing to note first from the above diagram is the confounding of process and products; here we have processes such as perception, comprehension and projections included within the "SA box", which is supposedly only a product of the behavior of achieving SA; secondly, the relationship between perception and mental models is unidirectional, indicating no effects of mental structures on guiding the perceptual processes. There seems to be some contradiction in what Endsley proposes and the diagrams of her model (perhaps the diagrams are poor). She acknowledges the influence of mental models on perception (i.e. expectation), and existence of perceptual "biases". The diagrams of her models are not reflective of this. According to this model, all incorrect updates of mental models have only perceptual mechanism as the plausible causal factors, so that all SA-loss error are at some point due to some external factor. The same applies to the operational definition of mental model and situation model: what drives the instantiation of a situation model from a mental model if not influenced by perception? There needs to be an element from perception influencing the selection of a given situation model based on the idea of pattern detection, in the way of the recognition-primed decision-making paradigm(Klein, 1989). This is the more substantiated approach in the literature for explaining skilled behavior.

The common limitation of the two previous models is that they talk about schemas and mental models as static representations of systems in the mind of operators, which are called upon during performance. Mental models are responsible for producing expectations, plans, predictions and guiding attention in the case of Endsley's model. In the case of Adams et al (in Barlett's sense, 1932, as cited in Adams et al, 1995) it is schemas that are seen as governing the mental flow of events. This last model does distinguish between schemas in general and active schemas, which are referred to as contexts or conceptual frames (a similar nomenclature used also by Endsley). Neither one of these two models puts any effort into explaining how contextual models are derived from static models, an area that has potential to explain certain aspect of SA-loss error. Furthermore, the concept of situation model and situation awareness is equated in Endsley's model, confounding aspects of process versus product again. In this view then the situation model (or SA) is a set of variables representing the perceived state of a system at a given time. It is the mental model that has the functional knowledge to make inferences in decision-making and prediction. Therefore it leaves SA out of the direct equation of dynamic behavior, and only considered as an input to it (see illustration 7 below).

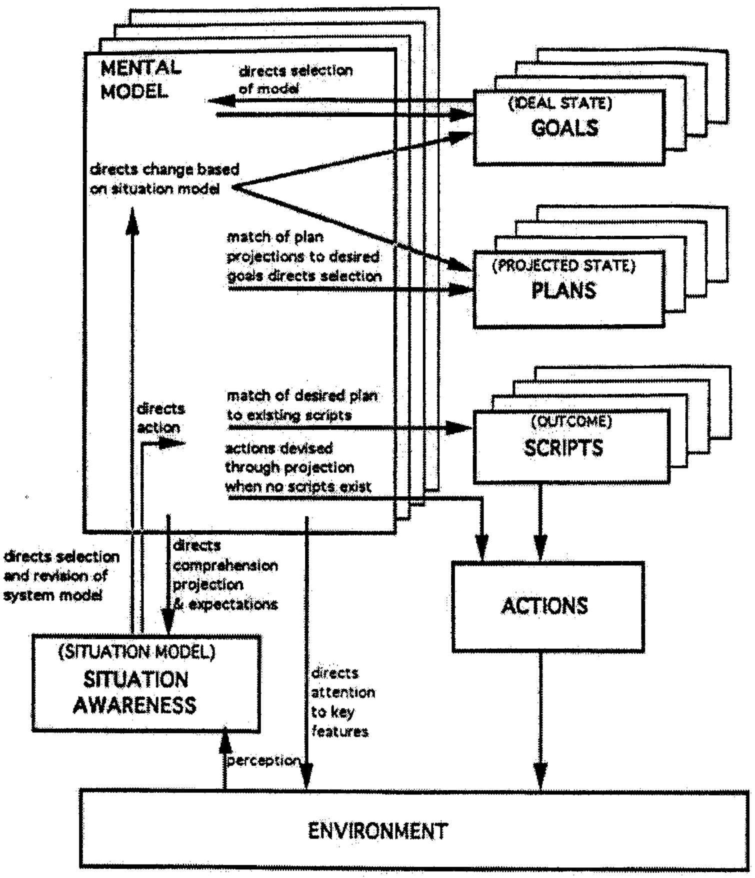

MENTAL MODEL
directs selection of model
(IDEAL STATE)
GOALS
directs change based on situation model
match of plan projections to desired goals directs selection
(PROJECTED STATE)
PLANS
directs action
match of desired plan to existing scripts
(OUTCOME)
SCRIPTS
actions devised through projection when no scripts exist
directs selection and revision of system model
directs comprehension projection & expectations
ACTIONS
(SITUATION MODEL)
SITUATION AWARENESS
directs attention to key features
perception
ENVIRONMENT

Illustration 7. Complete Endsley Model of SA

We argue that the influence between SA, processes and decision-making is subtly different from what Endsley's model describes, and closer to Adams et al version, based on Neisser's (1976) and Sanford & Garrod's models (Sanford & Garrod, 1981, as cited in Adams et al, 1995). This subtle difference is important for our review on the relation between misrepresentation and SA. We derived this analysis from the following theoretical work. There is an operational difference between Endsley's usage of mental model (Endsley, 1995) and the concept of mental models as studied in psychology (Wilson & Rutherford, 1989). The predominant sense employed in the Human Factors literature is that of a functional and *static* structure, stored in long-term memory. One definition of a mental model is that of a "*mechanism whereby humans are able to generate descriptions of system purpose and form, explanations of system functioning observed system states, and predictions of future system states*" (Rouse and Morris, 1986, as cited in Wilson & Rutherford,1989, and Endsley, 2000). Rasmussen (1986,1994) defines mental models as "*internal representations of environmental factors that determine the interrelationships among observable environmental data.*" The common ground among these definitions is their *functional* nature, or rather how they serve mainly to derive functional characteristic about a system.

On the other hand, the psychology-oriented version of mental models sees them as structures built *on-line* (in short-term memory), and subject to real-time contextual processing constraints. They can, therefore, be considered to become optimized through experience in order to achieve efficiency of functionality and encoding. Johnson-Laird's theory of mental models(Johnson-Laird, 1983), defines them as multiple in nature, capable of representing spatial relations, events and processes, and the operation of complex systems (Johnson-Laird, 2001). Furthermore his theory defines mental models as structures that yield inferential reasoning and more importantly representing possibilities and plausibilities, which are key factors for operating under uncertainty. In other words, mental models in this sense are seen as a fundamental part of what could be considered the "logic of the mind." It has also been proposed in the literature that these mental models are *homomorphic* systems, that is, reduced representations of complete systems, obtained through experience and training by way of many-to-one mapping reductions (Holland et al., 1986). These subsystems are usually productive and efficient for a given context, but fail during abnormal circumstances and produce *cognitive lockups (Moray, 1987).* This stands in opposition to the previous description of a mental model, where it was assumed to be a complete and static representations of the knowledge an operator has about a system or part of a system. Memory limitations alone would suffice to yield the isomorphic version unfeasible; furthermore, the homomorphic notion is aligned with a more general view in human performance of the *law of least resistance* and *model transformations* (Rasmussen, 1982). It is still not clear as to what are the selective constraints or processes which guide the formation of these homomorphs. The task of classifying them could range from using methods for system decomposition based on information theory (based on Information Theory, as cited by Moray, 1987), or in more qualitative terms using an abstraction hierarchy as the one proposed by Rasmussen (1986). What is almost certain, taking the perceptive that skilled performance requires pattern recognition behavior, is that perception has a more complex role to play in the SA loop. In this sense, perception processes both update mental models and are influenced by them (modulated), due to the prior set up of expectations which enable fast pattern recognition. This is the difference between recognition and perception per se, which are both perception-based processes.

In sum, what is suggested here is twofold: On the one hand, the predominant definition of mental models used in SA studies is not compatible with established notions of memory constraints on performance and on the idea of expectation-driven perception; on the other hand, separating knowledge from behavior as Endsley proposes, not only is not aligned with the current ecological notions of human performance, but it also poses a problem in discriminating between causal mechanisms of error in SA. Therefore it is important to highlight the caveats in the current models of SA before proceeding to create a new method for extracting data from incident reports. In this way we can target the design of our approach in order to assist with the improvement of SA error taxonomies. We will contribute to this enterprise by providing a computational tool that improves the analysis of incident narratives, from the perspective that mental models act as *modulators* of perception, in a recognition-driven loop. Our tool should allow analysts to better understand the relation of SA to misrepresentation by assisting in answering the following question: Where can we draw the line between an external factor such as misperception versus poor representation (i.e. misrepresentation) as a causal mechanism in SA-loss errors? In other words, are all SA-loss errors perceptual at some point? The taxonomies seem to indicate that. This is not a question that any of the current tools or descriptive models can help to answer. We can contribute to eliminate the disproportions, or pockets, present in the current taxonomies by providing assistance in answering this question. These "pockets" are single categories (such as misrepresentation or perceptual-level-1 errors) containing a large percentage of the overall number of errors (about 70%) in the taxonomies. The next section will discuss this in more detail. We will design our new method for data mining of incident report narratives with this goal. Our new approach should allow analysts to better discriminate between perception-driven and representation-driven SA-loss errors. The rethinking will be presented in the final chapter entitled "Commentary and Future Directions for Research".

A hypothetical revised model of SA for the sake of our analysis and design is shown in the following graph (Illustration 8).

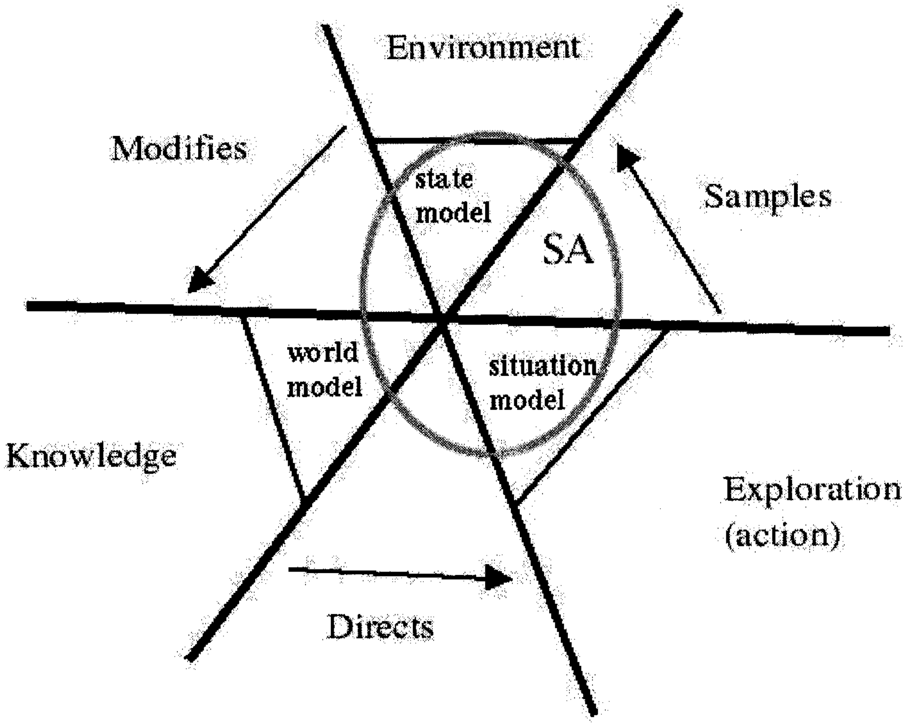


Illustration 8. Revised Model of SA

The overall picture above is one where mental models have the potential to *modulate* the perceptual component of the perception-decision-action loop, and they are by consequence, a potential causal mechanism in decision-making as well. Notice how the *situation model* drives the *sampling* process of the environment. This representation does not set a strict division between knowledge and behavior, therefore standing as a more ecological description of SA and its relation to performance as seen in the previous literature review. There are three types of mental models or schematic information present in the above diagram. The *world knowledge* would be equivalent to the static idea of mental models; the *situation model* is a homomorphic instantiation of world knowledge schemas, created real-time in LTWM, highly context-dependent and subject to local constraints (environmental and performance-related). The state model is similar to the situation model, but the emphasis there is on state information and not functional characteristics of the environment. In other words, moving from world to situation to state models, we lose representational emphasis on function and move towards information about environmental variables (states). This shares similarities conceptually to moving through a means-end hierarchy (Rasmussen, 1982) or homomorphic lattice (Moray, 1987), or Q-morph hierarchy (Holland et al., 1986). On real-time performance, knowledge about the overall operation of a system is hardly used, unless needed explicitly for inferences and decision-making. Even in most cases, reduced versions of those models are used to make decisions, i.e., the situation models. In other words, the combination of situation and state models provide a better perspective to give an account for the type of SA-related errors that occur due to real-time constraints; these constraints are imposed on the *selection* of representational models in use and therefore have influence on the complete perception-action cycle in human performance. From our diagram above we see that the situation and state models combined would be the equivalent to SA in the previously discussed models.

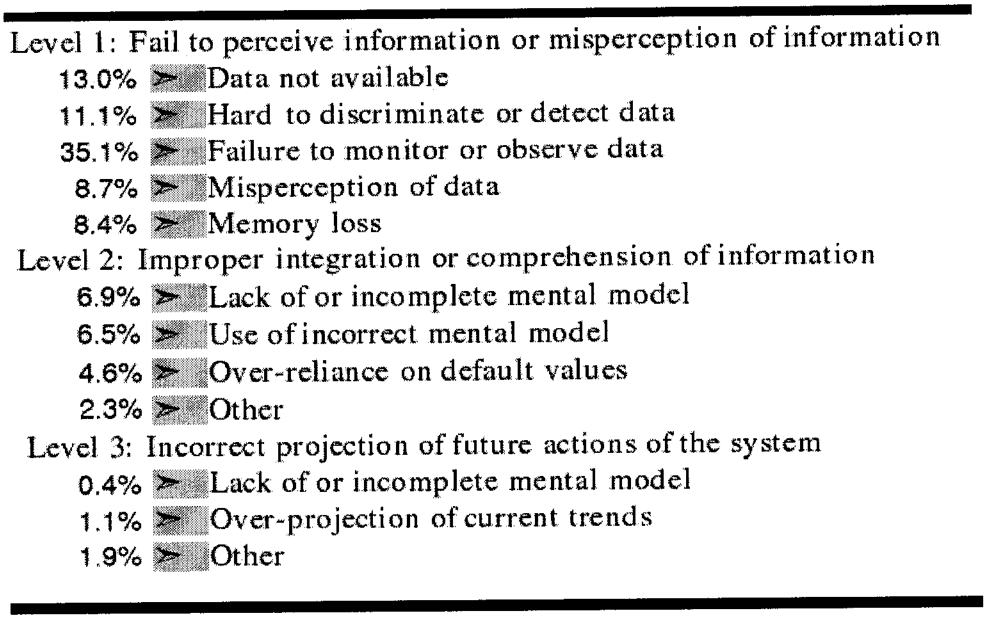
Error Taxonomies

The major problem with the current error taxonomies for SA-loss is that they contain large *pockets* in their distribution. A pocket is a category in the taxonomy where a disproportionate portion of the distribution falls. In other words the distribution of errors is uneven, which points to the inadequacy of the taxonomies in the first place. One key factor in this research is to provide a tool for better discrimination of SA-related error sources. For instance, to be able to detect in text differences between intrinsic human variability producing SA-related errors in performance, versus "general biasing" induced by the presence of *outdated* or *incorrect* mental and situation models present at the time of the incident, as described in the previous section. This distinction should then enable us to assist in the creation of improved error taxonomies by providing a diagnosis tool for discriminating these more subtle signals. The goal is to be able to better detect differences between misperception and misrepresentation, which form the pockets in the taxonomies. In this way our aim is to help enable the creation of SA taxonomies with better distributions. We will comment and expand on this idea in the final commentary chapter of this thesis.

Some well known surveys in the field of Human Factors regarding error taxonomies and their application to incident databases in aviation are: Shappell

& Wiegmann (1997) and Sarter & Alexander (2000). In the former, situation awareness and social variables were not included in the analysis, which was based on three different frameworks for errors. This indicates a clear factorization of SA and performance errors, something slightly different to what we are suggesting in this thesis. The latter work also does not include any mention of SA-related errors or misrepresentation. It is based on a performance-level breakdown in the lines of Rasmussen's ladder(Rasmussen et al., 1994) and Reason's (1990) general taxonomy. Again, the lack of SA mention or inclusion in the analysis and theoretical loop, indicates that a significant view of SA errors in the aviation field places it at the same level as envelope-shaping physiological factors, such as memory loss, and interference.

The most prevalent SA-based error taxonomy is that of Endsley's (1996, 1999). It presents the following distribution (Table 1 below) as obtained from a set of reports from incidents recorded by the National Transportation Safety Board (NTSB) over the course of four years. From this set, 88% were identified as containing a substantial human error component, and 71% as SA-related errors. The final set of 32 reports attributed to SA-loss was broken down according to Endsley's SA taxonomy. The results are shown inTable 1 below.



Level 1: Fail to perceive information or misperception of information	
13.0%	➢ Data not available
11.1%	➢ Hard to discriminate or detect data
35.1%	➢ Failure to monitor or observe data
8.7%	➢ Misperception of data
8.4%	➢ Memory loss
Level 2: Improper integration or comprehension of information	
6.9%	➢ Lack of or incomplete mental model
6.5%	➢ Use of incorrect mental model
4.6%	➢ Over-reliance on default values
2.3%	➢ Other
Level 3: Incorrect projection of future actions of the system	
0.4%	➢ Lack of or incomplete mental model
1.1%	➢ Over-projection of current trends
1.9%	➢ Other

Table 1. Error Taxonomy and Survey

Errors attributed to SA have been found to be a significant factor of performance errors in aviation (about 70%), as found through studies in the field such as this one. From Table 1 above, we note the appearance of the term mental model in levels 2 and 3 *exclusively*. We also notice from the percentages shown that about 76% of the errors fall under level 1. It has been recognized in the literature, that levels 2 and 3 are much more difficult to detect than level 1, which is related to perception almost exclusively and thus more evident. One argument in support of this fact is that level 1 errors are usually noticed by the protagonists themselves (i.e. the operators), and thus reported explicitly in the incident forms. This is then available in retrospective by the human factor analysts. This is one of the main argument for the approach we are proposing here, namely, that level

1 errors can be detected through keywords and phraseology, whereas levels 2 and 3 are much more subtle and need a slightly different approach (to be described in detail in the next chapter entitled "Method"). Endsley herself (Endsley, 1999) comments on the general distribution of SA-related errors in aviation studies. There is one very interesting fact about her commentary. In one instance, SA-related fields were added to the FAA report forms so that voluntary information could be collected. The distribution in this instance was higher at level 3 of the taxonomy (i.e when the operators themselves where evaluating the type of errors from the available categories). On the other hand, a similar study conducted by experts on the narratives using Endsley's SA model, found out that the distribution was much higher at the level 1 (i.e. when the experts where the ones making the classifcation of the errors from the narratives). Endsley fails to note this, but it is an interesting fact to mention. It is not clear whether the data came from the same sample pool or not. What is interesting nonetheless is that it points to the general difficulty about classifying these type of errors, but it also brings out a much more subtle point: the pervasive effect of mental models in SA-related errors. In other words, these errors are hard to detect by both the protagonists and the analyst experts, and they are detected *differently* (i.e. perceived differently). Endsley states "*there is evidence that people can fall into a trap of executing habitual schemas, doing tasks automatically, which renders them less receptive to important environmental cues*"

(Endsley, 1999). She is making reference to a very prevalent aspect about errors that has received considerable mentioning in the literature: "*cognitive lockups*" (Moray, 1987), "*diabolical errors*" (Carmino et al., 1990), "*habit capture*" (Reason, 1990), are among the names used to described this phenomenon. This begs the question then of *why* these errors occur and are so prevalent, and more importantly, why the current taxonomies don't seem to capture this. We will suggest that better taxonomies can be developed, or are possible, with the contribution of methodologies specifically designed to detect this general biasing present at the time of an incident. A taxonomy of SA-loss errors should reflect the fact that *expectation* could be a source of error at all levels of performance (i.e. perceptual and decision-making). We will argue so and discuss how this could yield a better distributed taxonomy in the concluding chapter.

Our argument in support of this research rests on two main assumptions. First, we are restricting the analysis to highly skilled operators (experts), who have many hours of working experience on top of their formal training. Therefore we assume that their mental models are well developed and have the potential to be complete in terms of their functionality. Secondly, and partly because of the first assumption, these reports should have enough information to distinguish between the types of error we are seeking to analyze (i.e. environmental, performance envelope and representational causal factors). Therefore our assumption is that causes of biasing or blocking should be detectable in light of the expertise level considered.

For instance, given a preliminary review of incident reports from the ASRS database, pilots and air traffic controllers are usually willing to admit of an error due to work conditions, stress, memory mishaps and communication between operators. Of course this is no guarantee of that being accurate to the real causal mechanism of the error in any given situation (due to the psychological limitations people have with introspect), but it could prove effective in discriminating among the two. In other words we can assume that for highly skilled operators, performance errors will be *highly* visible, whereas biasing ones will be much less so. So given a seemingly performance-related error, if the narrator expresses surprise or puzzlement about the causal nature of events, without mention to extrinsic factors, we have an indication that the real cause of the incident could lie somewhere back in the development of events. Namely, at some point in time an unrelated event prevented the correct updating of the situation models. This idea will surface again when we present our method for the analysis of text and the discussion of the results.

Cognitive Linguistics

Cognitive Linguistics is a fairly recent branch of linguistics that originated in the early eighties, and that has rapidly gained grounds in Europe and the west coast of the United States. It differs radically from established formal views which model language with either a mathematical or statistical approach, as a set of rule-based constructions. In the words of one of its scholars, "t*he original impetus for Cognitive Linguistics came from the pioneering research of psychologist Eleanor Rosch on the nature of human categorization. Throughout its brief history, Cognitive Linguistics has maintained a lively dialog with allied disciplines such as psychology, anthropology, neurobiology, motor control, artificial intelligence, philosophy, and literary criticism*" (Janda, 2000). Although it has no central tenet or single authority, it does have some consistent theoretical grounds, and an increasingly large number of empirical and theoretical work to accompany it. Among its principles, it considers the processes and structures of language to be no different than those of cognition, but only a subset of them. It acknowledges the existence of basic-level structures and meaning devices, which are present also in language through closed functional categories (Lakoff, 1987, Talmy, 2000). These are Cognitive Linguistics' answer to the classical problem formal logic of 'meaning primitives'. These basic elements are schematic in nature (i.e. highly abstracted) and enable the construction of concepts in cognition from basic perceptual elements, such as space and time. Furthermore, this theoretical field posits that meaning is *embodied*, and therefore conceptualized only through the interaction between a cognitive being and its environment, much in line with the ecological perspective in psychology of Gibson (Gibson, 1977), and the current trend in the

Human Factors field. Within the body of work in Cognitive Linguistics, this research project will draw its inspiration from the idea of *schematic systems* by Talmy (2000), more substantially on the formalization of *discourse modes* by Smith (2003) and more loosely on various scholars of the field from Europe (Couper-Kuhlen & Kortmann, 2000).

Talmy's view of language is that of asystem (a cognitive system), with a set of overlapping subsystems responsible for structuring meaning according to common and basic underlying principles. The overlapping of components gives rise to shared features of language, which then can be considered as common elements of language and cognition. These common features can be thought of as the foundational *strata* in language*:* configurational (conceptual), movement (space and time), attentional, perspectival and force-dynamic (causal). Among this set of components, there is pattern-forming subsystem, which is the *narrative* system. Its main characteristic is the temporal integration of events via the production of coherence. It is from this perspective that we will address the challenge presented by the AvSP project. How can this fundamental characteristic of the narrative system be exploited to understand meaningful variations? In engineering terms, we can conceive of this cognitive subsystem to posses a transfer function, and our study will exploit this idea. The following illustration exemplifies this idea.

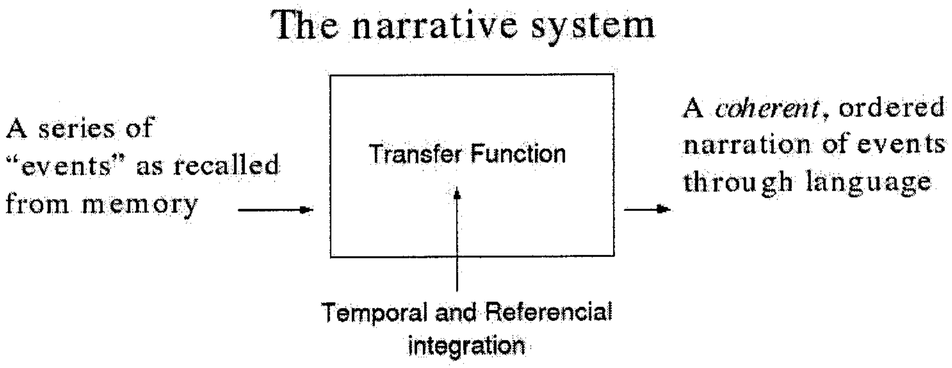

The narrative system
A series of "events" as recalled from memory
Transfer Function
A *coherent*, ordered narration of events through language
Temporal and Referencial integration

Illustration 9. The Narrative System

Smith's work on temporal segmentation of discourse(Smith, 2003) will be key to the implementation portion of this work. Our focus will be on the particular discourse mode which corresponds to the narrative style. This mode is characterized as a temporal pattern forming system, in agreement with Talmy's view. The different styles of discourse in general have been studied extensively in the literature of rhetoric and discourse analysis. The structure of narrative discourse, in particular, has a very definite set of characteristics which constitute a given style of writing and of discourse structure in itself. The main characterization of narrative coincides with the cognitive perspective, which is to define it by a cluster of cognitive-related functional features. Smith describes what she calls *discourse modes* as the compositional set of features that identify a given discourse structure. She states the following features as defining the narrative mode:

- Situations described. Primarily specific Events and States
- Temporality. Dynamic, located in time
- Progression. Advancement in narrative time

Furthermore, there has been empirical data supporting the distinction of the proposed modes at the psychological level. In a study conducted by Faigley & Meyer (1983) as cited in (Smith, 2003), subjects were able to discriminate among different styles of narrative based only on temporal and situational characteristics of the sources. This suggests some psychological reality to the idea of narrative modes. Also, the fact that most of us are able to recognize genre in text, indicates some validity to this idea.

Added to the previous description of the narrative style is our own domain-specific style arising from the constraints imposed by the incident reporting system (the ASRS) and the current analysis model of behavior. The conceptual model employed for this portion of the AvSP effort is labeled the "Scenario" model (Maille et al., 2004). The Scenario model treats the ASRS narratives as consisting of a series of *events*, which are transitions among a series of *states* about the system (the NAS). Each narrative story consists of an initial state, at least one compromised state, and a series of normal states. This sequence is meant to represent the development of an incident, from a normal to a risk loaded state, and back to a normal state after a resolution is effected (none of these reports are about actual accidents). This model will be further described in

Chapter 3. We will characterize the temporal structure of our narratives using a mix of this notion and the canonical style of the narrative discourse mode itself. This will become clear in the next chapter when we describe in detail the formalization of our style measures.

T*ense* and *aspectual* information will lay the formal foundation for the analysis of the text. We will establish a grammatical and computable distinction between *Events* and *States* using syntactic templates, and based on the Situation Type Theory work done by Smith based on Vendler's matrix(Vendler, 1957). Situation Type theory was introduced in nineteen fifties by Vendler as a way to characterize a space of event types that appear cross-linguistically. These linguistic events portray how language conveys the temporal contour of a predication in language, and therefore, of their mental representation counterpart (from the Cognitive Linguistic perspective). For instance, events are characterized by their inherent dynamism, expressed through a predication's main verb dynamic qualities (such as movement, or change). They are also characterized by the internal shape of the event, such as its instantaneousness, or its repetitiveness. Events are described in terms of their boundedness. These semantic characteristics will form the basis of our syntactic segmentation. We will formalize it in detail in the next chapter.

Discourse Theory

Whether discourse has an inherent structure that can be characterized formally is not a question that we will address in this work. Even though there has been a considerable body of work and effort poured into this enterprise (Smith, 2003; Mann & Thompson, 1988; Reichman, 1985; Grosch, 1986), it remains one of the most challenging areas of linguistics and artificial intelligence.

From the psychological standpoint, there are fundamental processes that seem to glue sentences together to form discourse. These are *coherence, plausibility* and *causality* (Johnson-Laird, 1983). Johnson-Laird has proposed that *temporal* and referential coherence enable the production of a consistent representation of discourse. This representation has two levels: the surface and the mental. The surface is the propositional representation close to the lexical form of text, and the mental refers to the internal structure and relational characteristics of the events and states that are described in discourse. Johnson-Laird shows how this idea implies the construction of mental models within a framework of plausible causality reflecting the structure and relations of the real world.

In linguistics, scholars have proposed a similar correspondence to this invisible glue, and loosely refer to it as coherence or *propositional relations.* There is no clear resolution as to whether these functional relations can be discretely enumerated and characterized for all discourse, or whether there are only some fundamental ones, such as temporal or causal, that then give rise to a large number of domain and style-specific ones. On the side of formal linguistics, Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) proposes a relational view of the structure of text, that provides a hierarchical representation of discourse and a set of finite structurally and lexically-triggered rhetorical relations. In the field of pragmatics, it has been proposed (Couper-Kuhlen & Kortmann, 2000) that discourse is best characterized by only assuming a small set of fundamental relations, which are State-of-the-art by the principle of *Relevance* (Sperber & Wilson, 1986) as the basic communication goal. These are: topic, time, clarification, causality, and persuasion(Couper-Kuhlen & Kortmann, 2000). All other relations are derived from these, and more importantly, these are the *generators* of relevance in discourse.

For the sake of this work, we will not concern ourselves with such debate, but rather focus on one fundamental relation: temporal coherence. We will argue that temporal coherence manifests itself in the overall grammatical composition (style) of our incident narratives, and we will define an approach to measure this formally in the following chapter. We will not adopt any specific discourse representation theory.

Information Extraction

This section discusses two areas of technology and conceptual approaches for extracting information from text. On the one hand, Information Retrieval (IR), and more specifically Information Extraction (IE). On the other hand, a more specialized cross-disciplinary set of methodologies known as Content Analysis, which originate from the fields Sociology and Clinical Psychology. The main differences between the two is in the techniques they employ for data mining, which stem from the different perspective they have on the nature of the informational content in language.

Information Retrieval

> *Information retrieval is a growing field that encompasses a wide range of topics related to the storage and retrieval of all manner of media... Most current information retrieval systems are based on an extreme interpretation of the principle of compositional semantics. In these systems, the meaning of documents resides solely in the words that are contained within them.*(Jurafsky & Martin, 2000, pg. 646-647).

The previous excerpt aptly describes the predominant paradigm in this field, which is known as the *bag of words* approach. This paradigm uses statistical methods to categorize documents in terms of words, sequences and combination of them. It takes no account of the intrinsic mechanisms of language, such as syntax and semantics, which have not yet been described in terms of any statistical model. The end result has two severe limitations: one is that the over-reliance on keywords and strict lexical form produces domain dependence and requires a significant amount of *a priori* knowledge for its setup; the other limitation is that it provides no explanatory power for studying and exploring behavior, intentions and psychological states. Conceptually, the difference between IE and IR is that in the former it is necessary to know beforehand the type of information that is being sought, and more importantly, having a working model of the phenomenon under study. The latter (IR) focuses in the categorization and extraction of documents (where a document is defined as any bounded set of information), given an *ad hoc* query for any given domain. Technically, the difference lies in that IE uses templates that are generated beforehand using tools for pattern matching, such as the Java Annotation Patterns Engine (JAPE), used by the Battelle team; in the case of IR, information is represented in terms of a *Vector Space Model,* which are vectors and matrices of weighted, context-sensitive lexical features, representing terms that occur in a document. By context-sensitive here it is meant linear context of word sequences. To quote from a source: " *The main problem with IE is the degree to which knowledge is template like, in the way history was once taught (but is no longer) as factual templates of kings, presidents, battles and dates. A major research issue is seeing how far the boundaries of templatability can be pushed out*" (Wilks & Gaizauskas, 2000, pg. 212).

One particularly interesting case study, is the application of IR techniques to a problem in our domain. This is in the work done by McGreevy (McGreevy,

2001) from NASA. QUORUM is an IR program specifically targeted to extracting information from aviation incident report narratives. Developed by NASA, QUORUM has capabilities to perform keyword and phrase searches, and phrase generation and discovery. The main advantage of this tool is that it has the capability of assisting the user in determining the most appropriated queries for any given topic. In other words, given a set of keywords and phrases, the discovery methods of QUORUM can provide associated keywords, phrases and topics to be consequently used as input for queries. Another feature of QUORUM is its use of the *Keyword In Context* method, which is an extension of the vector space – traditional of IR – used to compute scores for terms in contexts across documents. Simply put, the vector model is extended by adding features that measure the distance between the target word and its (linear) context, bidirectionally. These improved feature vectors are used to calculate a measure of *relevance*, which roughly corresponds to improving *precision,* in IR terms. Precision stands for

$$\frac{\text{\# of correct answers given by the system}}{\text{\# of answers given by the system}}$$

An example of a feature vector in QUORUM is the following:

word1	word2	A	B	C
CREW	REST	9241	264	50.9163

What the above table represents is the contextual association for the given word pair. Column A is the measure of the strength of the contextual association of the word pair across the database, B is within the current document (narrative in this case), and C is a combination of the two. The values in column A and B are obtained through a modified directional measure of the Relation Metric Value (RMV), and calculated as follows

```
for: term1 term2     term3      ...        termX
left RMV for term1 and termX = C - 1- N
```

where C stands for the length of the sentence, and N represents the number of words between the two target words.

As noted previously, this type of technique proves very useful in extracting information with a given model in hand for the phenomenon under study. It is also very useful when there is a substantial amount of information available about a specific type of incident in a database. Nevertheless, it falls short of allowing the detection of latent patterns that could potentially emerge that better discriminate causal mechanisms in human performance errors. What we are seeking in this thesis work is a methodology that is not tied to any specific *context* or domain-specific knowledge, such as what can be obtained through exclusive reliance on words.

The next interesting and relevant case study for our literature review is one involving the analysis of style in text. Karlgren (Karlgren, 1999) explored several experiments on stylistic variation for information retrieval purposes. A style in this case is defined as a " *consistent and distinguishable tendency to make...linguistic choices*" (Karlgren , 1999, pg. 147). Some of these linguistic choices that Karlgren refers to are choices regarding the *organization* of the narrative, syntactic constructions and choices of synonyms. He declares that style is not necessarily perceivable always by the reader of a text, but that it establishes a predisposition toward understanding the content in one way versus another. We will refer to this here as *phenomenological or latent* informational content. One of the main purposes of these experiments was to test new and complementary methodologies to the current existing techniques of IR. This is akin to the driving force behind our work, and it will be noted throughout this document. More specifically, as Karlgren states, this type of complementation was not be aimed at improving precision or recall measures (standards in IR) directly, but at improving the "quality" of the selected documents (in terms relevance, as judged subjectively by a panel of experts). In our case, this translates to assessing a measure of the overall subjective level (from the point of view of the narrators) of a given record and to the distributional characteristics of the subjective elements themselves (i.e. the segments).

The selection of dependent variables for these experiments was based on previous work done on authorship and readability. There were two axes for classifying these variables: lexical versus text-level statistics. Lexical variables were average word length, type-token ratio, use of digits , capitalized words and personal pronouns. The text-level variables were mainly text-tiling, sentence length and parser depth. Text-tiling refers to an available algorithm for identifying the number of topical units within a sentence. Parser depth refers to the depth of the syntactic trees obtained from using a syntactic parser to process the data. Here also, our approach will share similarities with these experiments in that we will also employ a syntactic parser to process our data, and that we will also consider clause complexity as an indicator of local relevance. One of the main results from this article is that syntactic complexity *is* indeed directly correlated to relevance for any given document. That is, the more complex sentences a document contains, the more likely it is to yield a higher degree of relevance within a given topic. Recall that relevance in this case was assessed subjectively by judges.

The statistical techniques employed to analyze the data in these experiments were nonparametric, given that no assumption can be made about the underlying distribution of the data. Again, the same applies in our case. The author was interested in measuring degrees of relevance within a given topic pool. He employed ranking algorithms such as the Spearman rank order correlation (rho) and the Mann Whitney U rank sum test. In our case, we are interested in identifying groups within our data, or in other words, of discovering the underlying distributions per se, i.e., the baselines. Once these are available, they can be used to detect meaningful deviations. Therefore, we will also employ a nonparametric tool: clustering.

An interesting review of all the computational methods available for stylistic measurement of text is Biber (Biber, 1993). Biber produced a compilation of all the different variables considered when studying style in free text. Some of these are: tense and aspect , place and time adverbs, pronouns, questions, nominal forms, passives, stative forms, subordination features, prepositional phrases, modals and negation. Our approach has much in common with this list; it also takes into account tense and aspect, adverbs and prepositions of time and place, subordination, modals, negation and stative forms. The particular *method* employed will be very different however, motivated by different analysis goals. In an experiment Biber produced categories of features (dimensions) by means of aggregations of variables, and assigned to each one of these dimensions a given significance for the purposes of his study. For instance, dimensions would be indicative of patterns of persuasion, information, involvement or abstract description styles. He employed factor analysis to identify significant co-occurrences leading to these feature clusters. The co-occurrence of lexical signals was done using conditional probabilities on bigrams, a standard approach. He could have just as well employed clustering techniques like we will do, such as agglomerative hierarchical clustering, to discover this pattern visually.

Finally, another interesting sub-area in data mining of text is that of word dis-abbreviation and typo correction. This was of particular interest for this review since it will be the basis for the preprocessing stage of the implementation program for this work (see Chapter 3). The two main works examined were that of Godden et al. (Godden et al., 1993) and that of Akira & Tokunaga (Akira & Tokunaga, 2001). Godden et al.'s work dealt with a similar problem in the context of the General Motors Research Laboratory. Faced with the need to extract knowledge from information repositories, part of the challenge was the intrinsic variability that comes with spontaneous and unrestricted natural language. One major task for this team was to come up with heuristics and robust tools for dis-abbreviating and standardizing the wide range of variation that appears in the form of acronyms and typos, which was widely present in their data. For this purpose the team developed a separate component of their program, named Lexfix. This component is "a *vocabulary correction and standardization system ... designed to improve keyword-based retrieval on ... free form fields*" (Godden et al., 1993, pg. 27). Each word is assumed to be an inflected form of a known word, a spelling error or an abbreviation. The plausibility of each choice is ranked and the winner selected as a replacement. Standardization occurs as the last step and from a dictionary. The work of Akira et al. (2001) used a statistical technique based on the context of the usage of a word (similar to the feature vectors in IR), but also implemented a preprocessing stage, where heuristics about common typos and abbreviations were employed. The set of heuristics made available from both of these works forms the basis of the heuristics employed in our current work. They are described in detail in Chapter 3 under the heading "Preprocessing" in the "Implementation" section. Together with this, we will employ a software tool named Aspell that implements a widely used typo correction algorithm (see Appendix D).

Content Analysis

More than a specific discipline, Content Analysis is a set of techniques and models employed for the analysis of communicative text corpora. At the beginning of the twentieth century Content Analysis was predominantly the domain of mass communication researchers. Political Science adopted it in the nineteen thirties and by the forties it was part of the techniques used by clinical psychologists, anthropologists, literature scholars, historians and linguists. It became known as an interdisciplinary effort after the 1955 Conference at the Allerton House at the University of Illinois, which was hosted by the Committee on Linguistics and Psychology of the Social Science Research Council. The main outcomes of this meeting were published in a book by Pool in 1959 (Pool, 1959). The main reason to include Content Analysis in the literature review for this work, is the similarities in the problems-at-hand. More specifically, their conceptual models are relevant to our task as well.

The original force behind this movement came out of the social sciences during the study of Nazi propaganda from World War II. In the efforts to do so, new models for analyzing text came about. One of those main concepts is that of *manifest* versus *latent* meaning in communication. By manifest it is meant the type of content that is readily accessible to human readers. Latent on the other hand refers to information (largely phenomenological) that resides in the structure and form distribution of language. These concepts give rise to two opposing conceptual models for encoding textual data: *instrumental* and *representational.* Instrumental refers to the interpretation of textual communication using a predefined research theory. That is, language is just the instrument for communication, and there is some other theoretical model that is applied to extract meaning from it. It ignores the original intended meaning by the author of the text. It usually places emphasis on syntax, lexicon and structure. The representational model refers to the use of language to extract the original intended meaning from text, therefore focusing on the context and background of the communication setting. Psychologists have argued about the need to focus on the instrumental aspect of language to infer cognitive structures and states, claiming that it is closer to the behavioral side of language (in the psychological sense), as opposed to the anthropological and sociological side of it (related to ethnography or cultural anthropology for instance); see Mahl's article in (Pool, 1959) for more on this. Psychologist Charles Osgood was one of the first scholars to propose a quantitative methodology based on the instrumental model. *" As a matter of fact, we may define a method of text analysis as allowing an 'instrumental' analysis as if it taps message evidence that it is beyond voluntary control* *of the source and hence yields valid inferences despite the strategies of the source"* (Pool, 1959). He was also the first to present a view of language communication that breaks down the perspective of these two models into a larger picture. It is shown in Illustration 10 below.

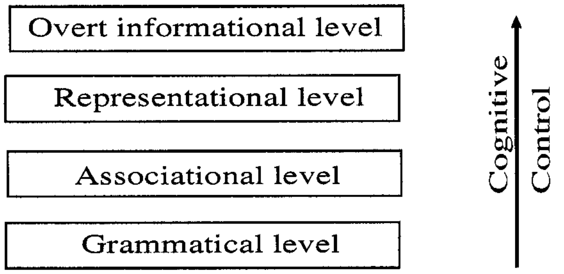


Illustration 10. Levels of Communication

The most interesting thing to notice here is the upward pointing arrow to the right, which suggests levels of cognitive control and therefore of awareness by the text producer and reader. This general idea is part of the basis of our argument for this thesis. The *overt* information level refers to the actual content of text as conveyed and intended by the author. Natural Language Understanding (NLU) technologies are still in their infancy, so we can assert that *computers don't understand meaning*even if we could define meaning operationally. Therefore, all that is available nowadays are techniques for deep parsing of text, based on hybrid architectures of machine-learned rules and statistical models. The *representational* level, which can also be called affective level, is the most common level approached in the social sciences; " *Here, word counts of varying levels of complexity and sophistication are employed, with researchers assuming, either explicitly or implicitly, that the occurrence of words, phrases or categories of words is an indicator of meaning, at the least, and of effect upon the reader at most*" (West, 2001). IR technologies would fall under this category. The *associational* level is where most of the current computer-assisted text analysis work is being done. Concepts are extracted from text using clustering models and techniques as well as neural networks. The categories involved in coding the data in this case are based on word co-occurrence and contingency of phrases and tokens. Finally, the *grammatical* level deals with type/token ratios of words and morphological information. They use statistical techniques to find deviations from expected values in a given corpus. One way to conceptualize this hierarchy, is to think of the two top levels as representational, meaning that their purpose is to convey a message "as is" (as it was intended by the author). The lower two levels could be considered instrumental, in that they represent language as a *vehicle* for conveying information, regardless of the original message. They provide *mechanisms* for conveying a message in a given style, to achieve different levels of effectiveness in terms of suggestion, expression of belief, outrage, etc. An example of this is reading a novel. If ask, one can clearly convey the storyline of the novel, its characters and name its author. But one cannot exactly describe what makes that particular novel different from another one (besides the story line), in terms of the grammatical features that identify its author, or the genre of the novel.

Numerous techniques have been devised to detect and measure affective and mental content from text using these models. Among the most interesting and relevant to us are: the Gottschalk-Gleser approach and the Perspective Analysis (Popping, 2000). Both methodologies are distinguished from the rest in that they employ parsing techniques for encoding their data, which is similar to the approach proposed here. The first one, the Gottschalk-Gleser approach, is a method for measuring psychological states based on an affective scale and on categories. These categories are encoded *a priori* into schemas of clausal constructions. Verbal interviews are transcribed and text units of Subject-Verb-Object are assigned into the categories. Summation of weighted scores on categories is then aggregated into a scale score. These scales are used in clinical psychology, psychiatry, medicine and psychosomatics. The Perspective Approach, developed by Bierschenk from Lund University in Sweden, also employs Actor-Action-Object relations as basic coding units. The difference is in the underlying model. This method is based on Kant's philosophical schema axiom, which distinguishes between analytic and synthetic propositions. Bierschenk argues that " *the analytical proposition gives clauses a formal structural definition*" (Popping, 2000). Based on this he proposes that intentionality can be detected structurally and therefore his method can characterize the mental models that govern language production. He employs clustering techniques to identify his categories and he proposes the use of four *viewpoint* modifiers to indicate distance in intentionality.

What these two methods have in common with our proposed methodology is the use of NLP parsing tools and techniques, and the underlying model that mental states and processes can be detected structurally in language. They were part of the inspiration for our approach.

CHAPTER 3

METHOD

This chapter will describe the theoretical basis and the computational formalization of our approach. The main purpose of this thesis is to prove that this novel approach is feasible and potentially useful for the specific problem of analyzing SA-loss errors within the AvSP project. Therefore the scope of this work is limited to the description of the approach in terms of its theoretical motivation and its formalization as a computer program used to validate its robustness.

The chapter will be broken down as follows. First, we will describe the psychology-based motivation for devising our novel approach. Then, we will proceed to explain how the theory served as the basis for the design and formalization of the conceptual structures that we will employ. Finally, we will describe the computational implementation of such structures and the overall architecture of the program. Our program will produce *segmented* narratives as its output, based on grammatical features to be defined. Our main qualitative and creative hypothesis is:

> *To demonstrate the feasibility and robustness of an new automated approach to analyzing textual narratives, using linguistic semantic theory, without reliance on keywords but on syntax instead; to show that it has potential applicability to studies of behavioral models of human performance.*

The rest of this chapter will describe our approach and prove its computational feasibility. In Chapter 4 we will describe the clustering method employed in order to uncover the underlying distribution of our data, as coded with our new tool. Chapter 5 will comment on the applicability of our approach and on improvements for the future.

Motivation

Maille et al (2004, pg. 33-34 cited from Ericsson and Simon, 1993) note:

> *People have no awareness, that is, no ability to verbalize, their own perceptual and cognitive processes... People have little or no ability to provide any accurate information about their performance or cognition after a short time has passed...*

The two main sources of noise in our data come from *inference* processes and memory *time* decay. As noted also by Ericsson and Simon in their work in *Protocol Analysis* (Ericsson & Simon, 1993), inferential processes rather than memory retrieval can play a significant role in the creation of noise in the data. The time decay factor is proportionally related to such inference noise. The aim here is to investigate cognitive processes which are presumably more stable across time and that contain less inferential noise. It is important to investigate methods for analysis and data mining which take into account this type of limitations. Therefore our main goal is to develop a methodology that will try to access a basic process: the temporal integration of mental events during narrative production. Theoretically this process remains consistent with the time of a given incident. In other words, the fundamental process of integrating memories, remains to a certain degree intact and available to the narrator at the moment of the narrative production. This phenomenon could potentially contain explanatory information about the source of the error. Thus our methodology will enable a first attempt at the detection of a pattern that relates to behavior: we will refer to this generally as biasing. By biasing we mean a measure of the degree of deviation from the canonical temporal distribution of the course of events as portrayed in the narratives. In other words, are there different ways (with respect to the order of events) in which these stories are told, and does that yield useful information? Of course given that each incident is different and also due to intrinsic human variability, it is not possible to obtain this canonical form directly, particularly since we do not have access to the true course of events for any given incident. What we will attempt to do here is categorize *styles* of these narratives in terms of their temporal characteristics and in relation to each other within a give pool of sample data.

We define style as:

> *a given arrangement of events in terms of their temporal distribution as they are introduced to the reader during an incident report, and the quality with which each of these individual events are portrayed to the reader*

The reader should note here the conceptual difference between "presenting" and "portraying." Presenting is meant as order of presentation, whereas portraying is meant as the quality of each event description. Quality will be expanded later on in this chapter under the rubric of "situation types" which correspond to "mental event" qualities.

For instance,

> "WE RECEIVED CLRNC BY GND TO TAXI TO RWY 26R AND HOLD SHORT OF RWY 7 ON DELTA. **DELTA TXWY IS ALSO RWY 7/25 AND VISE VERSA.** WHEN WE WERE TURNED OVER TO TWR WE WERE HOLDING SHORT OF RWY 7. WE THOUGHT TWR SAID TAXI TO RWY 26R VIA DELTA ... IT WAS VERY BUSY ON THE GND. MAYBE WITH ALL THE RADIO CONGESTION WE MISUNDERSTOOD. MAYBE THE CTLR MADE A MISTAKE?"

versus

> "WE RECEIVED CLRNC BY GND TO TAXI TO RWY 26R AND HOLD SHORT OF RWY 7 ON DELTA. WHEN WE WERE TURNED OVER TO TWR WE WERE HOLDING SHORT OF RWY 7. WE THOUGHT TWR SAID TAXI TO RWY 26R VIA DELTA ... IT WAS VERY BUSY ON THE GND. MAYBE WITH ALL THE RADIO CONGESTION WE MISUNDERSTOOD. **DELTA TXWY IS ALSO RWY 7/25 AND VISE VERSA**. MAYBE THE CTLR MADE A MISTAKE?"

In the two examples above we see the same story – an actual narrative taken from the ASRS database – slightly rearranged. Recall our example from Chapter 1. Here again we see a piece of extraneous knowledge (in bold) introduced at different stages of the narrative. The argument is that this knowledge reflects a potential error mechanism, namely biasing on the part of the pilot. In other words the pilot is familiar with the runway configuration and therefore is operating under a potentially biased situation assessment. The place in the narrative where this information is introduced can tell us something about the *expectations* present at the time of the incident on the part of the narrator. The blame seems to be directed differently depending on where this extraneous information is introduced. It also signals a language device used by the author in order to justify or explain through inference how the decision was triggered at the time: "**DELTA TXWY IS ALSO RWY 7/25 AND VISE VERSA.** WHEN WE WERE TURNED OVER TO TWR WE WERE HOLDING SHORT OF RWY 7. **WE THOUGHT TWR SAID TAXI TO RWY 26R VIA DELTA** ...". The confusion state here could have been possibly triggered by the familiarity of the pilot with the runway configuration. These are the sort of signals that we are looking for here.

In order to measure these temporal deviations in our data, we will find *clusters* of styles that will become our *baselines* for comparison purposes. From these cluster we will identify *outliers* and tag them according to our method as either potential cases of biasing errors or general high subjectivity content. The *physical* product from our work will be uncovering the underlying distribution of our data, based on our coding parameters. The *qualitative* output of our work will be the hypothesis and commentary about the usefulness of the results for studying a human performance modeling problem.

One important aspect that we are assuming here is that this type of coding remains largely unavailable (in real-time and to consciousness) to the reader and writer of a narrative, making our effort effectively one of measuring a phenomenological feature of language. This is akin to measuring authorship or genre in text, which has been shown to work effectively. The information we seek remains largely unavailable to the reader and writer of the narratives in real-time; it can be potentially detected by them, but not precisely described while reading or producing text. Statistical methods in the literature (Karlgren, 1999; Biber, 1993) have shown that there are distributed features in language which make this detection possible. What we are proposing here is to look at *how* the information is portrayed (over time and in its use of grammar) to the reader during the development of a narrative story. In other words, even though readers and writers are not aware of these features in real-time, these feature are nonetheless processed by their cognitive system (to produce coherence, as we introduced in the previous chapter). The empirical and theoretical work on which we base and inspire our thesis stems from Cognitive Science, Linguistics and Psychology; in particular, from work on *cohesion* (Givon, 1995), on *iconicity* in narrative text production (Couper-Kuhlen & Kortmann, 2000), and the Long Term Working Memory model (LTWM, Kintch & Ericsson, 1995). The pivot is on the process of temporal integration of mental events, which occurs during narrative production, and which is argued to be the same in principle as the process that creates cohesion in the memory source. This relates to the *Interpretation* level in the DRICP model, and to the creation of *retrieval cues* in the LTWM model. It is argued that grammar in language is the fundamental manifestation of this process (Givon 1995; Kintsch, 1992).

Finally, and as a consequence of the problem and the solution-at-hand, our approach is not dependent on domain-specific knowledge, but rather on a general model of cognition and language and on linguistic knowledge about syntax. An approach based on keywords and phrases will have a model-related limitation, as it has been noted in the literature review of Chapter 2. Therefore we hope that our new approach will be welcome and used as a complement to the existing methodologies, given its advantage from the implementation point of view.

Conceptual Structures

Theory

The main concept behind the segmentation of narratives is the situation types proposed by Vendler (Vendler, 1957) and further developed and formalized by Smith (Smith, 2003). Situation type refers to the temporal contour and characteristics of how mental events are portrayed through language. These types are determined by the two fundamental building blocks for expressing time in language, *Tense* and *Aspect*. These two subsystems of language are at the very fundamental cognitive level that characterizes the narrative mode or style of discourse; that is, the coherent temporal integration of events. Tense refers to the anchoring of an event in time (past, present, future) with respect to some other temporal center, which in most cases is the time of the speech act, but can also be a time internal to the story of the narrative. This is part of what is known in linguistics as the *deictic* center. Aspect conveys a psychological perspective or *viewpoint* to an event. Aspect and tense, together with the intrinsic characteristic of verbs such as dynamism and volitionality, allow the definition of a taxonomy of situation types. We will use this as the basis for our temporal segmentation of the narratives. Before we can present this situation type matrix we need to further qualify aspect, tense and the properties of verbs we are going to employ for our method.

Aspect consists primarily of a viewpoint. There are two main viewpoints: the perfective and the imperfective. The perfective viewpoint is expressed grammatically through the simple form of verbs (i.e., no auxiliary verbs). The full theoretical definition of aspectual viewpoint is slightly richer including simple present, for instance, as expressing habituals which are also a type of imperfective aspect. For the purposes of this work we will adopt Smith's (2003) narrower definition. As an example of the perfective viewpoint consider the following:

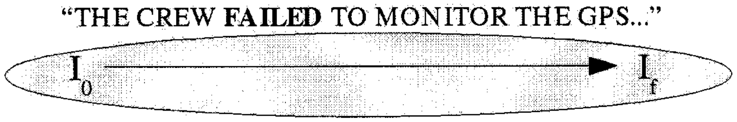
"THE CREW FAIL**ED** TO MONITOR THE GPS NAVIGATE DURING THE BUSY
DESCEND AT A CRUCIAL TIME A TURN"

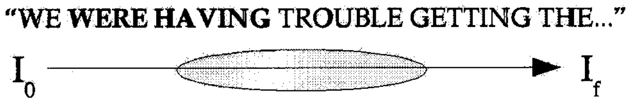
Imperfective viewpoints are others such as the progressive:

"WE **WERE** HAV**ING** TROUBLE GETTING THE FLIGHTMANAGEMENTSYSTEM TO
FLY THE HOLDING PATTERN SO WE **WERE** DO**ING** IT MANUALLY"

"Aspectual viewpoint is like the lens of a camera, it focuses on all or part of the situation

expressed, making the focused information visible. Only this visible information is available for semantic interpretation" (Smith, 2003). Conceptually, the difference between the two can be seen in the following diagrams (where I_0 = Initial state and I_f = Final state):





The shaded oval represents the temporal contour of the event, and it helps illustrate the portion of the event that is visible *semantically* in the speakers' and the readers' mind. More specifically, the last case shows a situation where it is not exactly known when the event started nor when it ended, but only that it was still occurring at the moment of description. The first case shows an event that includes both endpoints within its contour, making it a *bounded* event. In other words the start and the end of the event are visible semantically.

There are two semantic features that characterize the internal temporal structure of events: *dynamism* and *telicity.* Event semantic features are primarily determined by the main predication's verb. The dynamic quality of events is determined by the "agency" semantic feature of the verb in the main predication, together with the main subject of the clause. The other features that distinguish a dynamic event are the progressive viewpoint (i.e., imperfective "fly **ing**"), and locative and temporal adverbial constructions, such as "while", "when", etc. Telicity refers to verbs that have natural endpoints (bounded events, refer to as "Accomplishments"), such as to "turn" or to "pass"; these also include instantaneous or point events (Saeed, 1997), such as to "find" or to "start" something (also known as "Achievements"). The opposite type, atelic verbs, are those that do not have natural endpoints, such as "run" and "drive". Telicity as well as dynamism are features that can be overridden through *coercion* by other entities such as adverbials or prepositional constructions; e.g., "I drove **for 3 minutes**" turns an atelic verb into a telic-like event. The formalization of all of these features will be presented in the next section through syntactic templates. The full original matrix of situation types, is shown in Table 2 below.

Situation type	Static	Durative	Telic
State	+	+	Na
Activity	-	+	-
Semelfactive	-	-	-
Achievement	-	-	+
Accomplishment	-	+	+

Table 2. Situation Type Matrix

For the sake our analysis, we will amalgamate the original matrix into two generalized situation types called State and Event. Table 3 below shows the result of this.

Situation type	Static	Durative	Telic
State	+	+	-
Event	-	+/-	+

Table 3. Amalgamated Situation Type Matrix

The main motivation for this matrix design is the need for adequacy to the error analysis model employed in the AvSP program. This model is called the "Scenario" model and is shown below in Illustration 11 below.

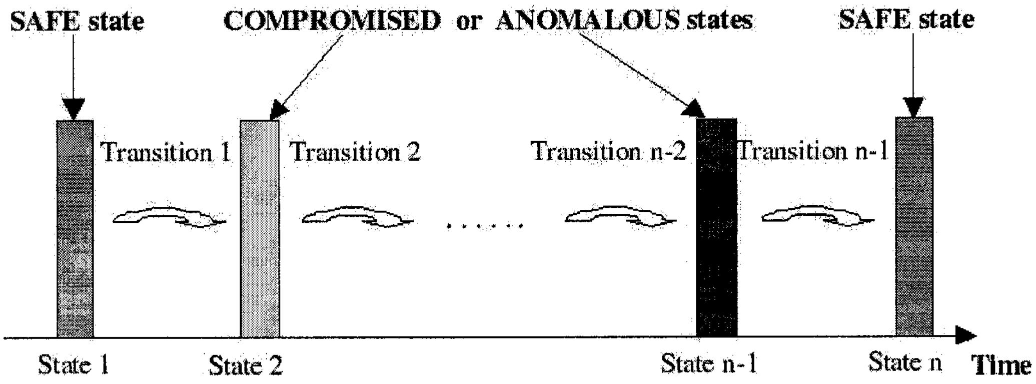


Illustration 11. The Scenario Model

The other motivation for the design stems directly from the work done by Smith (2003), where she defines the narrative mode as mainly containing these two types of events. From Table 3 above we see that a State is characterized mainly by the semantic features of stativeness, durativeness and atelicity. Events are dynamic and telic, and could be either durative or not. This is similar in part to Smith's description.The scenario model consists of representing each of the narrative stories in the ASRS database as a sequence of events and states about the system (the NAS). An event in this case is conceptualized as any action or occurrence in the operational environment that causes a transition from one set of values for the system's variables (a state) to another one. System variables are defined by the analysts, but they are always a close set, as available from the ASRS forms and the avionics instrumentation recordings. A state thus is a collection of variables representing the current state-of-affairs of the environment. Therefore a scenario is a sequence of interleaved states and events. Each one of these narrative reports consists of an introductory or initial state, where operation is normal, and a series of events and states leading to an *anomalous* state. Anomalous states are then followed by *compromised* states and eventually lead to a normal state again. The transition from compromised to normal state is usually effected via resolution on the part of the operator. For a complete description of the model and its use, see Maille et al. (2004).

We will operationalize our definition of situation types according to the scenario model and the narrative discourse mode. Therefore a State roughly corresponds to a state in the scenario model, in that it portrays an ongoing state-of-affairs in the environment. The quality of this description is then stative, atelic and durative, since it describes a snapshot of the state of the system at a given time. Events inherently advance the narrative time, therefore they possess equivalence to events in the scenario model, which also advance the story line via transitions. They are characterized by dynamism, telicity and possibly durativeness.

With respect to Tense, we will use the following paradigm. We will treat the narrative style as consisting of three timelines:

1. Speech time. The time of production of the narrative
2. Situation time. The time of the event being described
3. Reference time. The current time in the progression of story in the narrative

Speech time (Sp) is self-explanatory; it refers to the moment of communication or speech act, in our case, being the time when the author wrote the incident report. Reference time (Rt) refers to the timeline which anchors the narrative progression. Every narrative has a *linear* (canonical) timeline internal to the story being told. This is a crucial assumption for the purpose of this analysis, but one that is based on considerable theoretical and empirical work in linguistics. In other words, the narrative mode (Smith, 2003) has by default a linear incremental progression of events, where deviations from it can be detected through temporal shifts via grammar. So a narrative *advances* the reference time by default, unless there is an explicit signal otherwise. Situation time (St) refers to the internal time of the situation being described. It can coincide with the internal time of the narrative (i.e., the reference time) or not. It usually refers to events at a time before the current deictic center. For instance,

"THE FIRSTOFFICER HAD INADVERTENTLY ALLOWED THE AIRPLANE TO CLIMB TO 17300 FEET"

$$St < Rt < Sp$$

which is a case of past perfect tense.

"I BELIEVE THAT THE DISTRACTIONS AND STRESS FROM HAVING TO HOLD SO LONG , ..."

$$St=Rt=Sp$$

which is a case of present tense.

"I THEN CALLED SEALORD TO CHECKOUT WHAT WAS GOING ON"

$$St=Rt < Sp$$

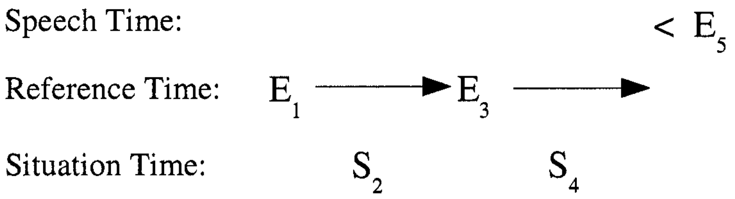
which is the simple past tense.

Certain types of events advance the narrative timeline by default, whereas others do not. This is the main phenomenon to be exploited here. Bounded events (Events in our situation type classification) advance the narrative time (Rt) by default. Unbounded events (States) on the other hand, do not. " *Narratives advance time dynamically. After the first sentence, the Events and States of a narrative are related to the previous events and times in the text, rather than to Speech time*" (Smith, 2003). Following is an example of a segment of narrative. The arrow "-

>" symbol indicates narrative story time advancement, i.e., Rt, and the brackets indicate situations. One thing to note here is that our minimum discourse unit, and therefore of events as well, is the syntactic *clause*. Clause boundaries will be determined through the preprocessing stage of the program, using a natural language parser. This example is taken from actual narratives as processed by the Battelle-PLADS tool; the "E" signifies Event, "S" stands for State.

" -> **E_1** [AFTER A COUPLE OF MINUTE , APPROACH ASKED OUR HEADING AND IF WE WERE GOING DIRECT TO SBJ .] **S_2** [I BEGAN TO SUSPECT A PROBLEM WITH THE COMPASS SYSTEM AT THAT TIME .] -> **E_3** [I INFORMED APPROACH THAT WE HAD A PROBLEM] AND **S_4** [I BEGAN TO TROUBLESHOOT WHILE THE FIRSTOFFICER WAS HANDFLY THE AIRCRAFT] ... **E_5** [AGAIN , IT IS IMPORTANT TO BE EXTRA VIGILANT UNDER THOSE CIRCUMSTANCES]"

The temporal progression of the previous passage is as follows:



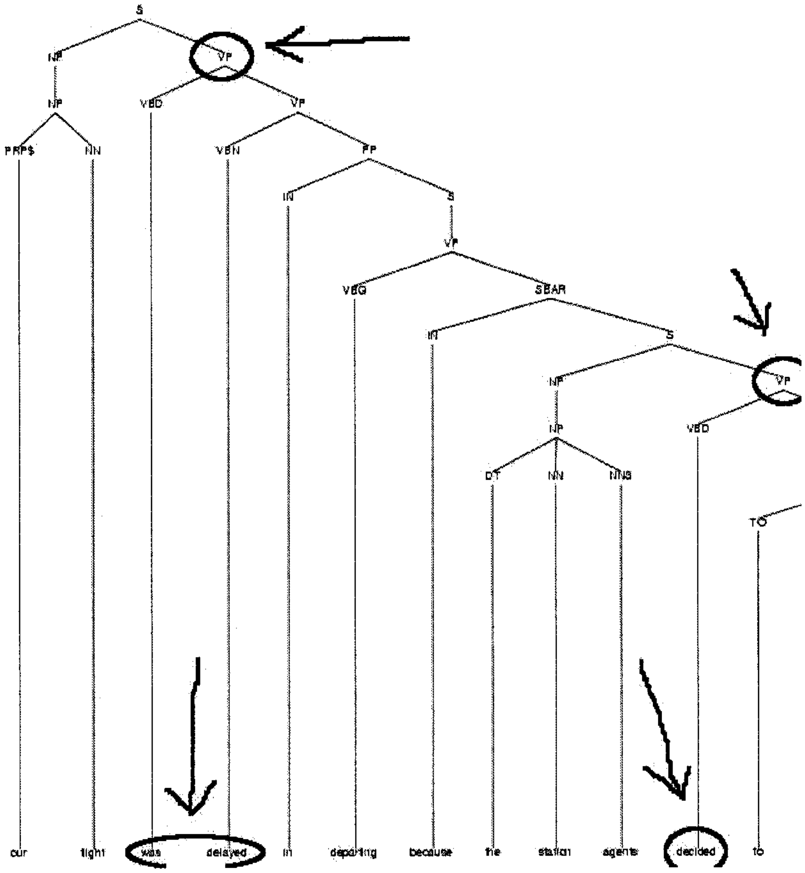
The main thing to notice in the diagram above is the inherent separation between Speech Time and the rest of the timelines. Reference and Situation times will always be previous to Speech Time, since that is an inherent property of the narrative mode of discourse and a constraint on language communication in general. The second thing to notice is that Situation time and Reference Time in this case are on the same timeline. The only difference is that E_1 and E_3 advance the narrative time, whereas S_2 and S_4 do not. This is the main difference between the two types of events, States and Events.

Syntactic Templates

There will be two sets of syntactic templates implemented: one capturing complete situation types at the *clausal* level, and the other capturing properties of the verbs, nouns and phrases, at the *lexical* level. Both sets will be based on a mixture from the formalization work done by Smith (2003), general linguistics semantics and pragmatic domain knowledge about the narratives. The most important features to be codified for tense and aspect are:

- The degree of volitionality of the subject noun.
- Telicity and dynamism of main verb.
- Spatio-temporal characteristics of complements and adjuncts (prepositional and adverbial phrases).

One thing to note here is that we are referring to main predicates. In other words, there will be an important distinction made between main clauses and subordinated clauses. This is one of the primary motivations behind using a natural language parser. Subordinated clauses will not be considered in the segmentation process, at least not as criteria for boundary detection. They will be used for determining subjectivity in the later stages of research (see Chapter 5, section "Future Research") . Consider the following parsed sentence



"Our flight was delayed ... because the station agents decided to ..."

The main predication verb is "was delayed" which reflects an imperfective viewpoint in the past tense (and in passive voice) therefore signaling a State. The other verb in the sentence belongs to an embedded clause which is linked to the main predication by a relational element ("because") and acts as a complement (causal) to the predication. Therefore it does not contribute to the main temporal progression of the narrative. We will explore this type of relation in future versions.

Our clausal syntactic templates will be subject to a phenomenon known as *coercion* (Smith, 2003), where optional entities override the main template and thus create so called *derived* situation types. The lexical templates consist of individual lexical entries (words) and constructions (phrases). These will be populated via the use of a lexical database and a set of external-file lookup tables. Templates, both lexical and clausal, will be implemented as data structures in a programming environment. Once populated these data structures (objects) will be processed by a matching algorithm to determine the narrative segment boundaries. All these processing steps will be described in the next sections. In sum, what this approach represents is a set of compositional rules based on semantic feature aggregates. These compositional rules represent situation types and are coded as syntactic templates. They are hand-coded *a priori* and matched in real-time to populated clauses (objects) in order to determine segment boundaries. This process will be described in detail in the next sections. Illustration 12 below illustrates the idea.

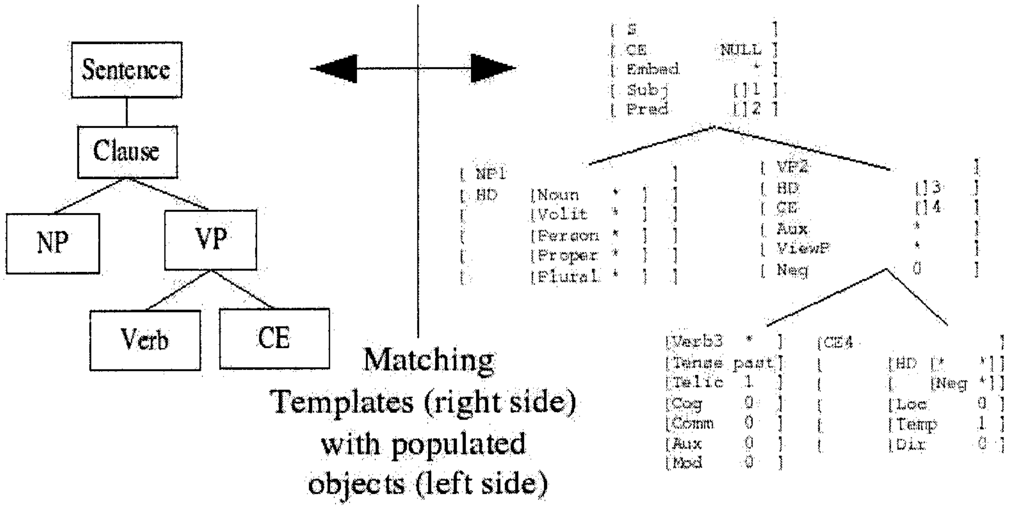


Illustration 12. Matching Templates to Objects

Templates are represented using feature structure notation. This is shown using square brackets "[]" and form matrices of attribute-value pairs. This is a very common notation for this type of conceptual structures and is easily converted into data structures in any programming language. Some examples of lexical templates are shown below; the full set of templates is shown in Appendix A. Note the distinction between a *template*, and a *populated* template (or object). A populated template is used as a live object to represent a narrative clause that has been parsed and populated by the program. A template on the other hand, is a pre-compiled structure that is used to match and produce a segment. That is, we have a set of manually-coded templates (Appendix A), that are matched against a set of real-time populated objects (Illustration 24 and

Illustration 25 of Appendix B and the run-time objects in Illustration 27 of Appendix C). When we make use of the word "template" in general, we are speaking of the conceptual structure without any values assigned to it, as shown in the following examples.

Some Examples of Lexical Templates

Noun

```
[ N        String  ]
[ Volit       +/-  ]
[ Person      +/-  ]
[ Proper      +/-  ]
```

The "+/-" indicates a binary feature that can take a value of either 1 (+) or 0 (-) exclusively. The need to have a separate "person" feature and "volitional"(Volit) feature has to do with the distinction between a *person* and an *entity* with volitional properties, such as the control tower or an airline call sign.

Adverbial Phrase

```
[ AP                   ]
[ HEAD      [[A String] ]
[           [[Neg  +/-] ]
[ Loc              +/-  ]
[ Temp             +/-  ]
[ Dir              +/-  ]
```

The square brackets within the square brackets indicate feature structure embedding (i.e., a reference to another structure of the specified type). The features above "Locative" (Loc), "Temporal" (Temp) and "Directional" (Dir) are mutually exclusive, i.e., constrained to only one value at a time being positive (+). The difference between locative and temporal is that of duration and not of physical versus temporal senses. Example: "Yesterday", "June 31 st", "presently", "now" are locative, whereas "meanwhile" and "after" are temporal, since they indicate temporal progression.

Verb and Verb Phrase

```
[ V          String ]
[ Tense      String ]
[ Telic        +/-  ]
[ Cog          +/-  ]
[ Comm         +/-  ]
[ AUX          +/-  ]
[ Modal        +/-  ]
```

```
[ VP                ]
[ HEAD         []V  ]
[ CE           []CE ]
[ AUX         []V  ]
[ ViewP   Perf/Imp ]
```

"ViewP" refers to viewpoint and is determined by the presence of auxiliary or modal verbs. Determining the telicity of a verb is non-trivial. Our heuristic will be that of considering verbs to be telic by default, except when a verb is either cognitive, communicative or has a coercive element (CE) that indicates so (more on this in the section "Object Buildup" of this chapter). CE elements could be adverbial or prepositional phrases. "Cog" refers to cognitive verbs, such as "think" and "believe". "Comm" refers to communication verbs such as "say", "tell", "inform", etc.

Some Examples of Clausal Templates

Clausal templates are responsible for determining the segment boundaries. This occurs through a matching process with a set of pre-compiled templates (as shown in Appendix A). In other words, once populated from the narrative text, lexical templates aggregate to form clausal templates, and are then matched against the pre-compiled set. The curly brackets "{}" indicate clause boundaries and the square brackets "[]" indicate templates (templates are greatly simplified here to show only their distinguishing feature, for ease of visualization purposes). The ellipsis indicates any number and type of elements. Note that the order of the elements is not critical here, any ordering will do. The important factor is that elements modifying a given entity are within that entities dependency reach and the we match the right number and types of entities. This notation is a simplification of the full expanded notation of Appendix A.

Events

```
{...VP[[Telic +]...[ViewP Prf]...[CE Ø]]...}
```

" WE LANDED NORMALLY ON RWY 19"

Events are the default classification of clauses. Here we have a typical event template, where there is a main predicate verb with a perfective viewpoint (aspect), no coercive element in the main verb phrase, and a telic verb, which indicates the boundedness of the event. Again, the full set is shown in appendix A.

States

```
1.  { ... VP[[Telic -][ViewP Perf]]...}
```

"BY THEN I WAS AT THE ..."

```
2.  {... VP[[Telic -][ViewP Imp]]...}
```

"WE WERE HAVING TROUBLE ..."

Cognitive verbs trigger a typical State, even though their viewpoint could be perfective. Such verbs can also be subject to any type of coercive elements and still retain their State classification ("any" here is denoted with an "X")

```
3. {...VP[[Cog +][ViewP Prf][CE X[]]]...}
```

".AT THE TIME I THOUGHT ANOTHER AIRPORT..."

Non-volitional agent as the subject of the predication

```
4. {...NP[[Volitional -][Person -]]...
        VP[ [Telic -]]...}
```

"OUR COLLISION AVOIDANCE SYSTEM SHOWED NO OTHER AIRCRAFT IN OUR VICINITY AT OUR ALTITUDE"

```
5. {...VP[[ Comm +][AUX Neg +][Tense Past]]...}
```

"HE DID NOT SAY WHY ..."

This last example is an instance of a negated verb phrase with a cognitive or communicative verb head. In this case, a communication verb such as "say" becomes a State as opposed to an Event, since it represents a "anti-fact", meaning, a state of knowledge about facts that did not occur. In other words, it is considered at the same level as a hypothetical construction such as

```
5. {... VP[[Modal +][ViewP Prf]...]...}
```

" HAD I MADE A RIGHT TURN IMMEDIATELY AFTER TAKEOFF I WOULD HAVE BEEN CLEAR OF THE TERMINALCONTROLAREA"

This constitutes information that is only present in the operators mind, therefore influencing his behavior. Even though it is not part of the system's state, at least not an objective parameter, we will still consider this as indicating a State (a state of the system in the Scenario model) and not an Event.

Implementation

The method presented in this document has been implemented computationally via a program. This program was written by the author using the PERL language (version 5.1), on a Linux RedHat Professional Workstation (version 3), running on an Intel Pentium III processor. The program consists of three main components:

1. Preprocessing. Prepares the data for syntactic parsing (including acronym expansion, spelling corrections and typos). This portion of the program was designed as a placeholder and therefore it does not represent the emphasis of this project. Ultimately, a more sophisticated tool such as Battelle's PLADS will be employed.
2. Object buildup. Builds the object trees based on the syntactic parsing output. Object trees are isomorphic to syntactic trees. In addition, they contain the information required for our analysis method. Objects are populated using a lexical database and external lookup tables.
3. Analysis. Populated objects are matched against manually-coded

templates. The text content from the object trees is output to external files. Matching is represented as XML tags around segments.

Each one of these components is organized as a series of PERL modules, which are in turn organized under separate directories. The total number of lines of code for the complete module set is approximately 6,500 including comments. The modules are named: preprocess, analysis and segmentation respectively. The file labeled "main.pl" is the main entry point for the program. It serves as a switcher to enable or disable any given portion of the program. For instance, by turning the right combination of switches on and off the user can run only the preprocessing stage of the program, and not the rest. This file also controls all output and input to external files. The command line arguments for main.pl are as follows

```
./main.pl -i <narrative-file-name> -s <random-sample-number or
  record-list-file> -o <output-directory-name> -a <analysis-
directory-name> -x <segmentation-output-directory-name> [-v to set
                         debug on]
```

The program accepts a large flat text file as the input, assuming it contains the narratives in some predefined format (either from the AeroKnowledge CD, or as produced by Battelle, see Appendix E). It also accepts an integer number representing a random sample size (for random sampling), or the name of a file containing a list of record numbers to be selected (this is mainly used for debugging). The user can specify a path to a directory containing the preprocessed narratives; this is used for the second and third portions of the program. Finally, a verbose (debug) flag can be turned on to produce diagnostics. All functionality for the main.pl file and for the rest of the PERL module files (xxx.pm) is very well commented, making it clear for the reader to follow the functionality of the code. It is important to note that each record narrative will be written out to an individual flat text file at the specified location, and the record number will become the file label. This was done in order to enable the final step which is analysis, so that each record can be considered as a discrete object. The final segmented XML output can be produced on a single file or separate files, as specified by the user.

The next three sections will describe in detail each one of the separate components of the implementation program. They will also discuss the Unified Modeling Language (UML) diagrams presented in Appendices B and C.

Preprocessing

The text is first converted into all lowercase and tokenized. Tokenization refers to the process of separating each word into its smallest units of functionality. For instance, the word "can't" is separated into "can" and "n't", two tokens with different functionality in language (verb + adverbial negation). It is important not to confuse this process with what is known as *stemming*. Stemming is the process of separating lexical stems from inflections, such as pluralization "s" or tense "ed". The tokenization portion of the program was implemented following the University of Pennsylvania's Treebank notation recommendation (Treebank, 2004). The reason for the conversion of the ASRS narrative text into all lowercase is the fact that the original text in the database is all uppercase (see Appendix E). Our program relies on several external files and program which are case-sensitive and thus would not work properly unless the text is converted. This comes into play at the second stage of the preprocessing. The second stage consists of the use of an external lexicon file in order to lookup every word in the text narratives. This lexicon file is a combination of the Brown University and the Wall Street Journal Corpus (Penn Treebank, 1999), which is included as part of the release of the Brill Tagger (see Appendix D). This lexicon file consists of a compilation of approximately 100,000 words (tokens), and a list of part-of-speech tags (POS) appended to each entry, ordered by statistical frequency of occurrence. This file is case-sensitive, so it is treated before using the program by merging capitalized words with the non-capitalized equivalents, giving priority to the latter since they are more frequent.

The next step is to use domain-specific lookup tables (LUTS) to expand common acronyms, abbreviations and codes. These lookup tables were compiled from material available in aviation sites in the World Wide Web (AirOdyssey) and Federal Aviation Administration material (FAA). The choice of source type was made for processing reasons, since text in the World Wide

Web is available electronically. These LUTS are also case-desensitized beforehand for consistency.

Next, the program attempts to identify and expand the rest of the words that were not processed in the previous two steps. For this, it uses the tool GNU Aspell (see Appendix D), which doubles as a full dictionary and a typo correction program. It has a fairly sophisticated algorithm that uses *sound-alikes* and heuristics on common spelling and typing errors. Some of these errors are due to keyboard key proximity, vowel omissions, and order (the reader is encouraged to refer to the literature on this program for more details). This part of the process is complemented with a set of heuristics obtained from the literature as described in the last section of Chapter 2.

These heuristics are as follows:

- Contractions
    - All vowels missing from a word, e.g., "bck" for "back"
    - Missing contiguous sub-strings with vowels and possibly consonants, e.g., "ctl" for "control"
    - Words beginning with 'x', which translate to 'cross' or 'trans' or 'out' in that order, e.g., "xponder" for "transponder"
- Truncations
    - Beginning of word (not very common)
    - Middle, eg., "apch" for "approach"

- End, eg., "capt" for "captain"

The final step in the preprocessing stage is that of assigning preliminary part-of-speech tags to the lexical entries and adding them to the main lexicon file. This is done with a set of hard-coded rules and with the use of WordNet, which is a lexical network implementation (see Appendix D). The rules employed are based on domain knowledge and are summarized below (for the meaning of each POS tag, refer to Appendix F). The "{}" represent lists of POS tags, in order of frequency.

- Numerical values are tagged as CD, including decimals
- Numerical ranges are tagged as { NN JJ RB }
- List Markers as LI
- Acronyms as proper nouns NNP
- Compound nouns as NN
- Dates as RB
- Fractions as { JJ RB }
- Special cases, such as "fl290" which translate as flight 290, pre-tagged as NNP

The complete flowchart of the preprocessing stage is show in Illustration 13 below.

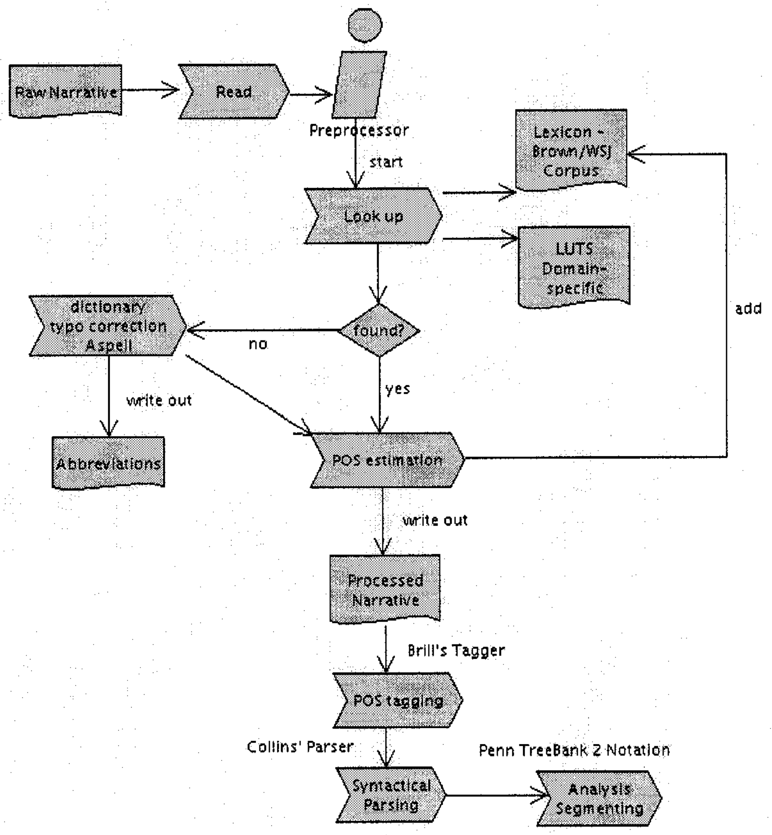


Illustration 13. Preprocessing Flowchart

Object Buildup

The second stage of the implementation program consist of building tree objects from the result of the syntactic parsing. The output from the parser must conform to Penn Treebank 2 notation (Treebank, 2004). This ensures that the main portion of the program, the analysis, can be used with any preprocessing software (such as PLADS). In this way the preprocessing stage remains a placeholder in the context of this implementation and is not the primary focus of this work. The syntactic parser used for this implementation is Michael Collins, a widely used world-class statistical context-free grammar parser. The output from this parser is for the most part compatible to Treebank 2 notation, except for some minor differences which are taken care through a small script provided in the release package.

The data structures required to build the object trees are shown in Appendix B, and the rest of the program structure and design are shown in Appendix C. All these diagrams are in UML standard notation omitting procedural information for simplicity sake. The data structures can be divided into two classes: *lexical* and *clausal*. Lexical refer to individual words and phrases, and clausal refer to single clauses and complex sentences. Following is a description of the main features captured by each of the lexical data structures.

- Noun and Noun Phrase: captures the volitional qualities of a noun and its modifiers. A given noun can represent a person or an object which stands for a person or a group of people, therefore having volitional qualities, e.g., control tower. Other objects like instruments are non-volitional even though they might appear as subject in the narratives, e.g., *" ... the tcasi informed us that..."*
- Verb and Verb Phrase: capture the dynamic and cognitive qualities of the predication. The distinctions made are between cognitive,

perceptual and communicative verbs, and verbs of movement and duration.

- Adverbial and Prepositional Phrase: these act as coercive elements to the main predication. What is captured here is the quality of Location, Temporality and Directionality, which can coerce a static verb into being dynamic or vice versa.

The population of these structures occurs via two processes: WordNet and lookup tables (LUTS). The use of WordNet is the primary source of knowledge in the population process. The idea behind WordNet seen as a lexical network, is the existence of certain nodes which are called unique beginner *synsets* (synonym sets). These synsets are nodes which give origin to the hierarchies present in the network. They correspond to the more basic and abstract concepts that are represented in WordNet (a.k.a. *semantic fields* or domains). These beginner nodes are used in our implementation to determine the basic properties of nouns. The total is 25 and we are using 13 of them. They are shown in Table 4 below as obtained from Miller (Miller, 1990). They show the mappings for our population process.

Noun feature	WordNet synset

Volitional	{ person, human being }
	{ cognition, knowledge }
	{ communication }

Event (dynamic)	{ event, happening }
	{ process }
	{ act, action, activity }

Object (non-volitional)	{ artifact }
	{ collection, group }
	{ attribute, property }
	{ natural object }

Location	{ location, place }
	{ time }

State (stative)	{ state, condition }

Table 4. Unique Beginner Synsets

The searching algorithm for nouns is the following. For any given noun, if inflected, the stem is first obtained through a method available in the WordNet implementation interface. If the word is a compound (i.e. hyphenated), then the last word is evaluated first and if no classification is found, the next word up is evaluated in turn (e.g., for "instrument-landing", first evaluate "landing," then "instrument"). The program searches all possible senses available for the target word, starting with the most frequent. This occurs by default given the frequency count and ordering inherent in the WordNet database. For each sense evaluated, the program searches the WordNet space by following the *hypernym* chain. Hypernyms are part of a two-way semantic (as opposed to lexical) fundamental relationship represented in WordNet. This relationship is that of Hyponyms/Hypernyms (a.k.a. ISA relationship).

> *Hyponymy is transitive and asymmetrical ... and, since there is normally a single superordinate, it generates a hierarchical semantic structure... Such hierarchical representations are widely used in the construction of information retrieval systems, where they are called inheritance systems ... a hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate and from any other hyponyms of that superordinate ... This convention provides the central organizing principle for the nouns in WordNet (Miller et al., 1990)*

The process stops when the program encounters one of the specified synset labels and assigns it the corresponding category as shown in Table 4 above.

Verbs are evaluated in the program in a similar way, except that the organization of verbs in WordNet is slightly different. Verbs originate from 8,400 synsets and are grouped into 14 lexicographer files, which correspond to different semantic fields or domains. These correspond to concepts such as bodily care and functions, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social interaction, and weather (Fellbaum, 1990). There is a 15th file, which does not correspond to any semantic field, and that groups verbs normally considered statives (i.e. States). All other files contain verbs of action or events, which posses dynamic qualities. Our program algorithm performs a similar search as for the case of nouns, following the hypernym chain for verb stems from most frequent to least, until a specified label is found. It is important to note one difference here. Hypernym for the case of verbs has a different meaning than that of nouns. It is impossible to defined an ISA relationship among verbs, so instead hypernymity is defined as "troponymy", which stands for "X is a form of Y". The end result then is that verbs are related via hypernymity through a *manner* relationship. In other words, X is a hypernym of Y if Y is a manner of doing X. The following table (Table 5) shows the relationships used in our implementation and their mappings to the semantic fields in WordNet.

Verb Feature	WordNet Semantic Field
Cognitive	Cognition and Psych Verbs (comprehend,cogitate,cognize, think,believe,decide,act,do)
Perceptual	Perception Verbs (perceive,sight,notice,observe)
Action/Causal	Verbs of Change, Contact and Consumption Verbs (cause,change,execute,alter,modify,revert, attach,touch,hit)
Communication	Verbs of Communication (verbalize,utter,communicate, express)
Movement	Motion Verbs (locomote,move,travel,displace)
Existential	Verbs of Possession and Stative Verbs (be,become,exist,have)

Table 5. Verb Features and Semantic Fields

Adverbs are much more limited than nouns and verbs in WordNet. They are also not well covered in the literature. Our program takes the following approach. For any given adverb, we obtain the adjective that originates it using the *pertainym* relationship. Pertainymity is the relationship that relates an adjective to a noun, in the case of relational adjectives, or in our case, an adverb to its adjective. If an adjective is found the program then proceeds to search for synonyms for each sense available. The main difference here with respect to the previous approach to nouns and verbs, is that the searching occurs among synonyms of a given sense, instead of through the hypernym chain. This is due to the fact that hypernyms are not defined for adjectives or adverbs. Adjectives are organized in WordNet around synonym/antonym clusters. The idea behind this organization is that adjectives are better *defined* through this type of relation. Each cluster is related to a noun through a pertainym relation as well, such as in the case of "tall/short" which relates to the concept "height". The population procedure via WordNet for adverbs, is complemented by a lookup table. This table is hand-coded after iterative runs of the program for any given pool of data. The population of prepositional phrases is done exclusively through lookup tables, since there is no support for prepositions in WordNet. A sample of two of these lookup tables is shown in Table 6 below.

```
# this file contains the major prepositions
# used. Each entry has a list of possible
# semantic attributes, used to populate
# coercive element structures
to locative
as manner
by agency
within locative
on locative
off directional
in locative determinative
with locative complementary
into directional
at locative temporal
from directional
after temporal
inside locative
about locative temporal
around locative temporal
before temporal locative
during temporal
because causal
for temporal relational
```

```
# this file contains the major adverbials used by the program.
# Each entry has a label indicating its semantic attribute, used to populate
# coercive element structures
always temporal
usually temporal
never temporal negation
actually manner
then temporal
```

Table 6. Lookup Tables for Adverbs and Prepositions

Analysis

The analysis portion of the program consists first of building a predefined set of clausal templates, as described in Chapter 3. Then the program matches these templates against each clause that was parsed and populated from the narrative text. This process is repeated for every record. It is, in effect, an iteration through all the trees obtained in the previous step (Object Buildup), and a matching operation on each one of them. The complete set of predefined templates for matching is shown in Appendix A. They can be grouped into three sets: States, Events and Assessments (corresponding to their respective segment types, to be defined shortly). Overall, the main distinction among the three types is that of the linguistic and mental event they are trying to capture. This was described conceptually and theoretically in Chapter 3. Assessments correspond to *general statives*, such as statements about common knowledge or knowledge present at the time of the production of a given narrative. Therefore it signals a shift in the temporal anchoring to Speech Time (as described in Chapter 3). It can be seen in Appendix A that all templates in this group have an element that is either in the present tense or in a modality known as *irrealis*, which stands for a plausible situation that has not occurred. The set under the label Events, are characterized by the lack of auxiliaries, the so called perfective viewpoint. They also show some cases where a coercive element, such as an adverb or prepositional phrase, turn an otherwise State into an Event; these are called *derived* situation types (Smith, 2003). This coercion occurs because an otherwise unbounded event becomes *bounded* by a modifier, such as a prepositional phrase (e.g., "at 2 o'clock") or an adverb (e.g., "immediately"). All Event templates have a simple form of past tense, indicating that they are anchored in the narrative timeline (Rt). States have combination of auxiliaries and tense types, indicating that States could also be anchored in Situation Time (St) (e.g. *"... i thought that we **had** completed the turn quickly..."*). These templates are more complex, showing a combination of viewpoints, tenses and auxiliaries. They were developed manually to capture events which are stative in nature and unbounded. For instance, we consider States to be any event which portrays a cognitive or perceptual activity, such as *"i **thought** that we had completed the turn quickly"* or *" i **noted** that the 2 segment distances prior to pm totaled 71 mi "*.

The matching of pre-compiled templates with populated objects occurs in the following manner. Part two of the program, the Object Buildup, produces isomorphic object trees from the syntactic trees obtained in the preprocessing stage, as shown in Illustration 27 in Appendix C. There is one object tree per each clause in a sentence, and sometimes there are multiple clauses per sentence (i.e. coordination). As it turns out, the narratives from the ASRS database contain a high density of coordinated sentences. Our algorithm separates each clause and processes it individually, i.e. with its own tree. Each one of these object trees is populated according to the algorithm described in the previous section. The next step is the matching. For this our program utilizes a common pattern in software modeling known as the Visitor behavioral pattern (Gamma et al., 1995). This pattern is shown in Illustration 14 below.

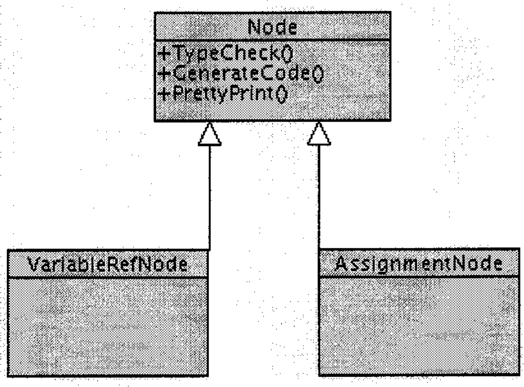


Illustration 14. Visitor Pattern

The implementation of this pattern in our program can be seen in Illustration 26 in Appendix C. The main purpose of this pattern is to allow for the traversing of a hierarchical structure, such as a tree, and to perform operations on the nodes of this structure without having to affect its logic. In other words, it decouples the operation logic from the data structure. It is a very common technique in compiler design, where syntactic trees are the main data structures. So in this sense it shares many similarities with our problem. The program uses this pattern to create a visitor class, which contains the instantiation of all the pre-compiled templates mentioned previously and described in Appendix A. The sub-classes of the visitor class, are visitors for particular entities in our program's data design. There are two visitors types: a clause and a sentence visitor. The latter simply iterates through sentences and invokes a clause visitor for every clause in the sentence. The clause visitor then tries to match each one of the pre-compiled templates against the current clause. The matching occurs through overloaded *equality* operators, implemented as methods of every class in the data structure hierarchy (Illustration 24 & Illustration 25, Appendix B). Illustration 28 of Appendix C, illustrates this process. We can see an object tree on the left hand side, equivalent to Illustration 27 in Appendix C. This tree is matched through the clause visitor by invoking the equal method on the clause and passing the template as the argument to the method. Effectively this causes the current clause to compare itself to any given template, which in turn, invokes the equal operator on every element within that clause, i.e. subject, predicate, coercive elements and so on. Each one of these elements in turn invokes the equal operator recursively until all elements are compared to all the equivalent elements for any given pre-compiled template. Wild card operators "*" are used through regular expressions to match any values. The procedure will fail if any one of the elements being compared contains a feature that is non-matching. It is important to note here that the algorithm matches all templates sequentially, so care needs to be exercised to ensure that the templates are mutually exclusive. In other words, any given clause must not match more than one template. This was done by careful examination of the output of the program. There are currently no overlapping sets of feature in the templates presented in Appendix A. This was done by not allowing features of elements to contain wildcards, unless that particular feature is non-discriminating among templates. Effectively this creates negative filters for the matching. More specifically, if two data structures differ only on one feature, for instance telicity, then they must not contain "*" in that feature. As an example we can see from Appendix A that Event 3 and State 2 are for the most part identical, except for the feature of telicity on the predicate's main verb.

The matching algorithm cannot be written in pseudo-code. This is because the program was not designed in that way. In the computer science literature, is it customary to design algorithms and explain them through pseudo-code. Algorithms are usually considered the "meat" of a program, or at least the most interesting part of it from a mathematical or computational standpoint. This is not the case here. Our program was designed from a software engineering perspective. This can be seen from the number of illustrations in the collection of UML diagrams included in Appendix C. In the software engineering field the concept of *Object-Oriented Design* (OOD) promotes the idea that data structures should be as devoid as possible of complex and lengthy logic. Instead, objects in an OOD architecture balance the computational logic load by distributing responsibility throughout the class hierarchies and using messaging and semantic interaction. In other words, the larger portion of the logic is embedded in the *design itself* of the data structures and the interactions of the program. This is the same spirit here, as can be seen from Appendix C. Therefore it is not as relevant to shown an algorithm for the working program or the matching process, but it is important to show its architectural design. Illustration 15 below shows the matching portion of the program using a sequence diagram in UML notation.

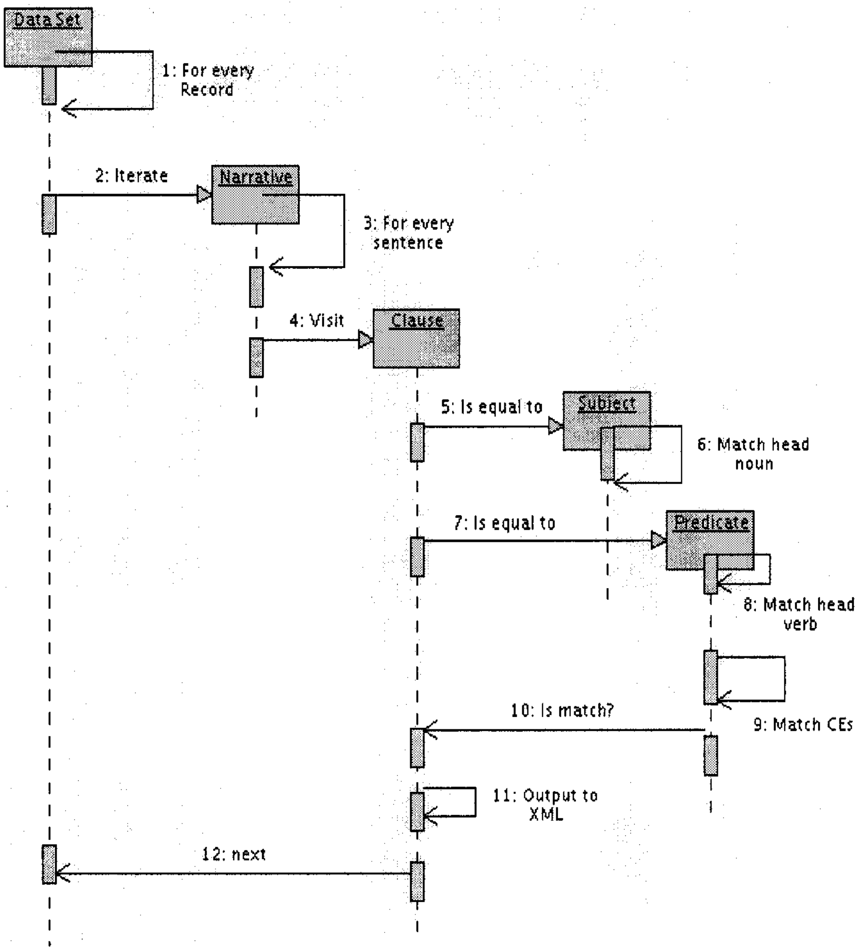


Illustration 15. Matching Sequence

The equivalent implementation to what could be considered an "algorithm" occurs *recursively*, by means of the Object-Oriented hierarchical design as shown in Appendix B and C and described in the last two sections. In effect, the program does recursion, instead of loops, and the messaging needed for the recursion occurs through the inherent data structures design.

Once a match occurs, the contents of the clause being evaluated are output to file in XML format. The XML Document Type Definition (DTD), is shown together with a sample output file in Appendix E. It is compatible with the DTD provided by Battelle-PNWD to ensure usability with their technologies, as well as promoting the validation process. Chapter 4 will cover this validation process, where a sample of 100 narratives obtained from the same sample pool as Battelle-PNWD is analyzed. The output DTD differs only in the declaration of five new XML elements (i.e. XML tags). These are:

- <St> represents a Setting segment, which stands for a State event type that occurs at the beginning of the narratives and forms a contiguous block.
- <S> represents a State segment, as defined by the templates presented in Appendix A and the conceptual structures defined in Chapter 3.
- <E> represents an Event segment, as defined by the templates presented in Appendix A and the conceptual structures defined in Chapter 3.
- <A> represents an Assessment segment, as defined by the templates presented in Appendix A and the conceptual structures defined in Chapter 3.

- <X> represents an Unknown segment, which stands for any clause that did not match any of the templates available, or a clause which has a bad parse tree.

The last type of element, "<X>", represents cases where none of the pre-compiled templates matched a given clause. This occurs in approximately 11% of the clauses parsed in a preliminary test trial of 15 narratives, randomly sampled from the ASRS database. We expect this ratio to remain roughly the same for the final trial of 100 narratives. It is due to what is known as *fragments* in syntactic parsing and to overly complex sentences. Fragments are sentences that lack a complete syntactic composition, such as a missing subject or predicate (not counting tacit subjects). Complex sentences can be grammatically poor sentences or non-sentences which occur in highly irregular free text such as in our case. The solution or improvement to this problem will come with the replacement of the syntactic parser. This will take effect in subsequent research phases. The new parser to be used will be Stanford's Lexicalized PCFG Parser, which performs more robustly in face of complex sentences (see Chapter 5).

The XML output from this section of the program will be run through a small script to count the number and type of segments in each record (i.e. narrative). The segment count will be used for the analysis of results and the validation portion of the work, as it will be described in detail in the next chapter.

CHAPTER 4

RESULTS

This chapter begins with a description of the statistical method we employed for uncovering the patterns in our data. It then covers the selection process and the description of the variables employed in our analysis. Following, the results from clustering analysis are presented with a review of the most prevalent characteristics of the clustering arrangement, in terms of emerging patterns. Finally, two new variables are introduced, which are aimed at detecting levels of subjectivity and potential for behavioral biasing. A selected set of narratives is evaluated and compared with respect to these two new measures.

Clustering Algorithm

In order to evaluate the results from our methodology and implementation program, we will employ a nonparametric statistical technique, known as Partitioning Around Medoids (PAM). This is a clustering algorithm developed by Kaufman & Rousseew (Kaufman & Rousseeuw, 1990). It is implemented in the statistical environment "R" (see Appendix D), and runs in the Linux platform (same version and processor type as the main analysis program. As a clustering technique, PAM makes no assumptions about the underlying distribution of the data, hence the term nonparametric. Instead, PAM is used as an *exploratory* tool for discovering patterns and groups in data. This is in contrast to the more common parametric statistical techniques, such as ANOVA, which are mainly used for inferential or confirmatory purposes. With clustering, there is a also considerable amount of latitude in terms of which type of technique to apply for a given data set, and in terms of the optimal values for the parameters in question. Within all the clustering algorithms available, there is a fundamental distinction between what is known as *partitioning* versus *hierarchical* methods. Both approaches work around objects which are characterized by a *set* of features. Each feature corresponds to a variable and each object has a set of values corresponding to each of those variables (observations). For instance, one could measure the per season average temperature, barometric pressure and average precipitation for a set of cities in a given geographical region. Then using a clustering algorithm, one could obtain clusters of cities which share similar meteorological characteristics. This experiment design would provide the analyst with a picture about groups of cities that share similarities in climate. Each city represents an object and each of the measures represent a value for each of the variables, for each and all of the objects. As it turns out, clustering techniques work better for *qualitative* measures rather than continuous quantitative measures.

Hierarchical algorithms proceed iteratively by pairing or partitioning objects until a hierarchy of clusters is obtained. The results of this type of technique is best visualized as a "tree" (in the mathematical sense). This is one of the reasons why this approach is favored by many. Partitioning methods on the other hand proceed by selecting representative objects, and then trying to partition the complete data set in the best way possible around these objects, using similarity measures. They try to find the best solution for a given number of groups, as specified by the analyst beforehand. The ultimate difference between the two clustering approaches is a subtle one, since the end result is the same, namely to expose natural groups underlying the data. The difference is in the way these groups can be visualized, and therefore conceptualized, and the road to that visualization. A hierarchical method will show on the one end *n* groups, one for each of the original *n* objects, and on the other end, one single group where all objects converge. A partitioning method instead, will partition the data set in the best way possible to find *k* groups, as requested by the user of the algorithm. Hierarchical methods allow the user to visualize *all* groups possible, and hence, give a better overview of the relations existing in the data. Partitioning algorithms provide access to *representative* objects, which are very useful in characterizing or reducing data.

We chose to use a partitioning method because of the idea of a *centrotype.* A centrotype could be considered an exemplar of a given group, in that it best characterizes the features of that group, and also because it provides a "norm" or baseline to be used for comparison purposes in our case. In particular the PAM algorithm is a very robust one, employing the idea of medoids instead of centrotypes (or centroids), which are more insensitive to outliers in the data. Medoids are calculated using the *mean absolute deviation* as opposed to the more common standard deviation (which uses squares of residuals instead of absolute values). Mean absolute deviation has been shown to be more stable in the face of large values, thus making the PAM algorithm also more stable in that sense. The PAM algorithm proceeds as follows: a number of clusters *k* is specified beforehand, and the program selects a set of k elements from the data. These are considered as the k most representative elements for each group. Each of these representative objects is selected as to minimize a measure of the distance (similarity) to the rest of the objects in the set. Then in order to obtain the corresponding clusters, each of the remaining objects are assigned to the nearest representative object in terms of the same distance measure. Clearly, not every selection of *k* groups yields the "best" or more "natural" grouping, so this is an iterative process and the representative objects need to be selected so that they minimize the average distance to every object in their group. This method tries to find *spherical* clusters, roughly ball-shaped groups if we consider the data space to be an Euclidian space. The program consists of two phases: *buildup* and *swap*. During the first one, the procedure just described is performed until all elements are assigned to a group. Then, the swap portion of the algorithm swaps elements in clusters and re-evaluates the average distances in the clusters in order to find improvements. This is also an iterative phase. The algorithm is deterministic in the sense that it does not depend in the order in which the objects are evaluated, as long as there are no overlapping solutions, which is very uncommon (i.e., same k with different average distances). The details of the algorithm are described in full length in the author's book(Kaufman & Rousseeuw, 1990).

Selection of Variables

Before proceeding to describe the results we obtained, we need to describe the types of variables we used for our analysis. Since the output of our program are segments, our variables consists of counts of those segments. In particular, we chose to employ variables which correspond to aggregates of those counts, so that they represent meaningful distinctions in our behavioral and narrative model. Given our four different segment types (Setting, Event, State and Assessment), one of our variables consists of the ratio of the count of contiguous blocks of segments of type Event and of type State. We will label this variable as <u>E to S Block Ratio</u> (E.to.S.block.ratio). More specifically, our analysis considers the *average* size of contiguous E-type segments, over the average size of contiguous S-type segments. What we are measuring here is one of the components of the *style* we defined earlier: grammaticalization of the portrayal of event sequences. This variable is of type *ordinal ratio*, which means that it is a fractional value and always positive. The PAM program actually treats this type of variable as continuous ordinal data, that is, as continuous data in some interval scale, by applying a logarithmic transformation and switching to rank. This transformation is done by the program and the details of the process are described in the literature (Kaufman & Rousseeuw, 1990). The next variable is labeled Canonicity (Canon), and reflects the degree to which a given narrative has a conventional structure consisting of an introduction, a development and a conclusion. In our case, this translates to the presence of at least one St type segment at the beginning and at least one A-type segment at the end of a given narrative. That is, it represents how well structured a given narrative is compared to the canonical form of the narrative style. This canonical form is present in our data, but the question is "to what extent?". Therefore our variable *Canon* can have values of 0,1,2 or 3, where 3 is the highest in terms of "canonicity" (3 = Setting+Conclusion 2 = Conclusion only, 1 = Setting only, 0 = neither).

The program DAISY, packaged together with PAM, is used to compute the similarity coefficient and obtain a similarity matrix used to compute the clusters. Actually, the program computes *dissimilarities* instead of similarity measures for mathematical efficiency. As it turns out, because all of our variables are of the same type, and because they are ordinal, DAISY transforms the values into ranks, normalizes them, and computes dissimilarities based on the Manhattan distance. The Manhattan distance is a variation of the geometric Euclidian distance, and consists of the sum

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

where each x_{ij} corresponds to a point in an Euclidian geometric space. Therefore this sum corresponds to the shortest path from point A to point B given that the only trajectories possible are orthogonal segments (i.e., city blocks, hence the name). Ranking and normalization are performed to neutralize the different weighting that each scale might have for each different variable. Then a 'źscore is obtained from

$$z_{if} = r_{if} - 1 \ / \ M_f - 1$$

where r_{if} is the rank for the i^th^ object in the f^th^ variable, and M_f is the highest rank for variable f. Z has values between 1 and 0 only.

Clusters

Illustration 16 shows the results of grouping the output data from our program. The graph depicts a "Silhouette" plot. A silhouette plot is a type of graph designed by the authors of the PAM algorithm, and it was designed as a way to represent, in two dimensions, a multidimensional statistical analysis such as partitioning clustering. The gray bars represent a measure called the "s" coefficient. This is a dimensionless measure plotted on the horizontal axis. Each of the groups of gray bars represents a cluster. Their proximity in the graph is *not* equivalent to their proximity in measure. The s coefficient is calculated with following equation

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$$

where a(i) is the average dissimilarity of object i to *all* other objects in a given group A, given that $a(i) \in A$. b(i) is defined as the minimum distance from object i to all other objects in group C, where $C \neq A$ (i.e., the distance between groups). Therefore,

$$-1 \leq s(i) \geq 1$$

A nice feature of these silhouette plots is that they are able to detect *artificial* or "forced" groupings. The average silhouette width is shown next to each cluster, and a grand total average is shown at the bottom of the graph (Illustration 16). The s coefficient can be thought of as a measure of the "compactness" of a cluster or clustering arrangement. That is, a measure of both the closeness of all the members of a group to their medoid, and also of their separation to other clusters.

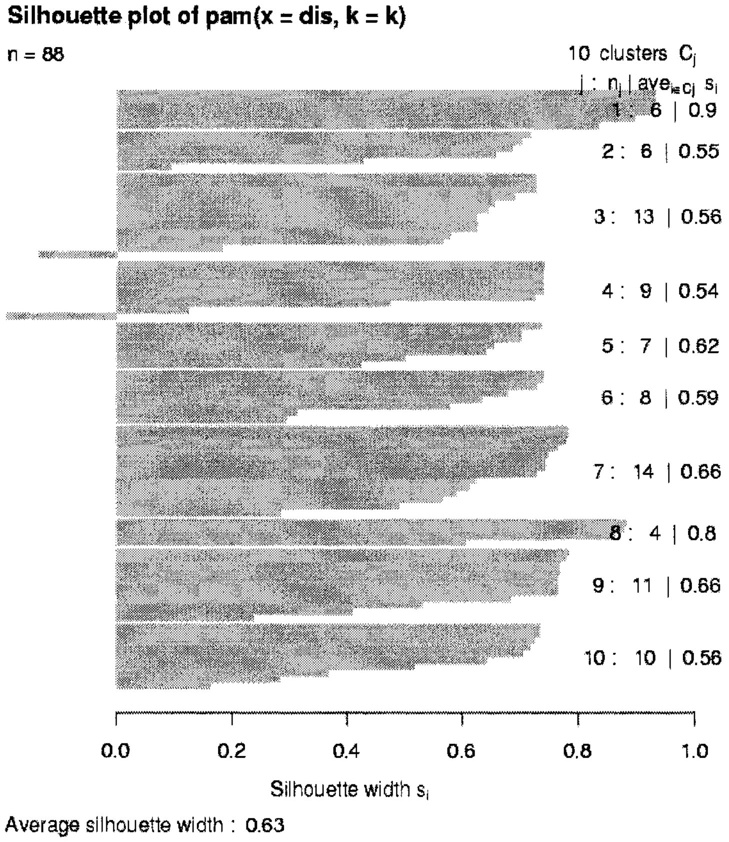


Illustration 16. Silhouette Plot of Cluster Data

Therefore when the number of groups is artificially too low, the plot will indicate so by showing narrow silhouettes. This is caused by the small a(i) factor arising from a large *within* dissimilarity produced by the forced "fusion" of naturally smaller clusters into larger ones. When the number of clusters k is too large, the result is also narrow silhouettes, this time caused by small *between* dissimilarities, and therefore small b(i) values which also cause a small s(i) value overall. We can expect to see this effect around the origin of the horizontal axis when the number of groups is not adequate. Conversely, we expect to see large gray bars when the number of clusters is close to the "natural" grouping of the data, i.e., s(i) values closer to 1. As stated by the authors (Kaufman & Rousseeuw, 1990) and based on their vast experience using clustering methods, values in between .5 and .7 are considered normal and satisfactory for clustering configurations. Values between .7 and 1 are considered optimal and anything below .5 is poor. Our plot shows ten clusters, with a total average width of 0.63. We can readily see that our plot shows at least 80% of our data falling between the .5 and 1 mark. We can also see that there is at least one extremely good cluster (1) consisting of six records and another one (8) with four members. None of the clusters fall below the .5 mark in terms of their averages.

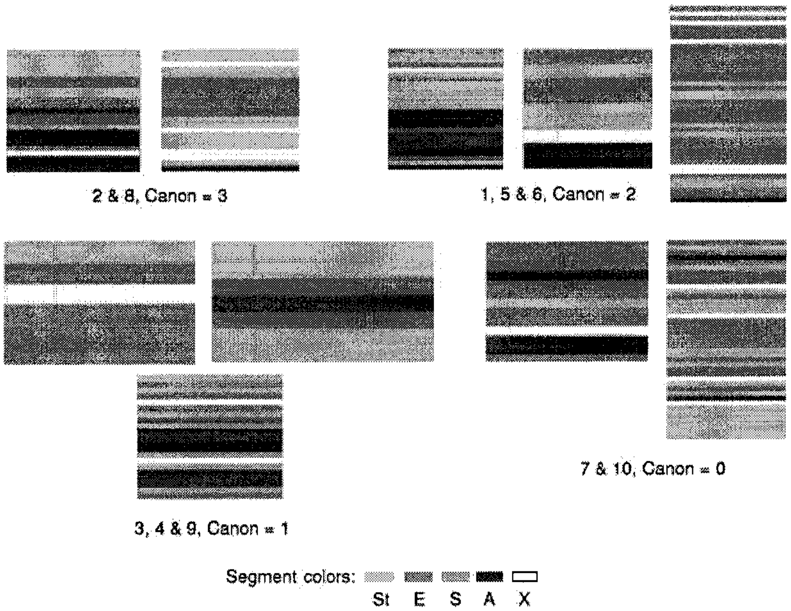
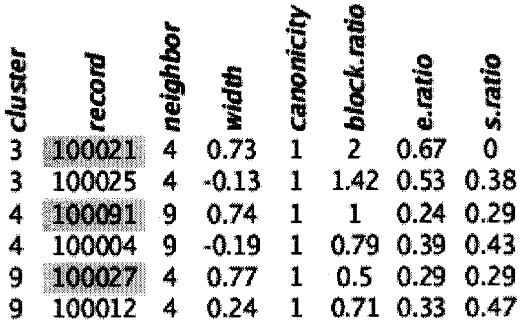


Illustration 17. Medoids Grouped by Canonicity

Illustration 17 above shows the overall distribution in terms of our two variables and in terms of the different segment types. The key at the bottom of the illustration indicates the color schema used to identify each of the segment types. It is meant as a visualization aid to understand the difference between the cluster exemplars. The text was removed for visualization purposes. The items shown are the medoids for each group, and the groups are blocked together by means of the variable *Canon*. Since only two variables were used for the clustering procedure, this arrangement yields visually clear distinctions between the members of each cluster. Namely, the ratio of the distribution of the E-type over the S-type segments. This is expressed in the E.to.S.block.ratio variable introduced previously, and it is obtained from the ratio of the average block size of type E over type S. For instance, noticing the first set in the upper left corner, which corresponds to clusters 2 and 8 with canonicity of 3, we see that the difference between the two clusters is in the overall proportion of dark grey lines (E ▓) over light gray lines (S ░). In this case, cluster 2 has a lower distribution of E segments and cluster 8 seems to have about the same. This is corroborated by their values of 1.17 and 3.5 respectively. It is a large difference considering that the maximum value obtained for the E/S ratio variable is 4 (Table 8). The overall patterns obtained in the clustering can be seen from examining Illustration 17 in the same fashion. Whenever there are three medoids per set, one of them has a value close to 1 for the E/S ratio, another one in the set has a low value, and a third one a high value (e.g., 1,5 & 6). Whenever we find only two medoids in a set, such as the case of groups number 2 & 8 just described, we find one group with a high E.to.S.block.ratio value and one with a value close to 1 (equal). This represents a clear pattern of division among the grammatical styles present in our narrative corpus (data set). It corresponds to the baselines we were set out to obtain earlier, in terms of the different temporal structure distributions that are present in the narratives. This now constitutes our "style" baselines, against which we will measure deviations in terms of other characteristics. We will use a new set of variables to study the distribution of *subjective* blocks of text in the following section, and we will do so within each cluster we obtained.

An interesting discovery from the results of this clustering process is that only ten out of a total of eighty eight narratives classified (12 were discarded due to a value of unmatched segments over 20%) present a value of 3 for the canonicity variable. This result is interesting from the point of view that one would expect the contrary to be the case, since the most typical format for the narrative style calls for the presence of an introduction and a conclusion. In other words, a story is usually told by introducing the setting, the participants, then narrating the sequence of events and outcomes that occurred, and finally concluding with some sort of moral, evaluation or opinion. It is the case for the ASRS database forms, that some of that information is already present in the rest of the form fields. Therefore it is logical that writers obviate some of that information, and move straight into the course of events. This is indeed the case. Also, the forms themselves do not contain information that would constrain the format of these narratives, or that would induce writers to follow a more conventional style of narrative writing. This result is a useful piece of information and part of the motivation behind the clustering process, i.e., to uncover precisely this type of distribution particular to our data set.

We notice from the plot in Illustration 16 above that there are two outliers: the last members of clusters 3 and 4. The following table (Table 7) shows their cluster *width* information. Highlighted record numbers correspond to medoids, the ones with a negative width value are the outliers.



cluster	record	neighbor	width	canonicity	block.ratio	e.ratio	s.ratio
3	100021	4	0.73	1	2	0.67	0
3	100025	4	-0.13	1	1.42	0.53	0.38
4	100091	9	0.74	1	1	0.24	0.29
4	100004	9	-0.19	1	0.79	0.39	0.43
9	100027	4	0.77	1	0.5	0.29	0.29
9	100012	4	0.24	1	0.71	0.33	0.47

Table 7. Outlier and Medoid Data

We can see in the table above that the assigned neighbors for the two outliers are group 4 and 9 respectively. That is, the outlier in group 3 is narrative record 100025 and is closer to group 4 than to the medoid in group 3. Record 100004 was classified in cluster 4, but the negative value width tells us that it might be better placed in cluster 9, its assigned neighbor. By inspecting the values of these two outliers for *canonicity* and *block.ratio*, we can see that they are *not* significantly closer to their neighbors than they are from their respective group medoids. All records have a canonicity value of 1 (i.e., presence of St-type, or "Setting", segments at the beginning), so the comparison applies only to the block.ratio variable. For record 100025 we have a 30% difference (with respect to medoid in group 3) versus a 40% difference (to group 4). Record 100004 shows 20% versus 60%, which is a stronger case against the reclassification of that record in cluster 9. What this means is that we cannot move these outliers to a different cluster, but instead that we need to treat them as special salient cases that do not fit with the rest of the groupings. This information could be used as a means to identify and measure saliency in the narratives. As it turns out though, what we are after in our analysis is better determined through measures of *dispersion,* as it will be described next.

Dispersion and Subjectivity Measures

In order to obtain measures of *biasing* and *subjectivity,* which was the original impetus for this work, we will use a different set of variables than those used for the clustering process. In this way, we will make our assessments, against, and in relation to, the clusters that we have obtained (i.e., orthogonally). Our new variables will be defined in terms of the distributional characteristics of one type of segment only, the A types. These are the most subjectivity-loaded and therefore most salient segments for our purposes. The variables used previously for the clustering step did include information on this A-type segment. The canonicity variable (when it has a value of 3) indicates the presence of at least one segment of this type at the end of a narrative. Here we will focus exclusively on this one type of segment. We hope in this way to have clusters that can be used as *baselines* against which to compare our new variable measures. In other words, we have a distribution for *all* four segment types across *all* records, and now, we will define a new variable measure of the distribution of *only one* particular segment type and analyze it only *within* a given group. Our newly defined variables are:

a.ratio = ratio of A-type blocks over the total number of blocks
dispersion = a measure of the average distance between all A-type segments present

The *dispersion* variable is obtained and normalized as follows

$$\mathrm{disp} = (\textstyle\sum_{i=0}^{i=n-1} \mathrm{ind}_i - \mathrm{ind}_{i-1}) \; / \; \# \text{ of A blocks}$$

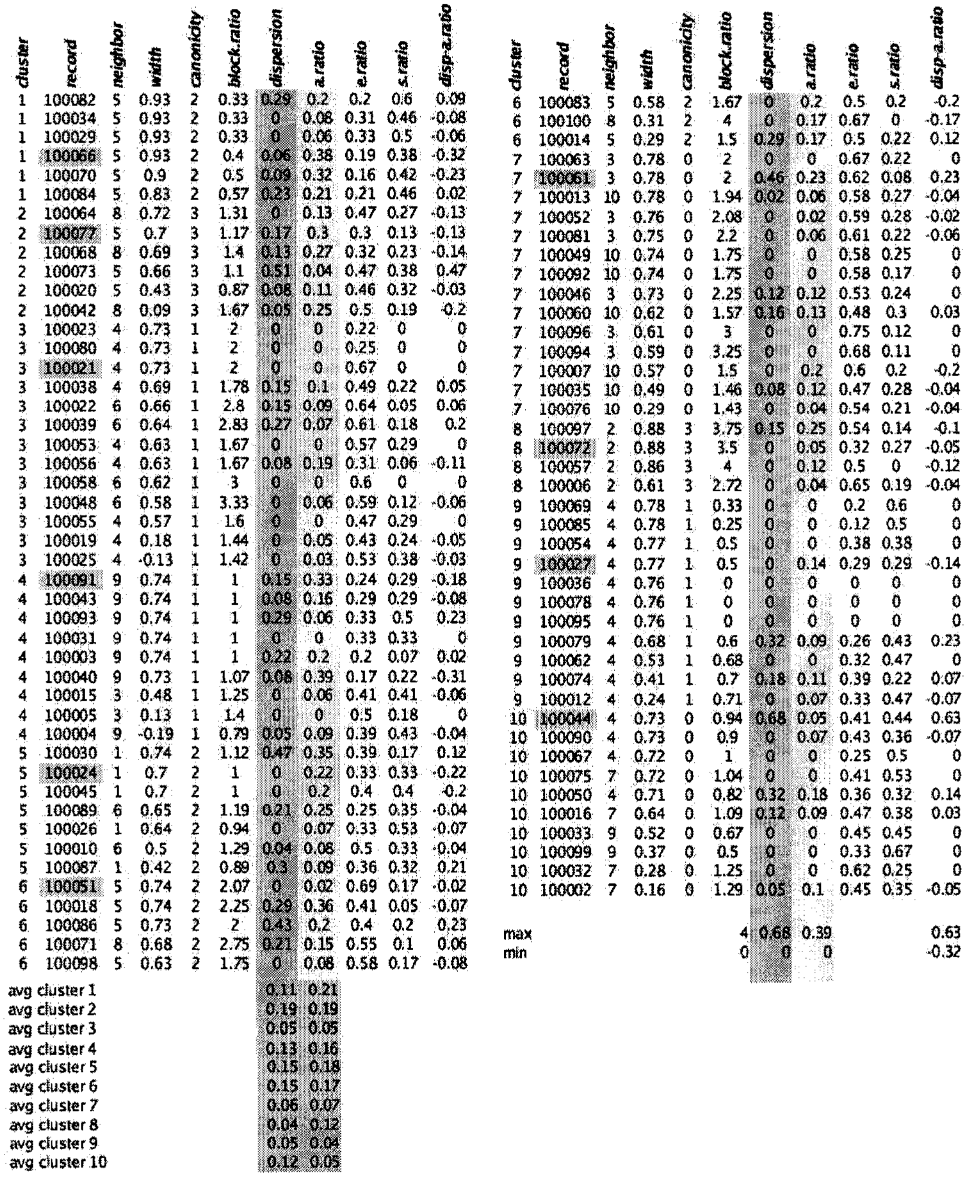
$$\mathrm{disp}_{norm} = \mathrm{disp} \; / \; (\mathrm{tot}_{S,E,St,A} - (\mathrm{tot}_{St} + \mathrm{fin}_A))$$

where *ind* represent the index of every A-type segment in a given narrative, starting from position 1 to *n,* the total number of segments in the narrative; tot_x represents the total number of segments of a give type and fin_A stands for the number of contiguous A-type segments at the end of a given narrative.

Given the definition of these new variables, Table 8 below shows the complete results for the count. The relevant columns are highlighted in two tones of gray for ease of visualization. There were a total of 100 records analyzed by the program, and a total of 88 records used for clustering. This is due to a filter that was set up in the program to eliminate narratives with a high percentage of X-type segments. That is, we eliminated from the final clustering and analysis, those narratives which for reasons of poor parsing (due to the complexity or poor grammaticality of the writing), had more than 20% of unknown-type segments. We will discuss the origins of this problem and ways to solve it in Chapter 5.

The purpose of these two new count measures is to be able to detect two things: one is an overall indication about the subjectivity level of a narrative, given by the total proportion of A-type segments in any given narrative (in comparison to the mean proportion value of the cluster it belongs); and the other one is a measure of the separation (dispersion) of A-type blocks present in a narrative. For the latter we are assuming that the more dispersed the A-type segments are, the more they indicate a deviation from the norm in terms of how narratives exhibit most of their A-type segments towards the end. In other words, the a.ratio variable will give an overall level of *subjectivity*, whereas the *dispersion* variable will indicate potential biasing. This is the core of our hypothesis for this work.

The numbers in Table 8 below reveal some interesting patterns. The clustering was performed independently of the a.ratio variable, or of any specific information related exclusively to the A-type segment. We see that certain clusters show a very low a.ratio value, as well as a low dispersion value *consistently* (groups 3,7 and 8). This comes as no surprise for clusters with a canonicity value of 0 or 1, but it is interesting when the same effect occurs for clusters with a value of 3. What this suggests is that there is an underlying "natural" grouping of the narratives in terms of our coding which has a dependency on the A-type segment distribution, even though that information was not included in the clustering process. This is significant since that information is the most relevant to our analysis model. Our style variables thus, have the potential to be predictor variables for our subjectivity and biasing measures. More interestingly, it reveals the fact that when narratives have no conclusion, in the sense of our canonicity variable, they also tend to have a lower degree of subjectivity in general. This can be seen in clusters 3,7 and 9 in Table 8, where a large portion of the entries have a 0 value for dispersion and a very low a.ratio value as well. Conceptually, this is interesting because it suggests a tendency on the part of the narrators to include personal assessment and opinions *only* at the end of a story. At least when they are significantly salient in the narrative.



cluster	record	neighbor	width	canonicity	block.ratio	dispersion	a.ratio	e.ratio	s.ratio	disp-a.ratio
1	100082	5	0.93	2	0.33	0.29	0.2	0.2	0.6	0.09
1	100034	5	0.93	2	0.33	0	0.08	0.31	0.46	-0.08
1	100029	5	0.93	2	0.33	0	0.06	0.33	0.5	-0.06
1	100066	5	0.93	2	0.4	0.06	0.38	0.19	0.38	-0.32
1	100070	5	0.9	2	0.5	0.09	0.32	0.16	0.42	-0.23
1	100084	5	0.83	2	0.57	0.23	0.21	0.21	0.46	0.02
2	100064	8	0.72	3	1.31	0	0.13	0.47	0.27	-0.13
2	100077	5	0.7	3	1.17	0.17	0.3	0.3	0.13	-0.13
2	100068	8	0.69	3	1.4	0.13	0.27	0.32	0.23	-0.14
2	100073	5	0.66	3	1.1	0.51	0.04	0.47	0.38	0.47
2	100020	5	0.43	3	0.87	0.08	0.11	0.46	0.32	-0.03
2	100042	8	0.09	3	1.67	0.05	0.25	0.5	0.19	-0.2
3	100023	4	0.73	1	2	0	0	0.22	0	0
3	100080	4	0.73	1	2	0	0	0.25	0	0
3	100021	4	0.73	1	2	0	0	0.67	0	0
3	100038	4	0.69	1	1.78	0.15	0.1	0.49	0.22	0.05
3	100022	6	0.66	1	2.8	0.15	0.09	0.64	0.05	0.06
3	100039	6	0.64	1	2.83	0.27	0.07	0.61	0.18	0.2
3	100053	4	0.63	1	1.67	0	0	0.57	0.29	0
3	100056	4	0.63	1	1.67	0.08	0.19	0.31	0.06	-0.11
3	100058	6	0.62	1	3	0	0	0.6	0	0
3	100048	6	0.58	1	3.33	0	0.06	0.59	0.12	-0.06
3	100055	4	0.57	1	1.6	0	0	0.47	0.29	0
3	100019	4	0.18	1	1.44	0	0.05	0.43	0.24	-0.05
3	100025	4	-0.13	1	1.42	0	0.03	0.53	0.38	-0.03
4	100091	9	0.74	1	1	0.15	0.33	0.24	0.29	-0.18
4	100043	9	0.74	1	1	0.08	0.16	0.29	0.29	-0.08
4	100093	9	0.74	1	1	0.29	0.06	0.33	0.5	0.23
4	100031	9	0.74	1	1	0	0	0.33	0.33	0
4	100003	9	0.74	1	1	0.22	0.2	0.2	0.07	0.02
4	100040	9	0.73	1	1.07	0.08	0.39	0.17	0.22	-0.31
4	100015	3	0.48	1	1.25	0	0.06	0.41	0.41	-0.06
4	100005	3	0.13	1	1.4	0	0	0.5	0.18	0
4	100004	9	-0.19	1	0.79	0.05	0.09	0.39	0.43	-0.04
5	100030	1	0.74	2	1.12	0.47	0.35	0.39	0.17	0.12
5	100024	1	0.7	2	1	0	0.22	0.33	0.33	-0.22
5	100045	1	0.7	2	1	0	0.2	0.4	0.4	-0.2
5	100089	6	0.65	2	1.19	0.21	0.25	0.25	0.35	-0.04
5	100026	1	0.64	2	0.94	0	0.07	0.33	0.53	-0.07
5	100010	6	0.5	2	1.29	0.04	0.08	0.5	0.33	-0.04
5	100087	1	0.42	2	0.89	0.3	0.09	0.36	0.32	0.21
6	100051	5	0.74	2	2.07	0	0.02	0.69	0.17	-0.02
6	100018	5	0.74	2	2.25	0.29	0.36	0.41	0.05	-0.07
6	100086	5	0.73	2	2	0.43	0.2	0.4	0.2	0.23
6	100071	8	0.68	2	2.75	0.21	0.15	0.55	0.1	0.06
6	100098	5	0.63	2	1.75	0	0.08	0.58	0.17	-0.08
6	100083	5	0.58	2	1.67	0	0.2	0.5	0.2	-0.2
6	100100	8	0.31	2	4	0	0.17	0.67	0	-0.17
6	100014	5	0.29	2	1.5	0.29	0.17	0.5	0.22	0.12
7	100063	3	0.78	0	2	0	0	0.67	0.22	0
7	100061	3	0.78	0	2	0.46	0.23	0.62	0.08	0.23
7	100013	10	0.78	0	1.94	0.02	0.06	0.58	0.27	-0.04
7	100052	3	0.76	0	2.08	0	0.02	0.59	0.28	-0.02
7	100081	3	0.75	0	2.2	0	0.06	0.61	0.22	-0.06
7	100049	10	0.74	0	1.75	0	0	0.58	0.25	0
7	100092	10	0.74	0	1.75	0	0	0.58	0.17	0
7	100046	3	0.73	0	2.25	0.12	0.12	0.53	0.24	0
7	100060	10	0.62	0	1.57	0.16	0.13	0.48	0.3	0.03
7	100096	3	0.61	0	3	0	0	0.75	0.12	0
7	100094	3	0.59	0	3.25	0	0	0.68	0.11	0
7	100007	10	0.57	0	1.5	0	0.2	0.6	0.2	-0.2
7	100035	10	0.49	0	1.46	0.08	0.12	0.47	0.28	-0.04
7	100076	10	0.29	0	1.43	0	0.04	0.54	0.21	-0.04
8	100097	2	0.88	3	3.75	0.15	0.25	0.54	0.14	-0.1
8	100072	2	0.88	3	3.5	0	0.05	0.32	0.27	-0.05
8	100057	2	0.86	3	4	0	0.12	0.5	0	-0.12
8	100006	2	0.61	3	2.72	0	0.04	0.65	0.19	-0.04
9	100069	4	0.78	1	0.33	0	0	0.2	0.6	0
9	100085	4	0.78	1	0.25	0	0	0.12	0.5	0
9	100054	4	0.77	1	0.5	0	0	0.38	0.38	0
9	100027	4	0.77	1	0.5	0	0.14	0.29	0.29	-0.14
9	100036	4	0.76	1	0	0	0	0	0	0
9	100078	4	0.76	1	0	0	0	0	0	0
9	100095	4	0.76	1	0	0	0	0	0	0
9	100079	4	0.68	1	0.6	0.32	0.09	0.26	0.43	0.23
9	100062	4	0.53	1	0.68	0	0	0.32	0.47	0
9	100074	4	0.41	1	0.7	0.18	0.11	0.39	0.22	0.07
9	100012	4	0.24	1	0.71	0	0.07	0.33	0.47	-0.07
10	100044	4	0.73	0	0.94	0.68	0.05	0.41	0.44	0.63
10	100090	4	0.73	0	0.9	0	0.07	0.43	0.36	-0.07
10	100067	4	0.72	0	1	0	0	0.25	0.5	0
10	100075	7	0.72	0	1.04	0	0	0.41	0.53	0
10	100050	4	0.71	0	0.82	0.32	0.18	0.36	0.32	0.14
10	100016	7	0.64	0	1.09	0.12	0.09	0.47	0.38	0.03
10	100033	9	0.52	0	0.67	0	0	0.45	0.45	0
10	100099	9	0.37	0	0.5	0	0	0.33	0.67	0
10	100032	7	0.28	0	1.25	0	0	0.62	0.25	0
10	100002	7	0.16	0	1.29	0.05	0.1	0.45	0.35	-0.05
avg cluster 1						0.11	0.21			
avg cluster 2						0.19	0.19			
avg cluster 3						0.05	0.05			
avg cluster 4						0.13	0.16			
avg cluster 5						0.15	0.18			
avg cluster 6						0.15	0.17			
avg cluster 7						0.06	0.07			
avg cluster 8						0.04	0.12			
avg cluster 9						0.05	0.04			
avg cluster 10						0.12	0.05			
max					4	0.68	0.39			0.63
min					0	0	0			-0.32

Table 8. Clustering Results

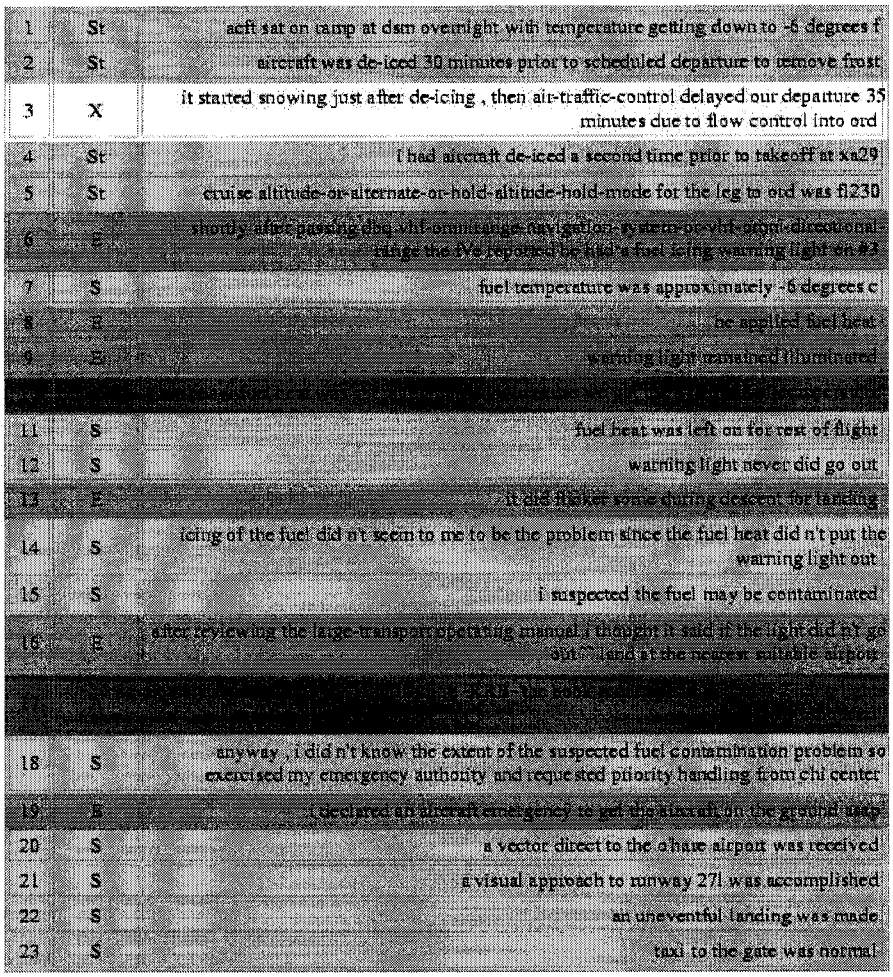
The next interesting pattern to notice in Table 8 is the difference between the value of the dispersion and the a.ratio variables for any given narrative (last column in Table 8). For instance, record 100079 in cluster 9 has a value of 0.32 for dispersion and a value of 0.09 for a.ratio. This is a salient difference considering that cluster 9 has an average dispersion of 0.05 (see bottom left of Table 8). This means that record 100079 is salient for two reasons: one is for being the highest entry in its cluster with a dispersion value significantly above the mean for the group, and also because it has a large difference with its a.ratio value, more precisely 0.23, well above the average of the measure in the group for either variable, 0.05 and 0.04 respectively. We will be analyzing this trend more carefully in the next section when we compare individual records.

In sum, we have assessed through our clustering procedure and our result values, that it is possible to identify narratives on the basis of their subjective measures, doing so with the variables we have defined and the data we have coded with our method. The next logical step is to carefully examine a representative set of narratives selected through these measures and comment on the potential of our method as a tool for error analysis. This process is equivalent to the validation for our method at this point. This is only a pilot test for our program; an expert evaluation of the results will be performed in the future. This aspect will be expanded on Chapter 5 as part of our commentary and suggestions for future research directions.

Comparison of Selected Narratives

In order to draw better conclusions about the validity of our approach as an automated detection and diagnosis tool, we need to examine closely some representative samples obtained through the variables defined in the previous section. The purpose of this section is also to set up the ground for a comparative discussion of the caveats and benefits of our approach, in terms of ways to improve it and ways to incorporate it with the existing approaches and technologies. This issue will be addressed in Chapter 5.

We start with the two most salient narratives in the complete set. Record 100079 exhibits the highest proportional dispersion measure in comparison with the average of its cluster (0.32 against 0.04 avg.). Record 100040 shows the highest value of the a.ratio variable (0.39). Illustration 18 below shows the contents of record 100079, using the coloring schema introduced previously for visualization purposes.

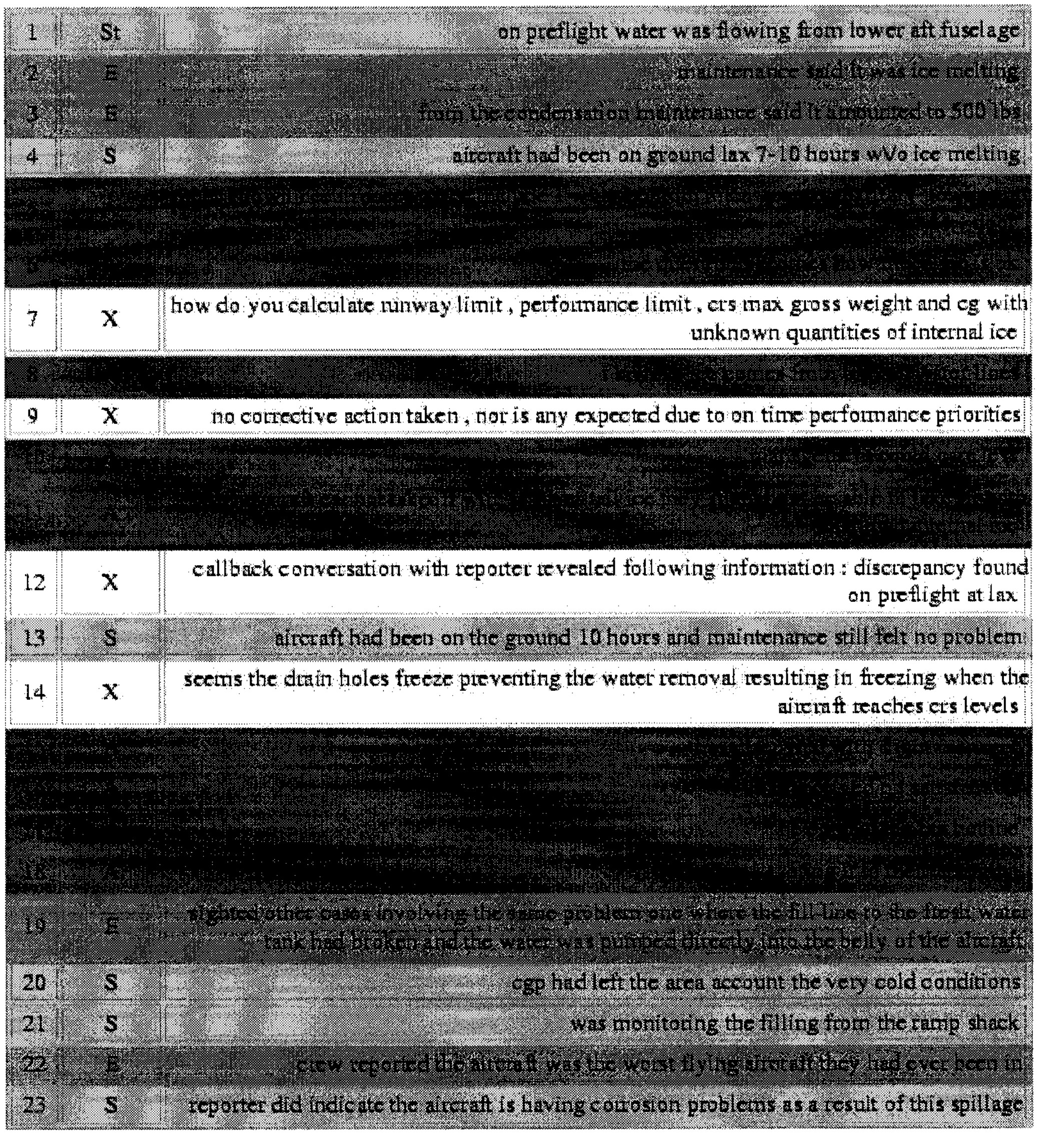


1	St	acft sat on ramp at dsm overnight with temperature getting down to -6 degrees f
2	St	aircraft was de-iced 30 minutes prior to scheduled departure to remove frost
3	X	it started snowing just after de-icing , then air-traffic-control delayed our departure 35 minutes due to flow control into ord
4	St	i had aircraft de-iced a second time prior to takeoff at xa29
5	St	cruise altitude-or-alternate-or-hold-altitude-hold-mode for the leg to ord was fl230
6	E	shortly after passing dbq vhf-omni-range-navigation-system-or-vhf-omni-directional-range the fe reported he had a fuel icing warning light on #3
7	S	fuel temperature was approximately -6 degrees c
8	E	he applied fuel heat
9	E	warning light remained illuminated
[illegible]	[illegible]	[illegible]
11	S	fuel heat was left on for rest of flight
12	S	warning light never did go out
13	E	it did flicker some during descent for landing
14	S	icing of the fuel did n't seem to me to be the problem since the fuel heat did n't put the warning light out
15	S	i suspected the fuel may be contaminated
16	E	after reviewing the large-transport-operating manual, i thought it said if the light did n't go out ~~ land at the nearest suitable airport
17	A	[illegible]
18	S	anyway , i did n't know the extent of the suspected fuel contamination problem so exercised my emergency authority and requested priority handling from chi center
19	E	i declared an aircraft emergency to get the aircraft on the ground asap
20	S	a vector direct to the o'hare airport was received
21	S	a visual approach to runway 271 was accomplished
22	S	an uneventful landing was made
23	S	taxi to the gate was normal

Illustration 18. Record 100079 Segmented

This narrative describes an incident where an airplane experiences fuel contamination due to icing. It starts off with an introductory setting, describing that the aircraft was stationed overnight at temperatures below 0°F. We notice this because of the initial block of St-type segments, which are captured in the record's value of 1 for the canonicity variable. The narrative contains no conclusion portion (no ending A segments), and exhibits a very low overall value for the a.ratio variable (0.09). This is evident by looking at the overall number of dark segments (A-type segments) in Illustration 18. The narrative progresses to the point where the first signs of a problem appear, in segment 6, marked as the first event of type E. The series of following events tell of an emergency situation declared, which ended in a safe landing. What is interesting here is to note the two A-type segments which triggered the high dispersion value in the first place: segments 10 and 17. We know from our psychology and linguistic model of the narrative mode (Chapter 3), that this type of intrusion in an otherwise linear narrative form, indicate inferential processes being used (as opposed to memory retrieval). According to our hypothesis, this may also indicate *biasing* present on the part of the operator at the time of the incident. Segment 10, the first salient segment, states *"we know fuel heat was operating normally because we got the rise in #3 oil temperature".* There is a shift here in the narrative timeline to Speech Time (Sp), indicating that the narrator is making use of his knowledge of the aircraft to draw an inference about the situation at the time. We could assume here that the same inferential processes were active at the equivalent point in time during the course of the incident. It also represents a bracket in the narrative where the author is revealing information about the current SA at the time. In this particular case it may not reveal anything new, but the author could have as well said *" i knew that fuel heat was on because...".* But the fact is that he did not, and because of his choice of the grammatical construction he used, he raised a flag in our method. This type of shift creates saliency in the narrative structure, which could be used in interesting ways by automated pattern induction technologies, such as machine learning algorithms, because it creates locality in the narrative text. That is, it brings focus to the segments around segment 10. As it turns out the next salient segment, number 17, correlates. Conceptually, the saliency of segments 10 & 17 arises from the fact that they are not placed at the end of the narrative in a single block, but rather, interleaved throughout the narrative together with the normal description of events. Therefore, these segments stand out in an otherwise linear sequence of events. We could think of these as *intermissions*, where the author brings forth his knowledge about the operational environment at the time (i.e., SA), to justify or explain his rationale for action. This is interesting from our point of view, and from our revised human performance model of SA, since it raises a flag that indicates potential for the presence of an incorrect mental model at the time of operation. In this case, the model was a safety one, which prevented an accident from occurring, but it could have been a different case. Ultimately this is the kind of information we want to evaluate through an expert review of our program's output. This is part of our future plan for research to be discussed in Chapter 5.

The next narrative, record 100040 is shown below in Illustration 19.



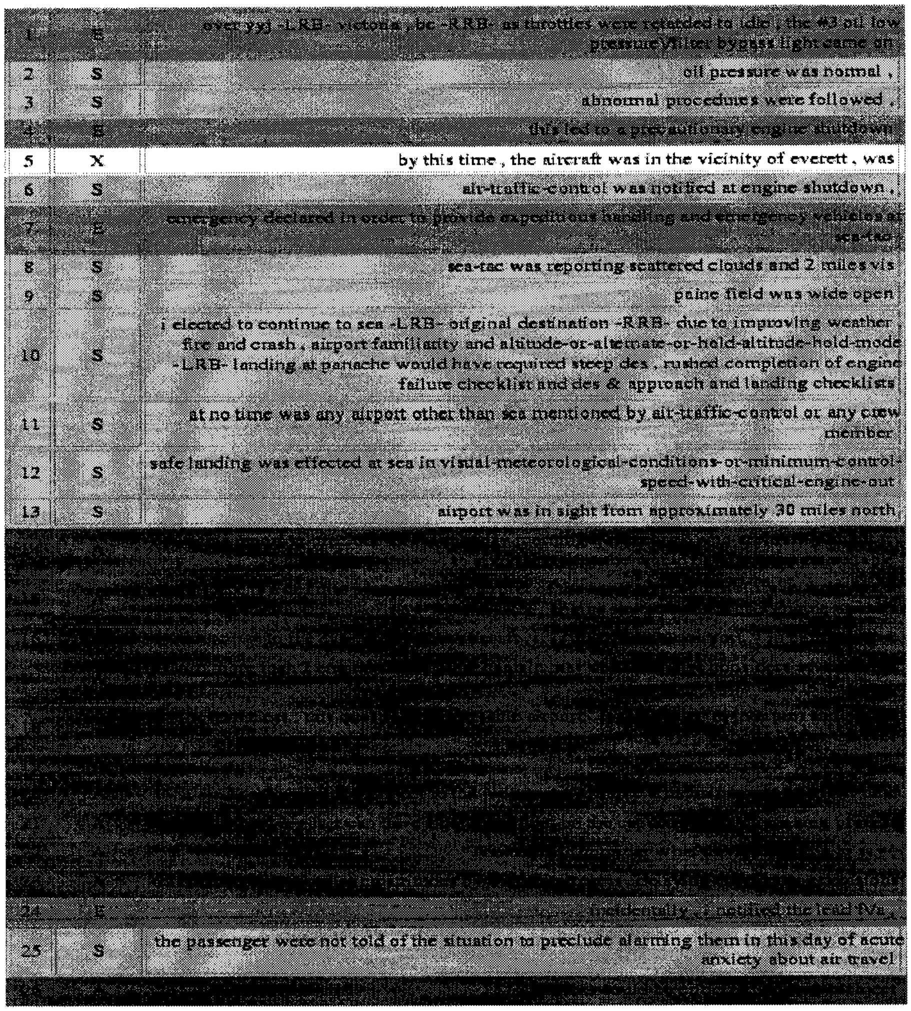
1	St	on preflight water was flowing from lower aft fuselage
2	E	maintenance said it was ice melting
3	E	from the condensation maintenance said it amounted to 500 lbs
4	S	aircraft had been on ground lax 7-10 hours w/o ice melting
5	[illegible]	[illegible]
6	[illegible]	[illegible]
7	X	how do you calculate runway limit , performance limit , crs max gross weight and cg with unknown quantities of internal ice
8	[illegible]	[illegible]
9	X	no corrective action taken , nor is any expected due to on time performance priorities
10	[illegible]	[illegible]
11	[illegible]	[illegible]
12	X	callback conversation with reporter revealed following information : discrepancy found on preflight at lax
13	S	aircraft had been on the ground 10 hours and maintenance still felt no problem
14	X	seems the drain holes freeze preventing the water removal resulting in freezing when the aircraft reaches crs levels
15	[illegible]	[illegible]
16	[illegible]	[illegible]
17	[illegible]	[illegible]
18	[illegible]	[illegible]
19	E	sighted other cases involving the same problem one where the fill line to the fresh water tank had broken and the water was pumped directly into the belly of the aircraft
20	S	cgp had left the area account the very cold conditions
21	S	was monitoring the filling from the ramp shack
22	E	crew reported the aircraft was the worst flying aircraft they had ever been in
23	S	reporter did indicate the aircraft is having corrosion problems as a result of this spillage

Illustration 19. Record 100040 Segmented

This narrative was selected because of its high a.ratio value (0.39). In contrast it exhibits a very low dispersion value (0.09). What we see in this narrative is an overall high proportion of darker segments (A-type). The white segments (X-type) indicate poorly parsed segments (see Chapter 5 for a commentary on this issue). If those segments were to be correctly parsed, the a.ratio value would be even higher (0.43) and approaching 50% of the narrative content. What this means is that this narrative is overall highly subjective, and that our program was able to single it out based on the clustering process and variables measured. The story pertains a water leakage incident. Water was coming out of an aircraft prior to departure, and it was attributed by the maintenance crew to the presence of internal ice melting. The narrative deviates from a normal sequence of events by introducing information about other incidents known to the author (segment 5). This is followed by an assessment (8) and strong opinionated statements (6,7,10 and 11). The rest of the story pertains to a callback conversation, and it is also marked as A-type. This is a pattern seen in many of these narratives, where a story is broken down into two sections, including a callback report. Our program was able to diagnose the presence of such narrative style that indicates a high degree of subjectivity compared to others in the cluster. This can be extremely useful for classifying narratives prior to using other automated methods. In particular, when using keyword and phrase-based approaches (which assign the same weight to any portion of the narrative indiscriminately), this type of pre-classification could help to identify and avoid matching words that are not directly related to the narration of the incident (e.g., areas of high subjectivity and extraneous information). An example here is segment 5 where extrinsic information is introduced; a keyword approach could fall in the trap of matching on words such as "main gear", "auxiliary-power-unit", causing a misclassification of the incident's context.

Another pattern worthwhile studying is the difference between the dispersion and the a.ratio variables. This difference is shown in the last column in Table 8 above. We will select the highest positive and the highest negative value as exemplars for commentary. More sophisticated analysis will have to be performed in the future to determine whether this or any other patterns of this sort are statistically significant (Chapter 5 will comment on this issue). The motivation here is to show how the results have the potential to act as detector signals with meaningful distinctions according to the SA model and the error taxonomy under study.

The selected narratives are record 100066 (disp-a.ratio = -0.32) and 100073 (disp-a.ratio = 0.47). The segmented output for narrative 100066 is shown below in Illustration 20. The most striking characteristic in the segmentation of this narrative is that almost half of it is of type A. Indeed as it turns out, this is a highly subjectively charged story. The writer 's intention is to make clear that his responsibility as a pilot is to make the most appropriate decision given the information available to him during operations, and not to be subject to an after-the-fact analysis by FAA staff. The story concerns a precautionary engine shutdown effected after an oil pressure indicator was signal

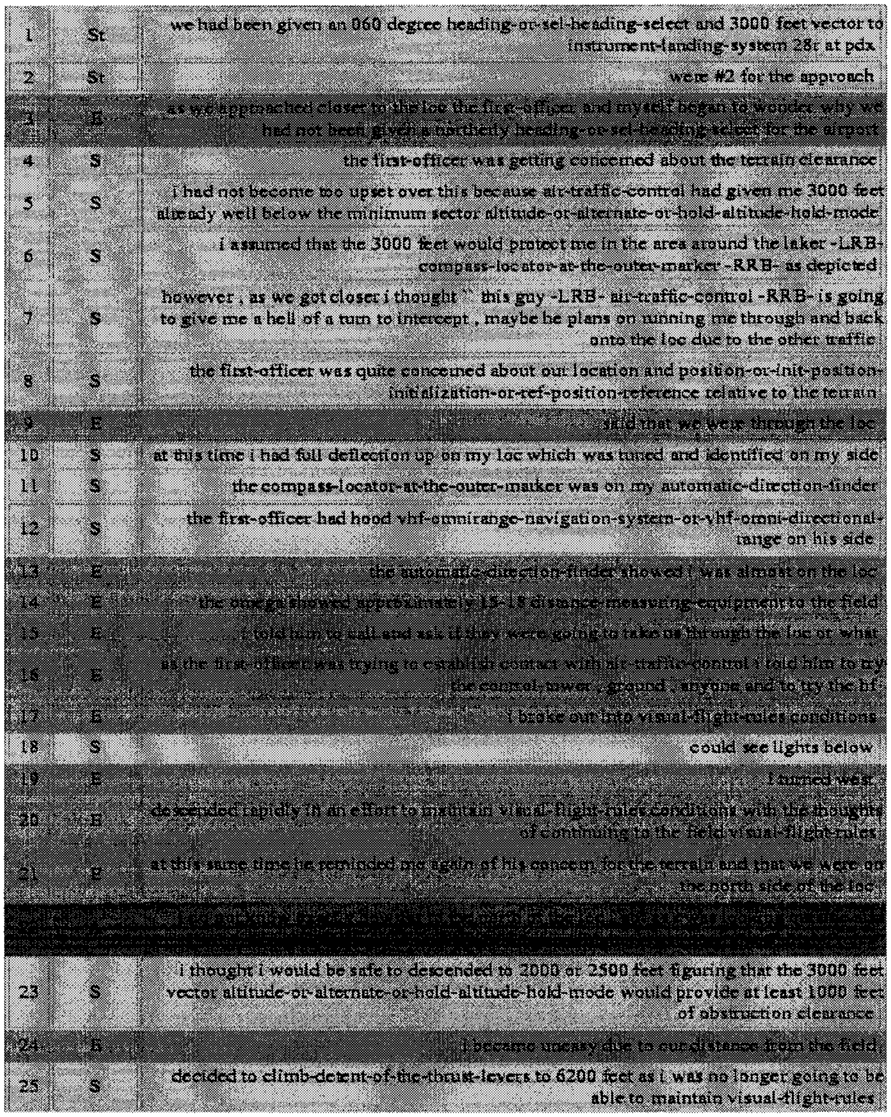


1	E	over yyj -LRB- victoria , bc -RRB- as throttles were retarded to idle , the #3 oil low pressure/filter bypass light came on
2	S	oil pressure was normal ,
3	S	abnormal procedures were followed ,
4	E	this led to a precautionary engine shutdown
5	X	by this time , the aircraft was in the vicinity of everett , was
6	S	air-traffic-control was notified at engine shutdown ,
7	E	emergency declared in order to provide expeditious handling and emergency vehicles at sea-tac
8	S	sea-tac was reporting scattered clouds and 2 miles vis
9	S	paine field was wide open
10	S	i elected to continue to sea -LRB- original destination -RRB- due to improving weather , fire and crash , airport familiarity and altitude-or-alternate-or-hold-altitude-hold-mode -LRB- landing at panache would have required steep des , rushed completion of engine failure checklist and des & approach and landing checklists
11	S	at no time was any airport other than sea mentioned by air-traffic-control or any crew member
12	S	safe landing was effected at sea in visual-meteorological-conditions-or-minimum-control-speed-with-critical-engine-out
13	S	airport was in sight from approximately 30 miles north
[illegible]	[illegible]	[illegible]
24	E	incidentally , i notified the lead fVa ,
25	S	the passenger were not told of the situation to preclude alarming them in this day of acute anxiety about air travel
[illegible]	[illegible]	[illegible]

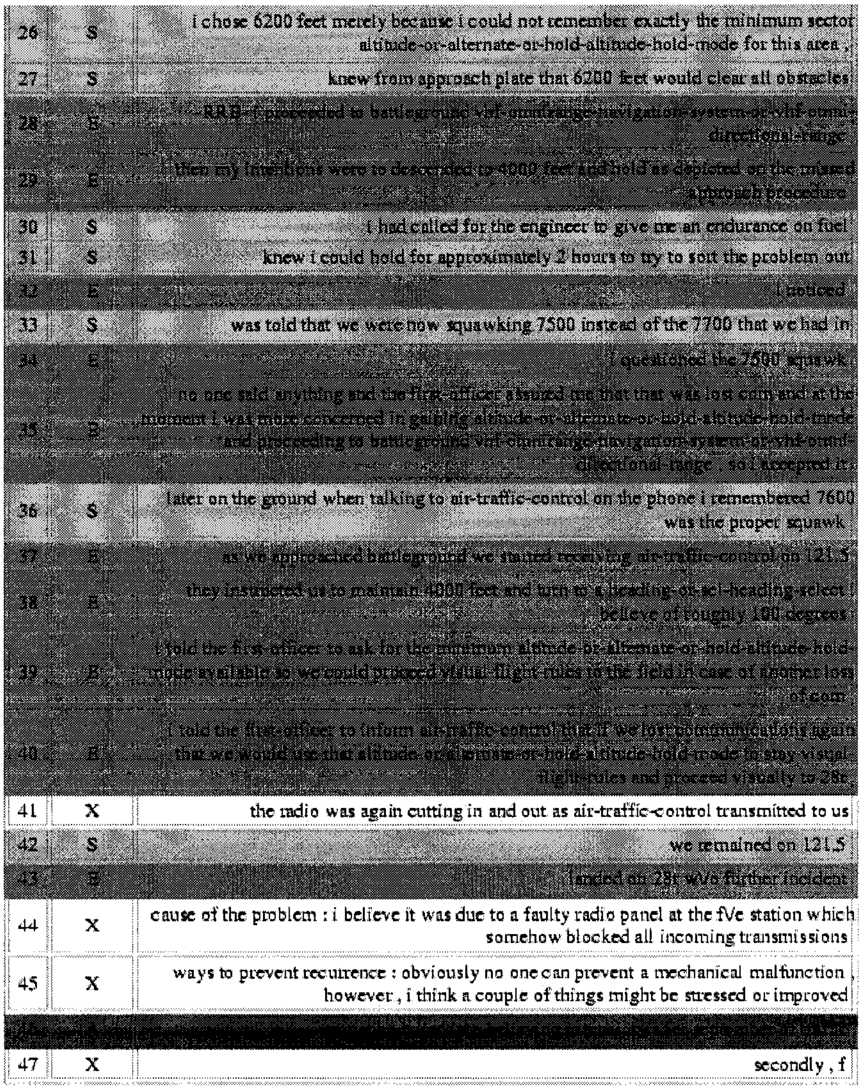
Illustration 20. Record 100066 Segmented

What might seem like the conclusion or assessment portion of the narrative takes up half of the total length of the record, starting at segment 14 and continuing to the end. The quantitative signal that triggered the selection of this narrative is the high value of the difference between the dispersion and the a.ratio variable (-0.32). Qualitatively, this represents a high proportion of A-type segments over the rest of the narrative and a low dispersion value, which means that the A blocks are highly contiguous. From the standpoint of our model, this means that all opinions or inferences are introduced at a *single* point in the timeline of the story. Furthermore, since our canonicity variable has a value of 2 for this record, we know that this contiguous type A block is located at the end. Considering these two aspects together, we have a way to detect a narrative which is highly opinionated, and most likely related to issues of rank, responsibility or regulations within companies and the FAA. This is because inferential processes are most usually manifested as *interleaved* segments distributed throughout the narrative timeline, as we saw for the case of record 100079 above. The distribution of our variables shows this pattern emerging.

Record 100073 is shown in Illustration 21 below. This narrative was selected due to its high dispersion-a.ratio difference value of 0.47. The difference with the previous example is that the dispersion measure is considerably higher than the a.ratio. Conceptually this indicates that there is a low number of A-type segments in proportion to the total number of segments, and that their separation is high (also in proportion to the narrative length).



1	St	we had been given an 060 degree heading-or-sel-heading-select and 3000 feet vector to instrument-landing-system 28r at pdx
2	St	were #2 for the approach
3	E	as we approached closer to the loc the first-officer and myself began to wonder why we had not been given a northerly heading-or-sel-heading-select for the airport
4	S	the first-officer was getting concerned about the terrain clearance
5	S	i had not become too upset over this because air-traffic-control had given me 3000 feet already well below the minimum sector altitude-or-alternate-or-hold-altitude-hold-mode
6	S	i assumed that the 3000 feet would protect me in the area around the laker -LRB- compass-locator-at-the-outer-marker -RRB- as depicted
7	S	however , as we got closer i thought `` this guy -LRB- air-traffic-control -RRB- is going to give me a hell of a turn to intercept , maybe he plans on running me through and back onto the loc due to the other traffic
8	S	the first-officer was quite concerned about our location and position-or-init-position-initialization-or-ref-position-reference relative to the terrain
9	E	said that we were through the loc
10	S	at this time i had full deflection up on my loc which was tuned and identified on my side
11	S	the compass-locator-at-the-outer-marker was on my automatic-direction-finder
12	S	the first-officer had hood vhf-omnirange-navigation-system-or-vhf-omni-directional-range on his side
13	E	the automatic-direction-finder showed i was almost on the loc
14	E	the omega showed approximately 15-18 distance-measuring-equipment to the field
15	E	i told him to call and ask if they were going to take us through the loc or what
16	E	as the first-officer was trying to establish contact with air-traffic-control i told him to try the control-tower , ground , anyone and to try the hf
17	E	i broke out into visual-flight-rules conditions
18	S	could see lights below
19	E	i turned west
20	E	descended rapidly in an effort to maintain visual-flight-rules conditions with the thoughts of continuing to the field visual-flight-rules
21	E	at this same time he reminded me again of his concern for the terrain and that we were on the north side of the loc
[illegible]	[illegible]	[illegible]
23	S	i thought i would be safe to descended to 2000 or 2500 feet figuring that the 3000 feet vector altitude-or-alternate-or-hold-altitude-hold-mode would provide at least 1000 feet of obstruction clearance
24	E	i became uneasy due to our distance from the field
25	S	decided to climb-detent-of-the-thrust-levers to 6200 feet as i was no longer going to be able to maintain visual-flight-rules



26	S	i chose 6200 feet merely because i could not remember exactly the minimum sector altitude-or-alternate-or-hold-altitude-hold-mode for this area ,
27	S	knew from approach plate that 6200 feet would clear all obstacles
28	E	-RRB- i proceeded to battleground vhf-omnirange-navigation-system-or-vhf-omni-directional-range
29	E	then my intentions were to descended to 4000 feet and hold as depicted on the missed approach procedure
30	S	i had called for the engineer to give me an endurance on fuel
31	S	knew i could hold for approximately 2 hours to try to sort the problem out
32	E	i noticed
33	S	was told that we were now squawking 7500 instead of the 7700 that we had in
34	E	i questioned the 7500 squawk
35	E	no one said anything and the first-officer assured me that that was lost com and at the moment i was more concerned in gaining altitude-or-alternate-or-hold-altitude-hold-mode and proceeding to battleground vhf-omnirange-navigation-system-or-vhf-omni-directional-range , so i accepted it
36	S	later on the ground when talking to air-traffic-control on the phone i remembered 7600 was the proper squawk
37	E	as we approached battleground we started receiving air-traffic-control on 121.5
38	E	they instructed us to maintain 4000 feet and turn to a heading-or-sel-heading-select i believe of roughly 100 degrees
39	E	i told the first-officer to ask for the minimum altitude-or-alternate-or-hold-altitude-hold-mode available so we could proceed visual-flight-rules to the field in case of another loss of com
40	E	i told the first-officer to inform air-traffic-control that if we lost communications again that we would use that altitude-or-alternate-or-hold-altitude-hold-mode to stay visual-flight-rules and proceed visually to 28r
41	X	the radio was again cutting in and out as air-traffic-control transmitted to us
42	S	we remained on 121.5
43	E	landed on 28r w/o further incident
44	X	cause of the problem : i believe it was due to a faulty radio panel at the fVe station which somehow blocked all incoming transmissions
45	X	ways to prevent recurrence : obviously no one can prevent a mechanical malfunction , however , i think a couple of things might be stressed or improved
46	A	[illegible]
47	X	secondly , f

Illustration 21. Record 100073 Segmented

We can infer that because the A segment separation is high, and because the record has a value of 3 for canonicity, that at least one of the A segments is at the end of the narrative, and that at least one is toward the middle or the beginning of the story. Again and conceptually, this is an indication of a deviation in the narrative timeline occurring as an isolated signal in an otherwise linear development of events. This deviation indicates a possible biasing case, that is expressed through inferential descriptions *interleaved* with the natural development of the story. Segment 22 in this case portrays an action performed under poor situation awareness, which becomes a causal mechanism in the development of the rest of the incident: " *i do not know exactly how far to the north of the loc i got as i was looking outside and descending trying to stay visual-flight-rules".* This segment creates an area of *local* saliency, where information has potential to be more relevant in explaining causal mechanism at play in the incident. This story concerns a breakdown in the communication channel between control tower and the crew of an airplane, possibly due to a faulty radio panel. Evidently it is an external factor that causes the loss of SA. What makes this story interesting from our point of view is segment 22 which suggests the idea that perhaps the SA-loss was also driven by a *decision* to stay in visual flight rules mode. This is debatable, and ultimately contingent on the behavioral and error model adopted, but it raises a flag in the development of the narrative from the point of view of our diagnostic tool. The saliency detected here is produced because the rest of the narrative temporal structure is linear and does not contain any other shift in its timeline, except for the last conclusion statement. It can be detected automatically by our program through the use of our coding method.

CHAPTER 5

CONCLUSIONS AND FUTURE RESEARCH

This final chapter will cover the following ground. First, we will comment on the potential application of our methodology and of the results toward revising the existing error taxonomies for SA. Following, we will discuss new research directions made possible by the current work, and also future research that is needed for the refinement of our program. Finally, we will comment on the robustness of the program by drawing a preliminary accuracy assessment, and discuss current NLP techniques and tools that can help improve the correct hit ratio for the output.

SA Error Taxonomies

The main goal behind this work was to device a new automated methodology to extract information from text narratives. The secondary motivation was to suggest ways to make use of this methodology for the revision and development of improved error taxonomies for SA. Therefore, this section is concerned with making such comments and suggestions for the Human Factors community. As we described at the end of Chapter 2 section entitled "Situation Awareness", the hypothetical model we found most suited in explaining *misperception* and *misinterpretation* provided a mechanism for mental models to influence the perception components of the perception-action loop. We argued that this type of "ecological" model is better suited to avoid the large *pockets* of misperception and misinterpretation error classification present in the current error taxonomies. The fact that there is such large proportion of errors that fall within these two categories, suggests that the taxonomy needs revision.

What we will propose here is inspired from general error taxonomies first devised by Reason (1990) and Rasmussen (1982), and more recently (specific to aviation) by Sarter & Alexander (Sarter & Alexander, 2000). The idea it is to take a different perspective and approach to creating SA error taxonomies than the one proposed by Endsley. We will use a fundamental distinction between what is known as "genotypes" versus "phenotypes", deriving the terminology from the natural sciences, which has been used widely in the Human Factors field (Reason, 1990; Sarter & Alexander, 2000). The main concept motivating this distinction is the idea that underlying cognitive processes that cause errors can have manifestations anywhere on the spectrum of human performance, and therefore this distinction is needed to differentiate causal processes from signals. This is particularly relevant in our case where signals are easily detectable by the operators themselves as described in the narratives, but the causal mechanisms for the errors are not. That is one of the challenges behind the analytical models employed in the AvSP project.

What we will propose here takes on this idea about the separation of error processes versus manifestations to suggest a slightly different (and in our opinion more suitable) approach to SA error taxonomies. The model, shown in Table 9 below, introduces a new genotype labeled "Representational" to account for the idea of mental models *modulating* perception (recall Chapter 2, section "Situation Awareness").

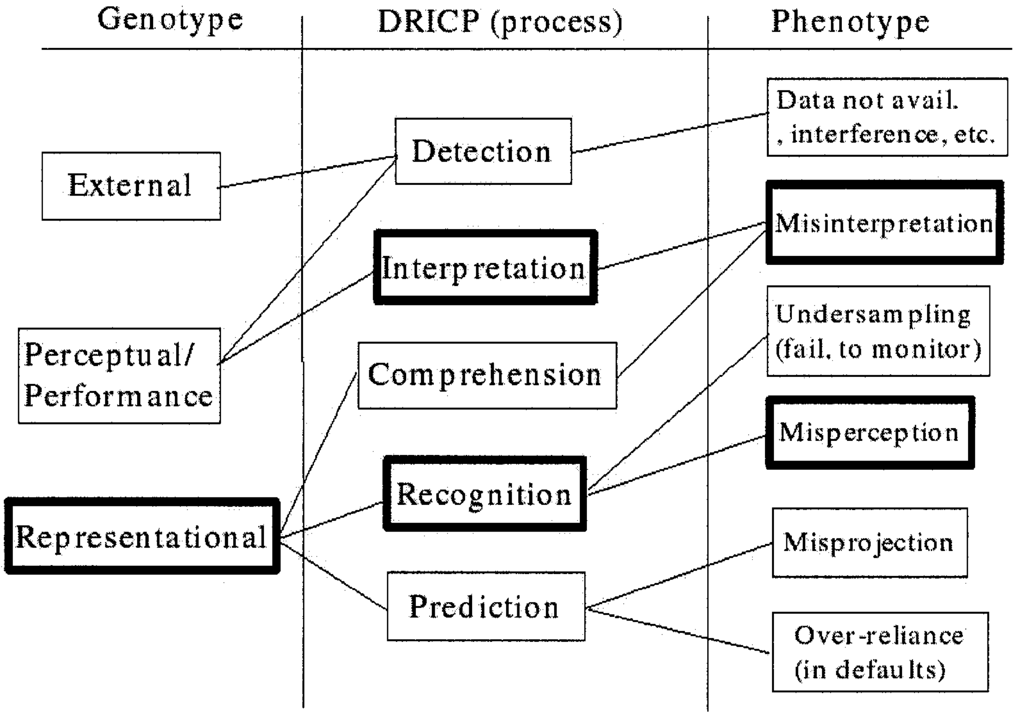


Table 9. Proposed Approach to SA Error Taxonomy

As we can see from Table 9, the representational genotype is linked to errors of misinterpretation, misperception and misprojection. At the same time, it can manifest itself *through the processes* of comprehension, recognition and prediction. The middle section was added to the model to facilitate the mapping with the DRICP model. The model emphasizes the separation between *recognition* and *interpretation*. This is a key idea because it addresses the question about the difference between misinterpretation and misperception, and its relation to SA. Note that a representational process can drive recognition of patterns in the environment, in accord with the idea of recognition-primed decision-making (Klein, 1989). This stands in contrast to detection of patterns, which is primarily a perception-driven process. In other words, there has to be some representational elements at work during human performance for signals in the environment to be *recognized*, otherwise, we are really concerned about detection (perception). This implies that an error arising from "poor recognition" is driven by an underlying model that has been poorly updated, or is simply incorrect for the given situation (i.e., at some point in the development of events, a missed signal triggered the selection of the wrong mental representation). The process of interpretation, as it has been modeled by Kintch and Ericsson (1995) and Doane (Doane & Sohn, 2002), represents the integration of perceived events into a network of " *retrieval cues*", which are then used to recall information from Long-Term Working Memory (LTWM). LTWM is an extended structure to represent Short-Term Memory in skilled performance. We see from Table 9 that the error (phenotype) labeled *misinterpretation* has a link to both comprehension and interpretation in the DRICP cycle. What this means is that this type of SA error really could be triggered by either one of two real-time effects: 1) some performance-envelope (or external) factor creates incorrect retrieval cues in LTWM in the first place, which are then used to retrieve incorrect information used for decision-making or 2) the retrieval cues were correctly formed during the perceptual portion of the model, but an incorrect representation is present in the operator for reasons other than perceptual (e.g., habit, prior knowledge or environment, social/personal expectations, etc). The misinterpretation occurs during the comprehension portion of the cycle for the last case. Reconsider the artificial example presented in page 4,

> "I WAS DRIVING TO L.A. LAST SUMMER IN MY OLD 1972 BMW. PRIOR TO DEPARTURE I HAD THE ENGINE CHECKED AND THE OIL CHANGED. ON MY WAY THERE THE OIL INDICATOR WENT OFF. I **HAD JUST CHANGED THE OIL, AND PLUS THIS KIND OF THING ALWAYS HAPPENS TO MY OLD CAR, SO I CONTINUED MY TRIP**. ABOUT 40 MI FURTHER MY ENGINE SEIZED. AS I FOUND OUT LATER, THE SERVICEMEN AT JIFFY LUBE HAD LEFT THE OIL CAP OFF! MORAL: I'LL NEVER GO BACK TO THAT SHOP AGAIN."

At the most evident level the cause of the incident could be attributed to the servicemen who left the oil cap off after servicing the vehicle. In the current error taxonomies employed for risk analysis, this type of incident would classify as having an external causal factor. Taking the perspective from our suggested model and utilizing our new method for detection, we can argue that the source of this error need be attributed to the operator's failure to follow procedures (in this case established by common sense). The causal factor that triggered this error should be classified as representational, and related to the process of comprehension, finally manifesting itself as an error of misinterpretation. The operator failed to follow the procedure triggered by the dashboard light, which by common sense would have been to stop the car and check the oil. The causal mechanism of this error was the operator's mental representation of the vehicle, which provided him with the background knowledge present at the time of the incident, and that triggered the wrong decision of not stopping to check the engine. In other words, if we constrain the error analysis to the operational environment (in this case the car and the driver) and to causal factors intrinsic to the incident report, we find the causal mechanism to be the operators mental representation of the environment, in that it affects his decision-making process. This type of signal is detected by our method and it discloses aspects of the *situation awareness* of the driver at the time of the incident. We see then how our paradigm allows for a shift from an external cause to a representational error, which is related to the process of comprehension in the DRICP model. This approach ,therefore, has the potential to reduce the number of errors due to external causes in an error classification, yielding a better distribution.

The primary goal of this work was to device a new methodology for automated detection of patters in text. Nonetheless, a very important motivation behind this effort was also to enable a technique to be helpful in furthering the understanding of SA and in improving the existing SA error taxonomies used in aviation risk management. The author feels that this goal was accomplished since a new and interesting methodology together with a computer tool were devised successfully. This tool should enable the extraction of information from the ASRS in new and interesting ways when combined with the existing methods and technologies. In the next sections, we will comment on the potential applications of this method as a complement to the current "bag of words" approach. We will also comment on how our new tool can augment current machine learning technologies by adding some *locality* around regions of saliency in the narratives, which can be used for the data mining process.

In sum, this approach can yield the kind of new information necessary to detect error patterns in a way that facilitates revised error taxonomies and SA models. In other words, for better models and better taxonomies to emerge and be properly validated with data, innovative technologies are needed. We hope to contribute in this fashion to the field. We also hope that is clear that our program can yield useful information in regards to SA by shedding light onto the subtle differences between misinterpretation and misperception, and by providing a way to detect potential causal mechanism in an automated fashion. This promotes statistical validation which is a key element in the emergence of new models, and it also promotes the development of better automated tools for preemptive error prevention through more accurate and detailed detection.

Future Research

There are two main areas of relevance for continuing the research initiated in this thesis. One is the introduction of new variables into the analysis process, in order to further refine the idea of subjectivity and increase the level of information available as output. The other area corresponds to the incorporation of this technology and the new conceptual approach into the current efforts underway in the AvSP program.

The main motivation behind introducing new variables is to further operationalize the concept of "subjectivity". This is the pivotal point of our research and therefore where we need to concentrate our efforts in order to increase the level of informational content in our output. This aspect is crucial because our hypothesis requires that we detect levels of high subjectivity successfully and that we are able to discriminate between different *types* of subjective content. In other words, having successfully been able to detect *overall* levels and distributions of subjectivity in ASRS narratives, we also need to understand the different *quality* of that subjectivity, as far as how it relates to the behavioral model. As an example, one could consider the overall subjectivity of a narrative reflecting organizational (e.g., rank-based) subjectivity. One can also consider *responsibility-related* subjectivity which pertains to the need of a writer to justify (excessively perhaps) the rationale or context for his decisions. The first stage of the research has shown successfully that subjectivity can be measured without recourse to keywords or phrases.

In order to introduce new variables into our program, the main area to expand is that of syntactic constructions and embedded clauses. The former one is a *vehicle* in language used to reflect what is called "propositional attitudes", which are the author's belief, evidential stance or affect, put toward the predication being transmitted. The latter is a also a syntactic device but that is used instead to put different levels of *emphasis* in cases of sentences with multiple predicates. Embedded clauses could have several relations with each other and the main predicate: subordination, coordination and complementation. Each one is used for a different functional purpose in conveying information, and together with syntactic constructions (i.e., templates) they can be powerful tools for capturing psychological states about the producer of the text. This is similar to the techniques used by clinical psychologists such as Bierschenk and his "Perspective Approach" (see Chapter 2, "Content Analysis"). The literature in Cognitive Linguistics and Discourse has a large body of work devoted to these so called "propositional relations". Therefore, a good amount of work is available to draw knowledge and inspiration (Couper-Kuhlen & Kortmann, 2000).

As an example, consider the following excerpt,

> HALFWAY DOWN RWY 7 (DELTA) TWR CALLED US AND ASKED US I**F WE HAD CLRNC TO TAXI ON RWY 7**. WE REPLIED **HE HAD CLRD US TO TAXI.** HE SAID **HE HAD GIVEN US CLRNC TO TAXI TO 26R VIA DELTA AND HOLD SHORT RWY 7.** WE WERE ALREADY HOLDING SHORT? WE APOLOGIZED AND THERE WAS NO FURTHER PROB."

versus

> HALFWAY DOWN RWY 7 (DELTA) TWR CALLED US AND SAID **"YOU HAVE CLRNC TO TAXI ON RWY 7"**. **"YOU HAD ALREADY CLRD US TO TAX"** WE REPLIED. HE SAID "**WE'VE GIVEN YOU CLRNC TO TAXI TO 26R VIA DELTA AND HOLD SHORT RWY 7"**. WE WERE ALREADY HOLDING SHORT? WE APOLOGIZED AND THERE WAS NO FURTHER PROB.

Here the use of "quotation" marks, instead of third person passive form, could indicate some form of "organizational" bias. For instance, the pressure or tendency on the part of the narrator to divert the blame toward a different group within the operation, pointing toward some work-related bias. It could also mean that the narrator is establishing some "distance" between the facts and his perception of them, trying to make the story less "subjective", another form of bias.

A beneficial side effect of using syntactic constructs and embedded clause analysis is that it allows us to measure the level of complexity of a sentence and therefore the overall degree of saliency of a given clause. The work on style in the literature (Karlgren, 1999) makes use of this same idea. This information alone is sufficient to enhance current methods of IE. We will discuss more on this shortly.

Following are the major ideas that will drive the development of the new variables, followed by an example and a comment on their significance for analysis.

Embedded general statements or assessments

> "I NOTICED HIS IDENT, AND DID NOT CONSIDER HIM A SECURITY THREAT, BUT **I QUESTION HIS ENFORCEMENT OF SECURITY** (OR AS AN AVIATION SAFETY INSPECTOR) PROCS IN THAT, WHY DID HE ALLOW 2 PEOPLE ONTO THE RAMP ..."

This variable will allow us to detect the introduction of potential biasing information in the form of a *contraposition* (a type of syntactic/discourse construct) embedded into the main predicate through a subordinated relation. Another example can be seen in segment 3 of record 100073, in the previous chapter. It contains causal information and a second level of subjectivity.

Embedded situation time anchor prior to current reference time anchor

"HE SAID HE **HAD GIVEN** US CLRNC TO TAXI TO 26R VIA DELTA AND HOLD SHORT RWY 7"

This type of embedding is useful for capturing a "missed" signal event. In other words, an event missed by the narrator at the time of the incident. A keyword approach could also detect this type of event (e.g., "I did not see" or "I did not notice"), the difference lies in that our approach not only generalizes but also enables the detection of patterns not discovered yet.

Main or embedded subjective verb

"WE **THOUGHT** TWR SAID TAXI TO RWY 26R VIA DELTA "

The idea here is to categorize verbs in a *subjectivity scale,* trying to capture "cognitive events", which are for the most part interpretative. We can then make a distinction between *perceptual* verbs, *communicative* verbs and strictly *cognitive* verbs, such as thinking, deciding, wondering and so on. This information can be obtained in a relatively straightforward manner from WordNet, as described in Chapter 3. Also, we could potentially use this information for detecting "brackets" of subjective interpretation in a given narrative (i.e., subjective areas across sentences). This last type of approach has been researched in the literature fairly successfully, and in a computational fashion (Wiebe, 1994). One interesting example from the narratives we examined in the previous chapter is record 100079, segment 18. The program currently does not capture the cognitive event described there:

> "ANYWAY , I DIDN'T KNOW THE EXTENT OF THE SUSPECTED FUEL CONTAMINATION PROBLEM SO EXERCISED MY EMERGENCY AUTHORITY AND REQUESTED PRIORITY HANDLING FROM CHI CENTER"

This event describes a decision made under poor knowledge of the environment, so it is crucial for our behavioral model to capture it as a *type*. The current implementation captures this sentence as a type S segment, which correctly identifies it as a location in the narrative where the author is describing aspects of the state of the system (in contrast to an event that advances the narrative time and that *affects* the state of the system itself). It would be a significant improvement to capture a fine-grained distinction within these type of events, such as a signal for "poor knowledge", since that information is beneficial for the analysis of SA-related errors.

Main or embedded **negated** perceptual and cognitive verbs

> " AS WE TURNED ONTO THE RUNWAY FOR DEPARTURE , WE **DID NOT NOTICE** AN APPROXIMATE 30 DEGREE HEADING ERROR IN OUR COMPASS SYSTEM"

This type of pattern signals a "missed information" event. It can be very useful to detect patterns where either the source (genotype) or the manifestation (phenotype) of the error are perceptual, without recourse to keywords or phrases. It is another way in which we can capture a concept without relying on *a priori* domain-specific knowledge.

Causal link

"OUR GPS TURNED TO GO DIRECTLY TO BOULDER
VERYHIGHFREQUENCYOMNIDIRECTIONALRADIORANGE **INSTEAD** ."

This is the only variable where keywords might have to come into play. These are domain-independent keywords, such as "instead", "but" or "therefore" and so on, and they are limited in number. The literature on Discourse Analysis has some valuable resources for the acquisition of these *causal link* keywords. Marcu's work on discourse parsing(Marcu, 2000) contains a compilation of over 400 of these words collected over the course of his thesis research. The idea here is to combine subjective areas and causal links to detect locality in the narrative where some form of cognitive biasing could have affected decision-making. The burden of using keywords to detect these patterns of causality could be lessened by employing a combination of syntactic patterns as well, but this is yet to be investigated. The goal here would be to detect *brackets* of subjectivity that span across sentences and that relate to a potential causal mechanism for the error.

Use of quotation marks

“21.7 WORST CASE ON THE GAUGES" THE F/E SAID

It can be argued that the use of quotation marks changes the level of subjectivity in a given portion of a narrative. In order to determine if this is significant, there should be no such *style* existing in the data sample pool. In other words, if there is no cluster of narratives present in our ASRS record selection which incorporates a significant usage of quotation marks, then we can quantify and use this measure as a potential detection signal. From the theoretical standpoint, this type of grammatical style has indeed been shown to indicate " distance of belief". That is, the author is placing some degree or distance about his belief with respect to the content of the predication. The nomenclature used in the literature to refer to this idea includes: "Subject of Consciousness" (SOC) distance and "epistemic" versus "content" relations. This type of information together with the previously described detection of "cognitive" verbs can be used to yield an interesting measure about *degrees* of subjectivity. It relates to the previously introduced concept of "propositional attitudes" in that it reflects a relationship of belief or attitude on the part of the author of the narrative toward what is being said. This is important information from the standpoint of the behavioral model used in the AvSP, and it allows us to establish another variable in the operationalization of the idea of subjectivity.

Embedded event of opposite type

"**A** [WE WERE ABOUT 10 MILES SOUTH OF BEATTY WHEN THE
CONTROLLER **B** [**ASKED** US TO TURN TO JOIN THE COURSE
OUT OF BEATTY]] "

This parameter captures a different way of interleaving States and Events. Its significance is not entirely clear at this point, but it nevertheless contains information about the clause *complexity* of a given segment.

The next area of interest for commenting is the integration of our approach with the current efforts under way in the AvSP. This involves both a conceptual integration as well as a technological one. The argument for this integration comes from a variety of reasons:

1. The "bag of words" approach requires a fair amount of *a priori* domain-specific knowledge. It also produces a high degree of specificity in terms its output (i.e., identifies events that can be directly correlated with other data). Our approach generalizes well but contains less specificity in terms of the level of the informational content of its output (i.e., sequences of segments, general types of events, no direct reference to any other data). On the other hand, it does not rely on domain-specific knowledge or requires any *a priori* knowledge besides general linguistic models.
2. The current effort underway at Battelle, employs a pattern language called JAPE (Java Annotation Patterns Engine) used to capture patterns consisting of a combination of keywords, phrases and part-of-speech tags (i.e., meta-language functional information). This process requires a considerable amount of expert input and manual coding to obtain reasonable results, but it has a high degree of specificity and accuracy. Our proposed method, on the contrary, does not require any input from domain experts. Again, and as in the previous point, this comes to the

expense of a much lower resolution and specificity in the output of the program.

3. The "bag of words" approach treats narratives in a homogeneous manner, not making distinctions on any particular portion of the narratives. The bag of words approach does not even make any distinctions based on syntactic structure or hierarchy. This can be seen in the description of the main algorithms behind tools like QUORUM (see Chapter 2 "Information Retrieval"), which rely on feature vector spaces that are blind to the intricacies of language syntax. Our method relies almost exclusively on syntax and on general semantic categorical features about words. In doing so, it gives "texture" to the narratives in terms of their temporal structure. This means that it creates areas of locality around salient "flags", that is, relevant syntactic signals. The output from our program therefore converts *flat* text into *structured* data.

Considering point 3 above, we can identify at least one area where our method could help improve the process of automated detection. A particular technology called "machine learning", is widely used at the moment for a great variety of NLP problems (including the part-of-speech tagger used in this work, see Appendix D). Machine learning technologies employ templates to induct patterns from language. These techniques are particularly powerful given the computational resources and the amount of language data available nowadays.

A technique such as machine learning could benefit tremendously from our segmentation output, for it can be used as a "guiding template" for the algorithms to induct new patterns not previously known or expected. The fact that the output from our program creates areas of saliency, further motivates the usage of such automated tools. This is in fact the reason why we incorporated the same XML DTD (Document Type Definition) as part of the output format. It will create a seamless transition to test and employ our technology.

Another area for synergistic interaction between the two approaches is acting as a filtering process for pre-classification purposes. Our method can serve as a filter records containing high subjectivity measures. Even within a single narrative we have seen (Chapter 4) how entire sections of the text appear to frequently contain opinionated or extrinsic information to the story of the incident. We have received very positive feedback on this idea both from Battelle and the Complex Systems and Data Mining Group at NASA. Therefore we feel strongly that our tool will indeed be of help in the AvSP project. Ultimately the decision is contingent on the field experts and members of the AvSP project.

Accuracy and Robustness

This section will cover three areas deemed important to evaluate and improve the robustness of our program and methodology. The first one is with regards to the statistical analysis of the data. It will involve the use of ranking techniques and correlation tests. Secondly, we will comment on accuracy tests to be performed on our program. We will use a gold standard (manually coded narratives) to evaluate the output of a selected set of narratives. Finally this section will describe the caveats and technical difficulties encountered in the development of our program. We will describe techniques that will help improve the accuracy of the program, and explain how to integrate them.

Statistical Tests

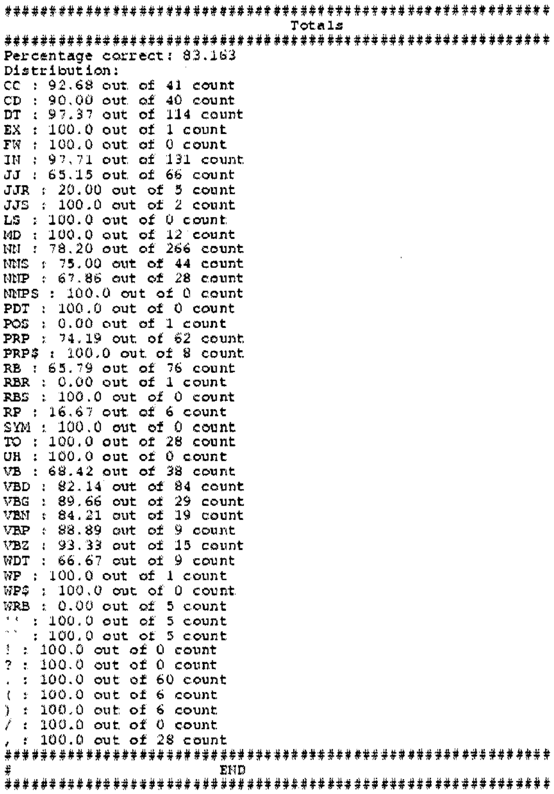
It seems to be common knowledge in the NLP community that stylistic elements in language do *not* follow normal distributions (Karlgren, 1999). This is particularly true when there is such heterogeneous and indirectly measurable set of variables, as is it the case here. Therefore the techniques employed have been nonparametric. This is part of the motivation for our decision to employ clustering to analyze the results from our program, but it is also because clustering is inherently an exploratory tool and our main goal was just that. We wanted to explore the data distribution that would occur given our syntactic definition of style, and to be able to use that as our baseline for comparison.

After examining the results in Chapter 4, the reader is left with a feeling that there is more there than meets the eye. We would like to further explore our output data matrix and clusters to be able to infer more detailed patterns. In order to do so we will have to employ other nonparametric or "distribution free" tests to find statistical significance in relationships among our variables. One of these could be using a correlation test, such as the Spearman Rank Order Correlation Coefficients (rho), to understand if there are any strong correlation among the complete set of variables. Once we are able to identify any correlations we can make a better decision on the selection of these variables for use in our clustering. In other words, we could revise our clustering based on a new set of variables so that we can maximize either the *isolation* or the *compactness* aspect of these groups.

Another test we can perform is ranking, such as the Mann-Whitney or Wilcoxon U Rank Sum Test. This could be done in order to establish a threshold for differences between clusters, and between variables across clusters, to better understand if the differences among the groups are significant or not. One thing to keep in mind here is that all of the nonparametric techniques mentioned here still have some assumptions underlying them. Even though no assumptions about normality are adopted for the data, the *random sampling hypothesis* must still hold. That is, we are still making the assumption that our experimental error is evenly distributed and consistent. In other words, we are assuming that the overall distribution of our sampling error is equal to the join distribution of each individual measurement error.

Accuracy Tests

A test performed on the *preprocessing* output of our program yielded a measure of 83% correct hits. This test was performed against a gold standard of nine randomly selected narratives from the ASRS database. The gold standard consists of manually POS-tagged narratives, following the guidelines of the Penn Treebank POS tag set and notation (Treebank, 2004). The complete tag set employed is shown in Appendix F. An 83% hit ratio is a satisfactory result considering that the literature indicates that percentages around 90% are state-of-the-art in NLP technology (Jurafsky & Martin, 2000). This is specially true due to the fact that tests are performed using training data of the same origin as the test data. This is not the case here where the tagger we employed was trained in a different corpus than that of the ASRS. The test was performed via a script written by the author, where POS tags are counted by type, and the hit ratio is calculated from the number of equal matches between the output of the program and the gold standard. The output of this program is shown in Table 10 below.



```
##############################################################
                               Totals
##############################################################
Percentage correct: 83.163
Distribution:
CC : 92.68 out of 41 count
CD : 90.00 out of 40 count
DT : 97.37 out of 114 count
EX : 100.0 out of 1 count
FW : 100.0 out of 0 count
IN : 97.71 out of 131 count
JJ : 65.15 out of 66 count
JJR : 20.00 out of 5 count
JJS : 100.0 out of 2 count
LS : 100.0 out of 0 count
MD : 100.0 out of 12 count
NN : 78.20 out of 266 count
NNS : 75.00 out of 44 count
NNP : 67.86 out of 28 count
NNPS : 100.0 out of 0 count
PDT : 100.0 out of 0 count
POS : 0.00 out of 1 count
PRP : 74.19 out of 62 count
PRP$ : 100.0 out of 8 count
RB : 65.79 out of 76 count
RBR : 0.00 out of 1 count
RBS : 100.0 out of 0 count
RP : 16.67 out of 6 count
SYM : 100.0 out of 0 count
TO : 100.0 out of 28 count
UH : 100.0 out of 0 count
VB : 68.42 out of 38 count
VBD : 82.14 out of 84 count
VBG : 89.66 out of 29 count
VBN : 84.21 out of 19 count
VBP : 88.89 out of 9 count
VBZ : 93.33 out of 15 count
WDT : 66.67 out of 9 count
WP : 100.0 out of 1 count
WP$ : 100.0 out of 0 count
WRB : 0.00 out of 5 count
'' : 100.0 out of 5 count
`` : 100.0 out of 5 count
! : 100.0 out of 0 count
? : 100.0 out of 0 count
. : 100.0 out of 60 count
( : 100.0 out of 6 count
) : 100.0 out of 6 count
/ : 100.0 out of 0 count
, : 100.0 out of 28 count
##############################################################
#                        END
##############################################################
```

Table 10. Part-Of-Speech Test

For the second part of our accuracy measure we will use the four narratives presented in the previous chapter, and manually evaluate the accuracy of each segment type against the author's criteria, based on the original design and linguistic theory. There will be two categories of errors:

1. Preprocessing errors (due to bad syntactic parses or poor dis-abbreviation).
2. Conceptual (syntactic templates need to be refined or new ones added).

It is important to note here that segments produced as type X, representing unknown events (white segments), are due exclusively to either one of these two error sources. The main program has a built-in filter to eliminate narratives containing more than an allowed threshold of X-type segments. For this work the filter was set to 20%. This setting yielded a total of 88 records out of the 100 original set provided by Battelle-PNWD. This means we are working with a baseline that has a 20% ceiling on this type of error (the approximate bottom was 11% as stated in Chapter 4). The next section discusses ways to improve this aspect of the program.

The first narrative, 100079 shown in Illustration 18, has a total of 19 sentences and 23 segments as processed by our program (recall that segments are produced clause-wise, so a sentence could consist of more than one segment). The most visible error here is segment 3, which is shown in white, indicating it is of type X (unknown). The debugging information produced by the program indicates this is a bad syntactic parse. In other words, it is a very complex or ungrammatical sentence that was not parsed properly by the parser, but it could also mean that our parser simply did not do a good job. Illustration 22 below shows the partial syntactic tree output for this clause (the picture shown has been cropped for visualization purposes). The black arrows and circles show the two top "S" nodes where the problem originates. These two S nodes should have been either coordinated by the comma or the connective word "then," or alternately, the second clause should have been subordinated to the first one.

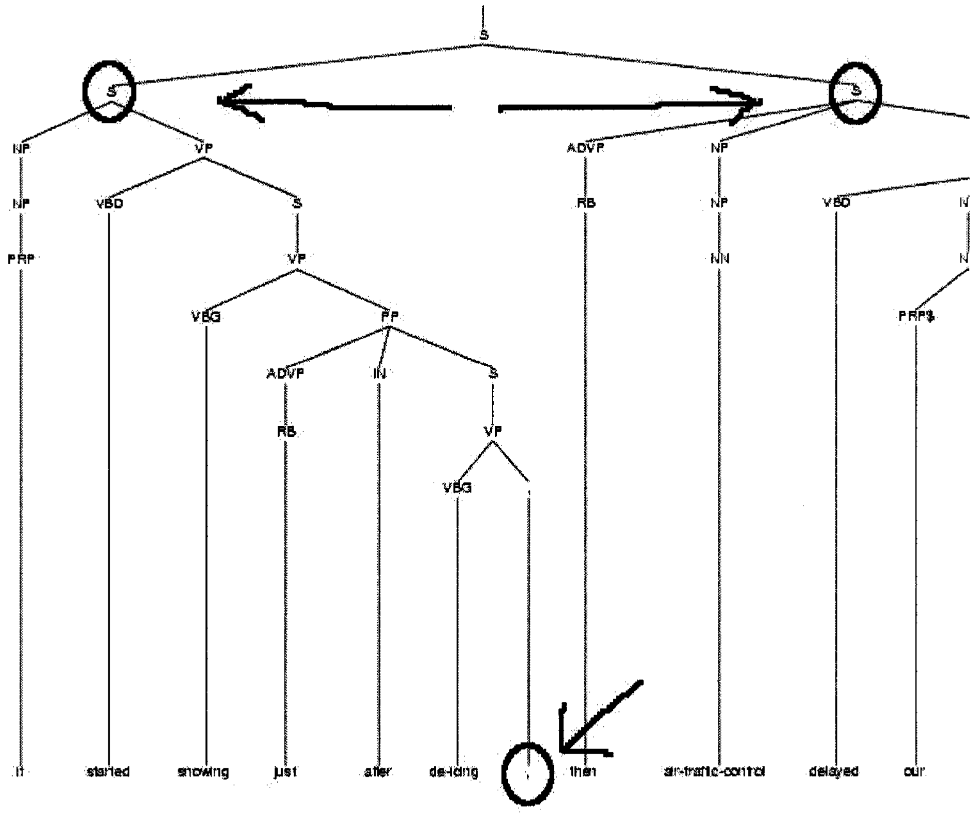


Illustration 22. Bad Syntactic Parse

The second error in narrative 100079 follows from the first one, where segments 4 and 5 should really be of type "S" instead of "St" (because segment 3 is really of type E). This makes the introductory portion of the narrative longer and consequently, reduces the value of the S-type ratio variable, which comes into play in our second clustering variable. This would have brought the S.ratio variable up to 0.52 from 0.43 and the E.ratio variable to 0.3 from 0.26. The E.to.S.block ratio variable, the second variable in our clustering procedure, would have change its value from 0.6 to 0.583. This is not a considerable change, and the overall percentage of errors in record 100079 is therefore low as well, roughly 13% (3 incorrect segments out of 23 total).

The next three narratives are record 100040, 100066 and 100073. The first one (Illustration 19) shows 4 X-type segments out of a total of 23 (17%). There are two segments, number 2 and 3, which could have arguably been segmented as type S instead of E. The decision of making communicative verbs State types rather than Events is still not clear at this point (i.e., " *maintenance **said** it was ice melting*"). Another potential source of error is a block of A-type segments located toward the end of the narrative. They correspond to the conclusion portion of the story; they are followed by a series of E and S-type segments. The information in those final E and S-type segments also belongs to the conclusion portion of the story. They are *grammatically* different from A-type because the author is bringing supplementary information from other sources. It would be fairly straightforward to incorporate logic into the program to detect and correct this problem. These occurrences could be measured by detecting contiguous A-type blocks in the proximity of the end of a narrative, and then converting any interleaved E and S-type segments into A-type. The program could measure the distance to the end of the narrative and the proportion of A blocks versus the other types, to make a decision based on a preset threshold and convert the segment blocks adequately. We do not consider this a true error for our accuracy measure because it really corresponds to an improvement to the program. Record 100066 (Illustration 20) shows only one X-type segment out of 26 total (4%). Segment 10, marked as S-type, shows an interesting case: " *i elected to continue to sea -LRB- original destination -RRB- due to improving weather , fire and crash , airport familiarity and altitude-or-alternate-or-hold-altitude-hold-mode -LRB- landing at panache would have required steep des , rushed completion of engine failure checklist and des & approach and landing checklists*". "LRB" and "RRB" stand for left and right round parenthesis respectively. This segment is matched successfully against State template number 2 in Appendix A. The main verb "elected" is populated as a cognitive verb via WordNet. What is interesting about this particular segment is its complexity and the fact that it really marks an "event" that changes the course of events in the story. Unfortunately it is very hard to determine this fact computationally without recourse to keywords or domain-specific knowledge. Ideally there should be a distinction within the cognitive verb categories to discriminate verbs with potential for changing the course of a story (i.e., decision-making related verbs). The complexity level of the clause will be detected an quantified when the type of ideas described in the previous section ("Future Research") are implemented.

Finally, record 100073 (Illustration 21) shows four segments of type X over 47 total (8.5%). Segment 3 presents an instance of the verb "begin". It is not clear whether this verb should be considered a State or an Event trigger. At this point it was decided to treat it as an Event type, because for the most part it does signal a change in the state-of-affair described in a narrative. One thing to note here is the large contiguous block of S segments, from 4 to 8. It is created by a series of descriptions of mental processes related to assessment and decision-making at a given point in the narrative time and for a set of system conditions. Segment 9 is an E-type because it portrays a communication act, and is then followed by a fairly lengthy series of E-type segments (13-17 and 19-21). What this signals, overall, is the discrimination between descriptions of mental states and descriptions of external events that move forward the story timeline. This is indeed the original motivation behind our segmentation and here it is seen at work. In other words, when we set out to design our situation type segments, our goal was to capture the difference between a portrait of the state of a system as seen in the mind of the narrator (States), and events which stand out as temporal landmarks in that they advance the story line and mark points at which the state of the system changed (Events). Segment 29 shows a true error. This segment should really be categorized as S-type. The reason why it is segmented as an E-type is because there is a typo in the verb "descended" where it should really be "descend", since it is located within an infinitival construction with a marker "to", i.e., "to descend". Unfortunately this is the one type of error that would be extremely difficult to prevent. Even sophisticated tools such as PLADS cannot capture this type of error robustly. This is because tense inflection tends to be eliminated during normalization when using pure statistical models of language. One last thing to note in this narrative is segment 44 and 45. Both exhibit the use of colon ":". Syntactic parsers seem to have difficulty in parsing this type of sentences correctly. Some manual intervention will be needed to solve this problem. This is feasible because the lexical element ":" is easy to detect due to its homogeneous grammatical function. Also a manually coded solution to this problem would still be generalizable, since we are not bringing any domain knowledge into the picture. By "manual" here it is meant hard-coding the program to detect this directly, not actual manual editing.

Improvements

There are two main areas in our method where improvements can be implemented. One is in the preprocessing stage of the program, aiming at improving errors related to parsing. The other one is in the development of more targeted, more problem-specific, and more fine-grained templates to match against narratives in order to increase the level of the informational content of the output. In the first area, that of preprocessing, there are two main steps to take. One is the use of a better syntactic parser, and the other one is to perform what is known in the NLP literature as "Terminology Extraction" or "Collocation Analysis" (Manning & Schütze, 1999).

For this work, we employed the Collins' parser and Brill's tagger combination. This combination of tools is one that has been the standard in the field for the last several years. This is because they were among the first NLP tools of their kind to be publicly available and work well together in terms of coupling their respective inputs and outputs. Even though they remain world-class tools, they have been surpassed by other implementations, such as the Stanford Lexical Parser. This is probably due to the amount of effort that has been put into these new technologies, and that is driven by the research volume underway in those institutions. Therefore, we have conducted preliminary test and comparisons with Stanford's parser, and have discovered significant improvements in the overall quality of the parsing output. This syntactic parser is lexicalized probabilistic context-free grammar, which is the state-of-the-art in NLP at the moment.

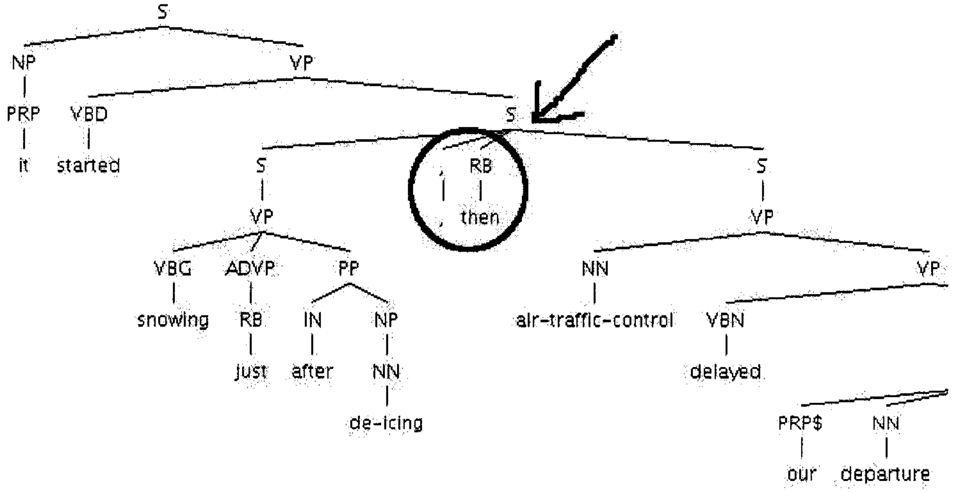
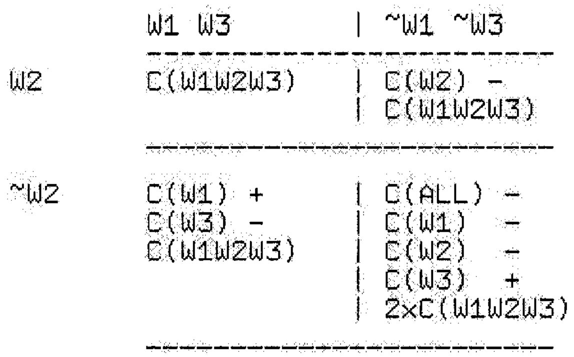


Illustration 23. Improved Syntactic Parse

Illustration 23 above shows the same sentence of Illustration 22, parsed by Stanford's parser instead of Collins. We can see how the problem of coordination between the two predications is solved by subordination, and the comma and the "then" adverbial are captured correctly as relational elements.

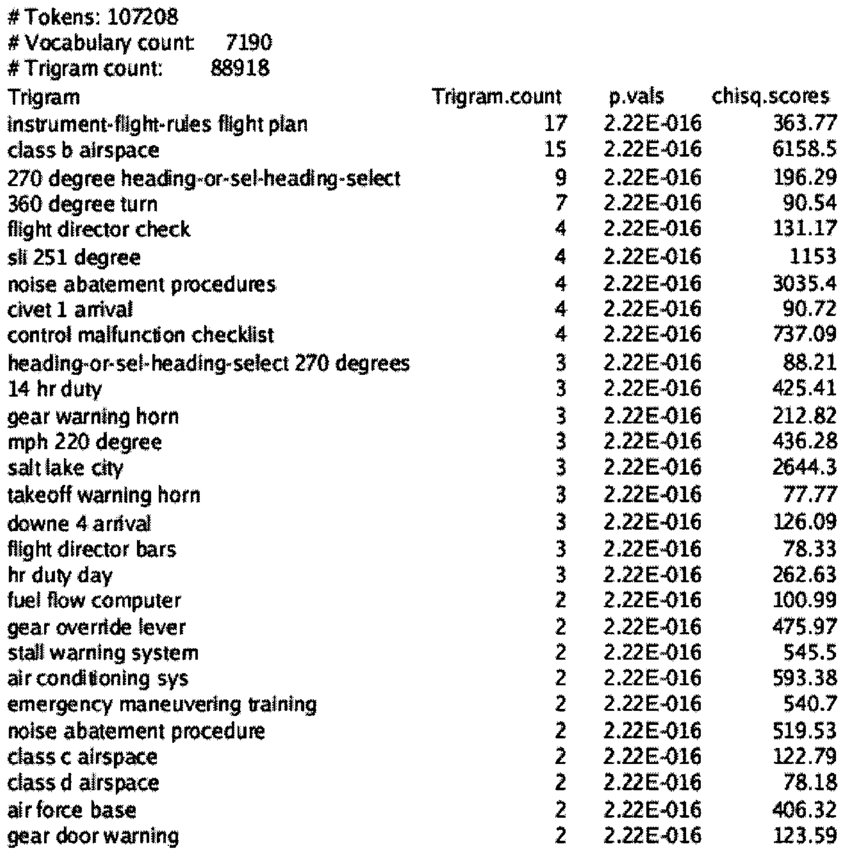
An NLP technique we can take advantage of is *terminology extraction*. This technique refers to the use of common statistical tests, such as the chi-square test, to discover *collocations* among words in a given corpus. These collocations could represent domain-specific terminology. This technique is particularly useful in a case like ours, since our corpus is highly technical and repleted with terminology. The use of the chi-squared test is to identify these collocations in terms of their *expected* values. That is, the test sets up contingency tables that allow for the comparison of expected values of collocations (in terms of the conditional probability of sequences and count of individual words) to the actual frequency count of bigrams and trigrams (i.e., two and three word sequences of words). Table 11 below shows an example.



	W1 W3	~W1 ~W3
W2	C(W1W2W3)	C(W2) - C(W1W2W3)
~W2	C(W1) + C(W3) - C(W1W2W3)	C(ALL) - C(W1) - C(W2) - C(W3) + 2xC(W1W2W3)

Table 11. Chi-Square Contingency Table

The contingency table shown above is for *trigram* collocations, that is, co-occurrences of three words. Normally bigrams suffice for this kind of analysis, but our domain is particularly heavy on rather lengthy technical terminology. W1 stands for "word 1", C(W1) stands for the count of occurrences of word 1, and C(W1W2W3) is the count of a trigram. A preliminary test was run with a relatively small sample from the ASRS corpus. A portion of the results is shown in Table 12 below.



Tokens: 107208
Vocabulary count 7190
Trigram count: 88918

Trigram	Trigram.count	p.vals	chisq.scores
instrument-flight-rules flight plan	17	2.22E-016	363.77
class b airspace	15	2.22E-016	6158.5
270 degree heading-or-sel-heading-select	9	2.22E-016	196.29
360 degree turn	7	2.22E-016	90.54
flight director check	4	2.22E-016	131.17
sli 251 degree	4	2.22E-016	1153
noise abatement procedures	4	2.22E-016	3035.4
civet 1 arrival	4	2.22E-016	90.72
control malfunction checklist	4	2.22E-016	737.09
heading-or-sel-heading-select 270 degrees	3	2.22E-016	88.21
14 hr duty	3	2.22E-016	425.41
gear warning horn	3	2.22E-016	212.82
mph 220 degree	3	2.22E-016	436.28
salt lake city	3	2.22E-016	2644.3
takeoff warning horn	3	2.22E-016	77.77
downe 4 arrival	3	2.22E-016	126.09
flight director bars	3	2.22E-016	78.33
hr duty day	3	2.22E-016	262.63
fuel flow computer	2	2.22E-016	100.99
gear override lever	2	2.22E-016	475.97
stall warning system	2	2.22E-016	545.5
air conditioning sys	2	2.22E-016	593.38
emergency maneuvering training	2	2.22E-016	540.7
noise abatement procedure	2	2.22E-016	519.53
class c airspace	2	2.22E-016	122.79
class d airspace	2	2.22E-016	78.18
air force base	2	2.22E-016	406.32
gear door warning	2	2.22E-016	123.59

Table 12. Chi-Square Test Results

Table 12 shows that for a relatively small corpus of 6,500 lines, 107,208 words (tokens) and 7,190 unique words (vocabulary), we can obtain interesting results with potential for improving our syntactic parsing significantly. We can see for instance that the phrase "class b airspace" has a very high chi-square score (i.e., low p-value), which is not surprising given the nature of our data (could be generalized to "class x airspace"). This type of information can be used effectively in our program by pre-tagging these collocations at the preprocessing stage. In this way the parser will have more information available and will perform better under ambiguity in the syntax. For instance, in record 100068, there is a segment (7) containing the phrase "*repeated restart procedures*" which is incorrectly classified as A-type whereas it should be of type E. The reason for this incorrect match is that the parser identified the verb "restart" as the main verb, instead of "repeated" which is in past tense. If our collocation work had identified "restart procedure" as significant terminology, then our pre-tagging scheme would have tagged the phrase as "restart-procedure//NN" which would then have been considered a noun (i.e., NN), forcing the selection of "repeated" as the main verb. A similar example is found in record 100044 segment 4. The clause *"after lift off the aircraft had a considerable right roll"* is segmented as an A-type, whereas it should really be an S-type. The reason for this incorrect match is the parser selecting "lift" as the main verb. If we had pre-tagged "lift-off//NN", then the parsing would have been corrected.

Conclusions

We set out to make our contribution in the AvSP program by devising a novel approach and tool for automated analysis of text narratives. We envisioned a method independent of domain-specific knowledge and that could serve as a complement to the undergoing efforts in the field. This could be a fairly daunting task. Fortunately we were armed with an interdisciplinary array of theoretical background and technologies. We feel that we have succeeded in the creative and technical aspects, and that we have opened new and interesting directions for future research. In this sense, we are confident that we have made our humble contribution not only to the ASMM and AvSP efforts, but to the NLP community in general. The problem of extracting information from language, and in particular information related to human behavior, is a very important one. It is a task still in its infancy, primarily due to the lack of unified models of cognition but also due to our scarce knowledge about the link between language and mental representations. Therefore our contribution should be welcomed as it provides a new tool for gaining insight into the problem. Our approach, we believe, yields some new and interesting possibilities for study, in particular for such applied problems and bounded domains as is the case for the AvSP project.

Our solution was primarily inspired by the increasing amount of scholarly work that has emerged since the "cognitive revolution" of the nineteen fifties and sixties. In particular, new linguistic theory and psychology models of memory have been key to the conceptual development of this work. We drew direct theoretical background from linguistic semantics, specifically from current work on situation types from the Cognitive Linguistics field. With this theoretical tool set in hand, we designed a set of syntactic templates to match different types of "mental events" as they manifest through grammar in textual narratives. These mental events portray different qualities about a story, and together form sequences which we labeled "styles". Our method studied these styles and how their distribution is indicative of certain type of "biasing" on the part of the speaker in the narration.

The results from this work showed that it is possible to use our approach and our computational tool to extract useful information from the ASRS narratives. In particular our approach is well suited to discriminate highly subjective narratives, that is, very opinionated reports with a high content of extrinsic information. We have also shown that there is tremendous potential in our data-coding method to further increase the level of granularity in the discrimination of "biasing" signals. These signals have great potential as a diagnosis tool for determining error causal mechanisms. We have also argued that improved error taxonomies are possible using methods specifically designed to detect subtle behavioral patterns in language. Our method is able to detect such subtleties. We presented a revised SA-based error taxonomy together with an example of a classification using our method.

We feel that we have achieved our goal of creating a method that is generalizable and robust, relying only on general models of language and memory. Our other goal was to provide a complementary tool to the current technologies in NLP. We believe we have done so by including these technologies in the design process of our program. Further research and collaboration with Battelle-PWND will bring out a better understanding about this integration. In sum, we feel that we have made a small yet important contribution to the NLP and Human Factors community, and to the AvSP program.

REFERENCES

Adams, M. J., Tenney, Y. J., & Pew, R. (1995). Situation Awareness and the Cognitive Management of Complex Systems. Human Factors, 37, 85-104.

AirOdyssey.net (2004). AirOdyssey.net - Reference of aviation terms. [On-line]. Available: http://www.airodyssey.net/reference/

Akira, T. & Tokunaga, T. (2001, November). Automatic Disabbreviation by Using Context Information. Proceedings of Natural Language Processing Pacific Rim Symposium. Information Processing Society of Japan, Tokyo, Japan.

Baxter, G. D. & Ritter, F. E. (1999). Towards a classification of state misinterpretation. Engineering Psychology and Cognitive Ergonomics, 3, 35-42.

Biber, D. (1993). The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation. Computers and the Humanities, 26, 331-345.

Card S. K., Moran T. P. & Newell A. (1983). The Psychology of Human-Computer Interaction. New York: Lawrence Erlbaum.

Carmino, A., Nicolet, J. L. & Wanner, J. C. (1990). Man and Risk. New York: Marcel Dekker Inc.

Couper-Kuhlen, E. & Kortmann, B. (Eds.). (2000). Cause,Condition,Concession,Contrast. Berlin: Mouton de Gruyter.

Doane S. & Sohn Y. W. (2002). Evaluating Comprehension-Based User Models. User Modeling and User-Adapted Interaction, 12, 171-205.

Endsley, M. (1999, May). Situation Awareness and Human Error: Designing to Support Human Performance. Proceedings of High Consequence Systems Surety Conference. Sandia National Laboratory, Albuquerque, NM.

Endsley, M. R. (2000, July-August). Situation Models: An Avenue to the

Modeling of Mental Models. Proceedings of Human Factors and Ergonomics Society 44th Meeting. Human Factors and Ergonomics Society, San Diego, CA.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors, 37, 32-64.

Ericsson, K. A. & Simon, H. A. (1993). Protocol Analysis:Verbal Reports as Data (2nd Ed.). Cambridge: MIT Press.

Federal Aviation Administration (2004). Aviation Acronyms and Abbreviations. [On-line]. Available: http://www.gps.tc.faa.gov/glossary.html

Fellbaum C. (1990). English verbs as a semantic net. International Journal of Lexicography, 3, 278 – 301.

Flach, J.M. & Warren, R. (1995). Active psychophysics: The relation between mind and what matters. In Flach, Hancock, Caird & Vicente, Global perspectives on the ecology... (pp. 189 - 209). Hillsdale, NJ: Erlbaum.

Gamma, E., Helm, R., Johnson, R. & Vlissides, J. (1995). Design Patterns. New Jersey: Addison-Wesley.

Gibson, J. J. (1977). The Theory of Affordances. In Shaw, R. & Bransford (Eds.), Perceiving, Acting, and Knowing (pp. 67-82). Hillsdale: Erlbaum.

Givon, T. (1995). Coherence in text vs. coherence in mind. In Givon, T., Gernsbacher, M.A. (Eds.)., Coherence in Spontaneous Text (pp. 59-116). Philadelphia: John Benjamins.

Godden, K. S., Uthurusamy, R. & Means, L. G (1993). Extracting Knowledge from Diagnostic Databases. IEEE, 7, 27-38.

Hancock, P. A. & Smith, K. (1995). Situation Awareness Is Adaptive, Externally Directed Consciousness. Human Factors, 37, 137-148.

Holland, J. H., Holyoak, K. J., Nisbett, R. E. & Thagard, P. A. (1986). Induction. Cambridge, MA: MIT Press.

Janda, L. (2000, February). Cognitive Linguistics. Proceedings of Slavic

Linguistics 2000. Indiana University, Bloomington, IN.

Johnson-Laird, P. N. (1983). Mental Models. Cambridge: Harvard University Press.

Johnson-Laird, P. N. (2001). Mental models and deduction. Trends in Cognitive Science, 5, 434-442.

Jurafsky, D. & Martin, J. (2000). Speech and Language Processing. New Jersey: Prentice Hall.

Karlgren (1999). Stylistic Experiments in Information Retrieval. In Tomek Strzalkowski (Ed.)., Natural Language Information Retrieval (pp. 147-166). The Netherlands: Kluwer Academic Press.

Kaufman, L. & Rousseeuw, P. J. (1990). Finding Groups in Data. New York: Wiley.

Kintch, W. & Ericsson, K. A. (1995). Long-Term Working Memory. Psychological Review, 102, 211-245.

Klein, G. (1989). Recognition-Primed Decisions. In Rouse, W. B. (Ed.), Advances in Man-Machine Systems Research (pp. 47-92). Greenwich: JAI Press. Linguistic Data Consortium (Producer). (1999). The Penn Treebank Project Release 3 [CDROM]. Pennsylvania.

Maille, N., Rosenthal, L., Shafto, M. & Statler, I. (2004). What Happened, and Why: Towards an Understanding of Human Error. Research Report, NASA, Mountain View, CA.

Mann, W. C. & Thompson, S. A. (1988). Rhetorical Structure Theory:toward a functional theory of text organization. Text, 83, 243-281.

Manning, C. D. & Schütze, H. (1999). Statistical natural language processing. Cambridge: MIT Press.

Marcu, D. (2000). The theory and practice of discourse parsing. Cambridge: MIT Press.

Marr, D. (1975). Approaches to biological information processing. Science, 190, 875-876.

Maturana, H.R. & Varela, F. G. (1970). Biology of Cognition. In , Autopoiesis and Cognition (pp.). Dordrecht, Boston: Reidel.

McGreevy, M. W. (2001). Searching the ASRS Database Using QUORUM Keyword Search. NASAReport Interal Report TM-2001-210913.

Miller, G. A. (1990). Nouns in WordNet: a lexical inheritance system. International Journal of Lexicography, 3, 245 – 264.

Miller G. A., Beckwith R. , Fellbaum C., Gross D. & Miller K. J. (1990). Introduction to WordNet: an on-line lexical database. International Journal of Lexicography, 3, 243.

Moray, N. (1987). Intelligent aids, metal models, and the theory of machines. International Journal of Man-Machine Studies, 27, 619-629.

NASA-FAA (2003). ASRS Database Report Set. [On-line]. Available: http://asrs.arc.nasa.gov/

Neisser, U. (1976). Cognition and Reality. New York: Freeman.

Pool, I. D. S. (Ed.). (1959). Trends in Content Analysis. Urbana: University of Illinois Press.

Popping, R. (2000). Computer-assisted Text Analysis. New Delhi: Sage.

Rasmussen, J. (1982). Human Errors. A Taxonomy for Describing Human Malfunction. Journal of Occupational Accidents, 4, 311-333.

Rasmussen, J., Pejtersen, A. M. & Goodstein, L. P. (1994). Cognitive Systems Engineering. New York: John Wiley & Sons, Inc..

Reason, J. (1990). Human Error. Cambridge: Cambridge University Press.

Saeed, J. I. (1997). Semantics. Oxford: Blackwell Publishers.

Sarter, N. B. & Alexander H. M. (2000). Error Types and Related Error Detection mechanisms in the Aviation Domain. The International Journal of

Aviation Psychology, 10, 189-206.

Shappell S. A. & Wiegmann D. A. (1997). Human Factors Analysis of Postaccident Data. International Journal of Aviation Psychology, 7, 67-81.

Smith, C. (2003). Modes of Discourse: the local structure of texts. Cambridge: Cambridge University Press.

Sperber, D. & Wilson, D. (1986). Relevance, communication and cognition. Oxford: Basil Blackwell.

University of Pennsylvania (2004). The Penn Treebank Project - POS tagging guidelines. [On-line]. Available: http://www.cis.upenn.edu/~treebank/

Vendler, Z. (1957). Verbs and Times. The Philosophical Review, 66, 143-160.

West, M. Ed. (2001). Theory, Method and Practice in Content Analysis. Westport: Ablex Publishing.

Wiebe, J. (1994). Tracking Point of View in Narrative. Computational Linguistics, 20, 233-287.

Wilson, J. R. and Rutherford, A. (1989). Mental models: theory and application in human factors. Human Factors, 31, 617-634.

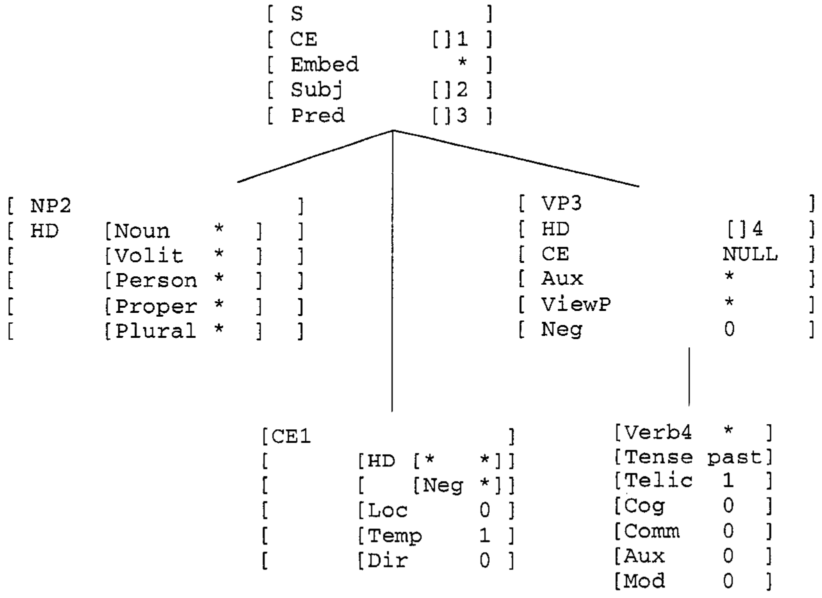
APPENDICES

APPENDIX A. Syntactic Templates

The following is a representation of the s used for matching in the program. The symbols represent the following: '*' stands for any value, 1 and 0 stand for present and not present respectively, NULL means empty.

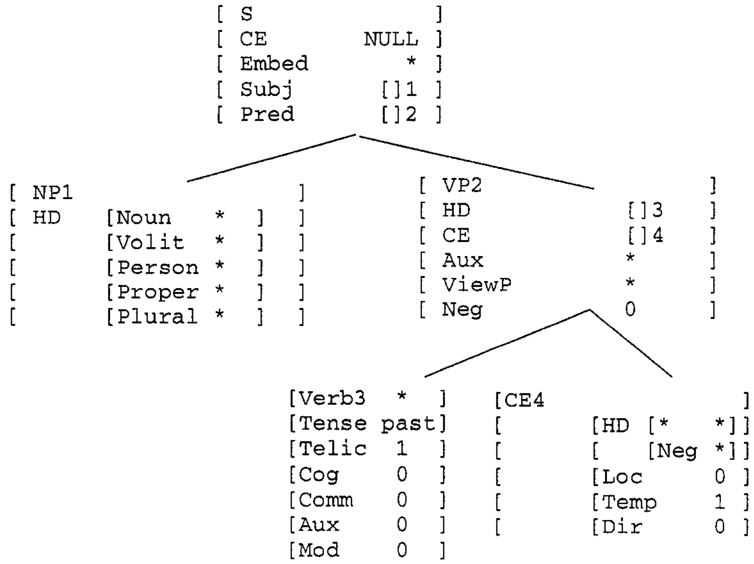
Events:

1. Derived type: from S->E . Any subject NP. VP head is past tense, regardless of the viewpoint, no modals, and the presence of a temporal CE

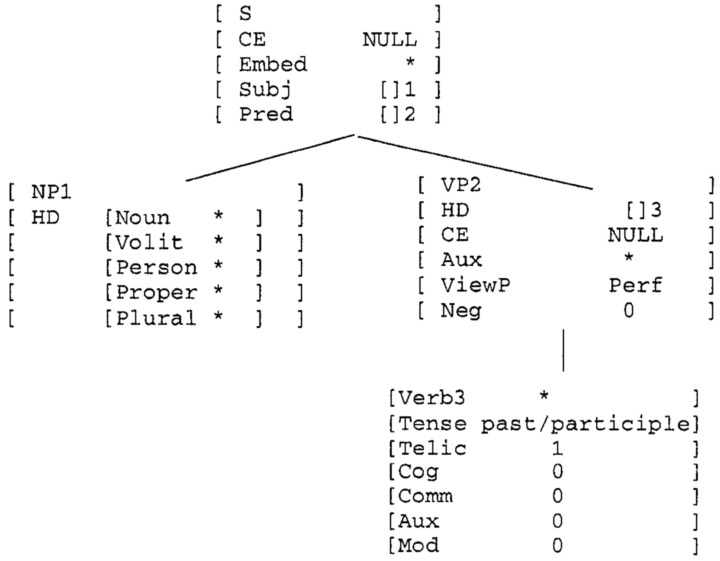


```
                          [ S               ]
                          [ CE          []1 ]
                          [ Embed         * ]
                          [ Subj        []2 ]
                          [ Pred        []3 ]

[ NP2                  ]                   [ VP3                  ]
[ HD   [Noun    *  ]   ]                   [ HD          []4      ]
[      [Volit   *  ]   ]                   [ CE          NULL     ]
[      [Person  *  ]   ]                   [ Aux         *        ]
[      [Proper  *  ]   ]                   [ ViewP       *        ]
[      [Plural  *  ]   ]                   [ Neg         0        ]

                          [CE1                ]     [Verb4   *  ]
                          [      [HD [*    *]]      [Tense past]
                          [      [    [Neg *]]      [Telic   1  ]
                          [      [Loc      0 ]      [Cog     0  ]
                          [      [Temp     1 ]      [Comm    0  ]
                          [      [Dir      0 ]      [Aux     0  ]
                                                    [Mod     0  ]
```

2. Derived type: from S->E . Any subject NP. VP head is past tense, regardless of the viewpoint, no modals, and the presence of a temporal CE <u>inside</u> the VP node

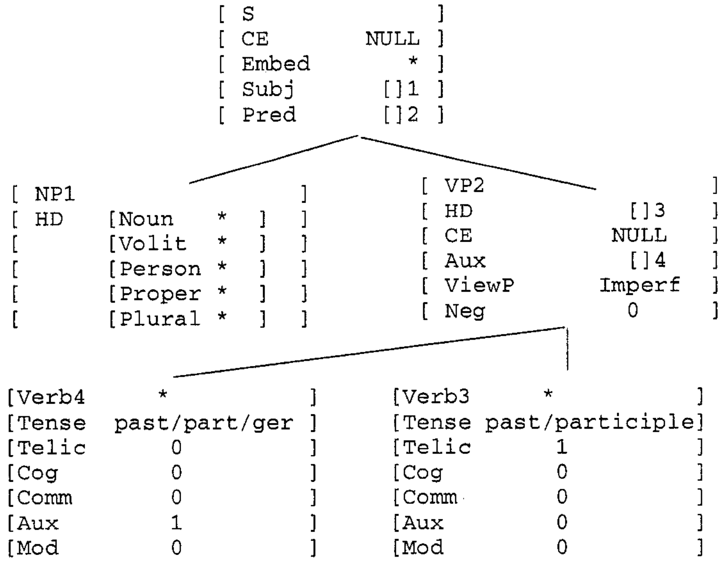


3. Any subject NP. VP head tense past or participle. Perfective viewpoint, no CE's, VP head not modal, VP not negated, main verb telic.

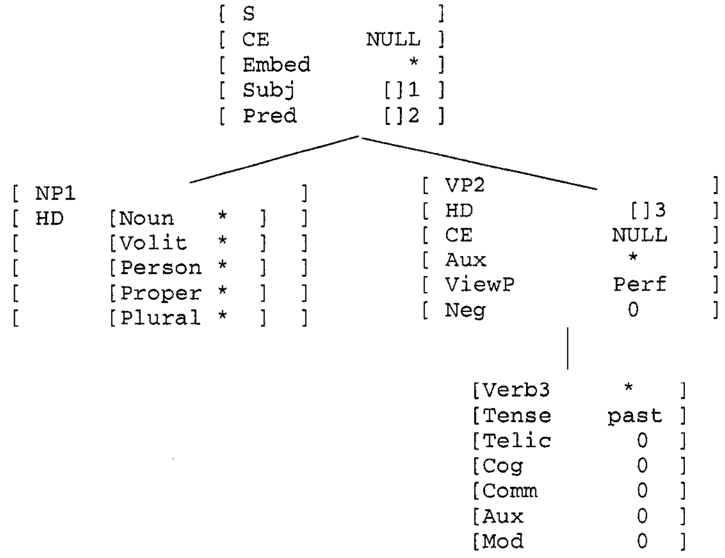


States:

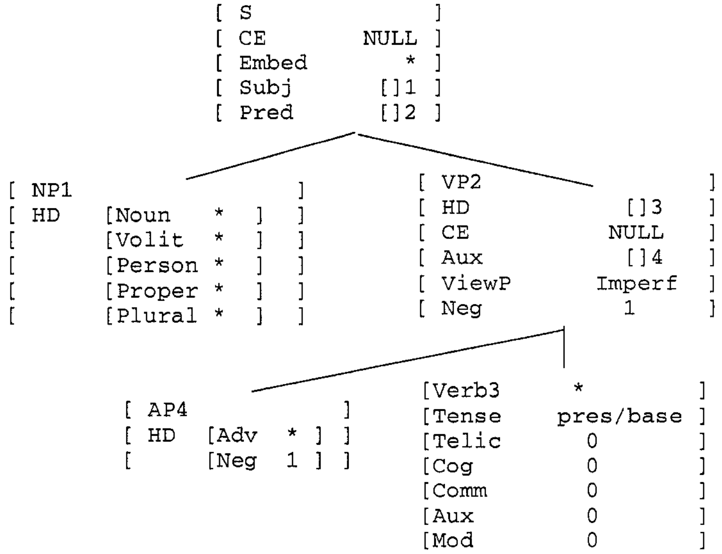
1. Any subject NP. VP head past or participle tense, VP imperfective viewpoint, no CE's, VP not modal, VP not negated, VP has auxiliary, but it is not in present tense or base form



2. Any subject NP. VP head past tense, perfective viewpoint, no CE's, VP has no modal, VP not negated, main verb is atelic (cognitives, perceptuals)

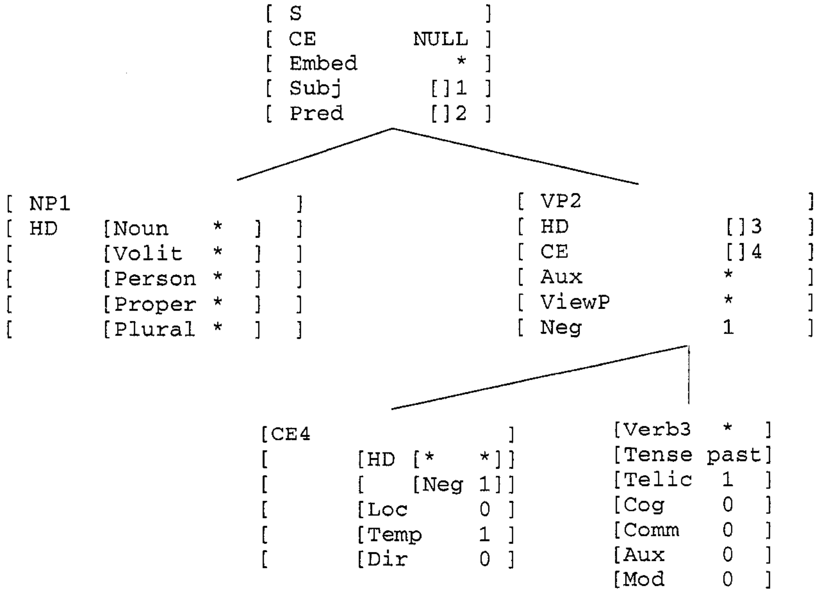


3. Derived type: from E -> S. Any subject NP. VP present/base form, imperfective viewpoint, no CE's, VP not modal, VP negated



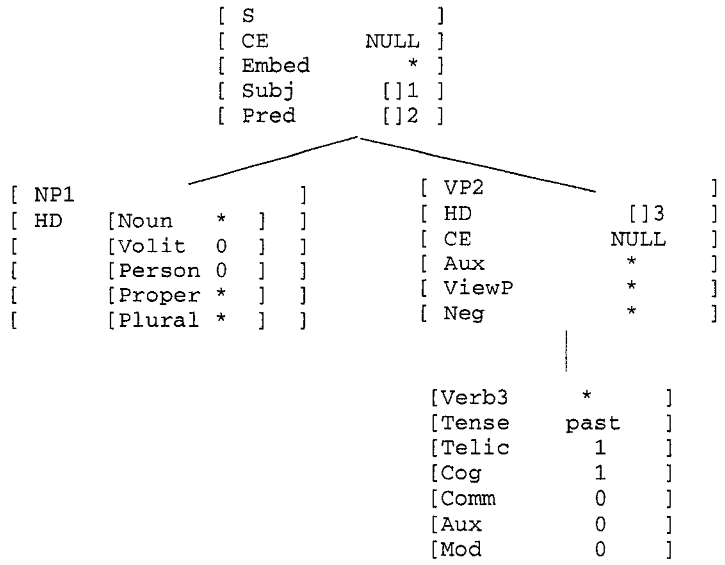
```
                          [ S               ]
                          [ CE        NULL  ]
                          [ Embed        *  ]
                          [ Subj       []1  ]
                          [ Pred       []2  ]

[ NP1                       ]    [ VP2                     ]
[ HD   [Noun    *   ]  ]         [ HD            []3       ]
[      [Volit   *   ]  ]         [ CE            NULL      ]
[      [Person  *   ]  ]         [ Aux           []4       ]
[      [Proper  *   ]  ]         [ ViewP         Imperf    ]
[      [Plural  *   ]  ]         [ Neg           1         ]

                                 [Verb3      *          ]
         [ AP4             ]     [Tense      pres/base  ]
         [ HD  [Adv  * ] ]       [Telic      0          ]
         [     [Neg  1 ] ]       [Cog        0          ]
                                 [Comm       0          ]
                                 [Aux        0          ]
                                 [Mod        0          ]
```

4. Derived type: fom E -> S. Corresponds to a derived Event type 1 (above). Any subject NP. The negation happens through a CE in the VP, with any viewpoint



```
                              [ S               ]
                              [ CE        NULL  ]
                              [ Embed        *  ]
                              [ Subj       []1  ]
                              [ Pred       []2  ]

[ NP1                       ]          [ VP2                  ]
[ HD   [Noun    *   ]  ]               [ HD            []3    ]
[      [Volit   *   ]  ]               [ CE            []4    ]
[      [Person  *   ]  ]               [ Aux           *      ]
[      [Proper  *   ]  ]               [ ViewP         *      ]
[      [Plural  *   ]  ]               [ Neg           1      ]

                    [CE4                 ]     [Verb3   *   ]
                    [       [HD [*    *]]      [Tense  past]
                    [       [   [Neg 1]]       [Telic   1  ]
                    [       [Loc      0 ]      [Cog     0  ]
                    [       [Temp     1 ]      [Comm    0  ]
                    [       [Dir      0 ]      [Aux     0  ]
                                               [Mod     0  ]
```

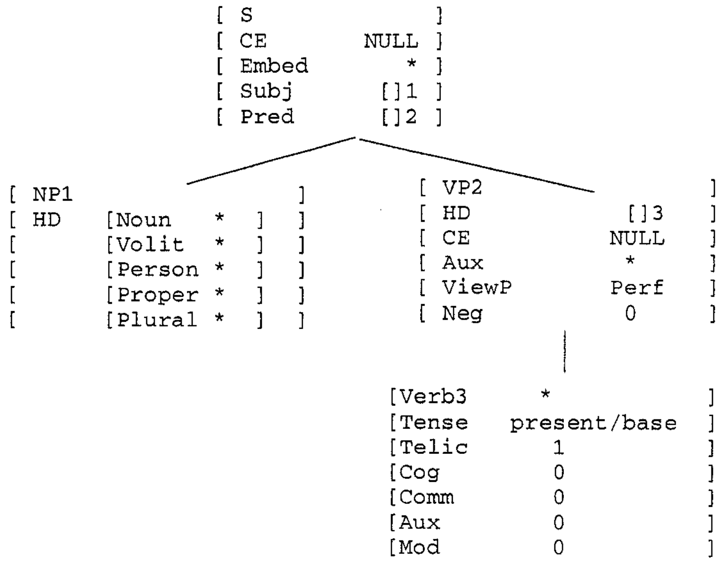
5. Derived type: from E -> S. Corresponds to a derived Event of any type. Non-volitional subject NP. VP with any viewpoint, past tense, cognitive verb



```
                  [ S               ]
                  [ CE         NULL ]
                  [ Embed         * ]
                  [ Subj        []1 ]
                  [ Pred        []2 ]

[ NP1                        ]    [ VP2                    ]
[ HD    [Noun     *   ]      ]    [ HD             []3     ]
[       [Volit    0   ]      ]    [ CE             NULL    ]
[       [Person   0   ]      ]    [ Aux             *      ]
[       [Proper   *   ]      ]    [ ViewP           *      ]
[       [Plural   *   ]      ]    [ Neg             *      ]

                                   [Verb3      *       ]
                                   [Tense     past     ]
                                   [Telic      1       ]
                                   [Cog        1       ]
                                   [Comm       0       ]
                                   [Aux        0       ]
                                   [Mod        0       ]
```

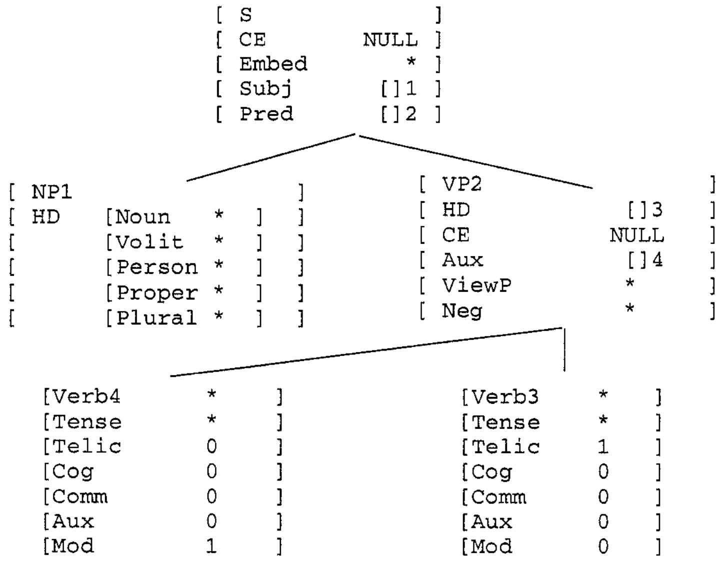
Assessments:

1. Any subject NP. VP head in present tense or base form, perfective viewpoint, no CE's



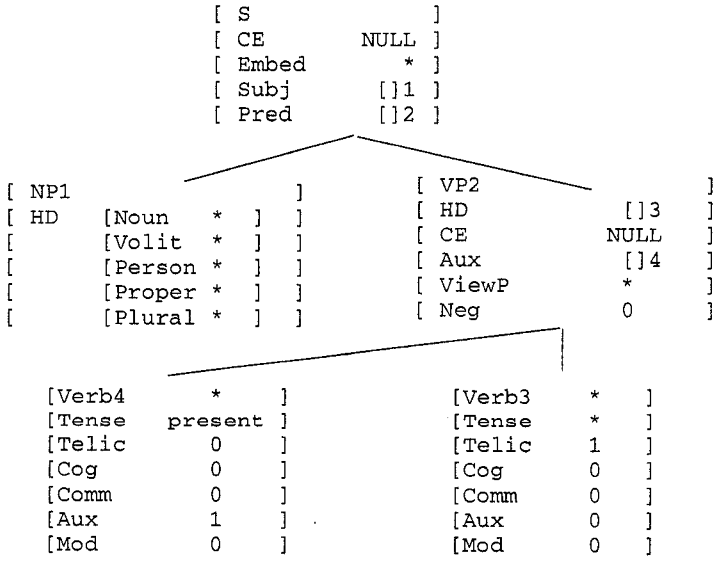
```
                  [ S               ]
                  [ CE         NULL ]
                  [ Embed         * ]
                  [ Subj        []1 ]
                  [ Pred        []2 ]

[ NP1                        ]    [ VP2                    ]
[ HD    [Noun     *   ]      ]    [ HD             []3     ]
[       [Volit    *   ]      ]    [ CE             NULL    ]
[       [Person   *   ]      ]    [ Aux             *      ]
[       [Proper   *   ]      ]    [ ViewP          Perf    ]
[       [Plural   *   ]      ]    [ Neg             0      ]

                                [Verb3      *                 ]
                                [Tense     present/base       ]
                                [Telic      1                 ]
                                [Cog        0                 ]
                                [Comm       0                 ]
                                [Aux        0                 ]
                                [Mod        0                 ]
```

2. Any subject NP. VP head in any tense, any viewpoint. VP has a modal in it



```
                    [ S                ]
                    [ CE          NULL ]
                    [ Embed          * ]
                    [ Subj         []1 ]
                    [ Pred         []2 ]

[ NP1                         ]        [ VP2                         ]
[ HD    [Noun    *   ]   ]             [ HD               []3    ]
[       [Volit   *   ]   ]             [ CE              NULL    ]
[       [Person  *   ]   ]             [ Aux              []4    ]
[       [Proper  *   ]   ]             [ ViewP             *     ]
[       [Plural  *   ]   ]             [ Neg               *     ]

  [Verb4       *     ]                     [Verb3     *    ]
  [Tense       *     ]                     [Tense     *    ]
  [Telic       0     ]                     [Telic     1    ]
  [Cog         0     ]                     [Cog       0    ]
  [Comm        0     ]                     [Comm      0    ]
  [Aux         0     ]                     [Aux       0    ]
  [Mod         1     ]                     [Mod       0    ]
```

3. Any subject NP. VP with any tense for the main verb. Auxiliary in present tense, VP with any viewpoint



```
                    [ S                ]
                    [ CE          NULL ]
                    [ Embed          * ]
                    [ Subj         []1 ]
                    [ Pred         []2 ]

[ NP1                         ]        [ VP2                         ]
[ HD    [Noun    *   ]   ]             [ HD               []3    ]
[       [Volit   *   ]   ]             [ CE              NULL    ]
[       [Person  *   ]   ]             [ Aux              []4    ]
[       [Proper  *   ]   ]             [ ViewP             *     ]
[       [Plural  *   ]   ]             [ Neg               0     ]

  [Verb4       *     ]                     [Verb3     *    ]
  [Tense    present  ]                     [Tense     *    ]
  [Telic       0     ]                     [Telic     1    ]
  [Cog         0     ]                     [Cog       0    ]
  [Comm        0     ]                     [Comm      0    ]
  [Aux         1     ]                     [Aux       0    ]
  [Mod         0     ]                     [Mod       0    ]
```

APPENDIX B. Data structures

The following diagrams are class diagrams produced in standard Unified Modeling Language (UML) notation. Class member functions are not shown for simplicity.

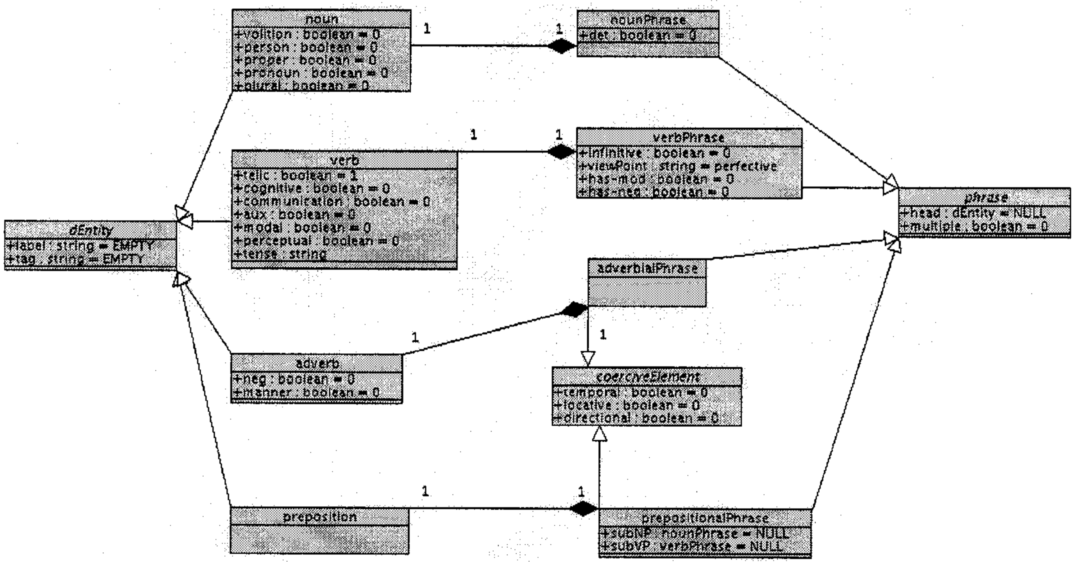


Illustration 24. Lexical Data Structures

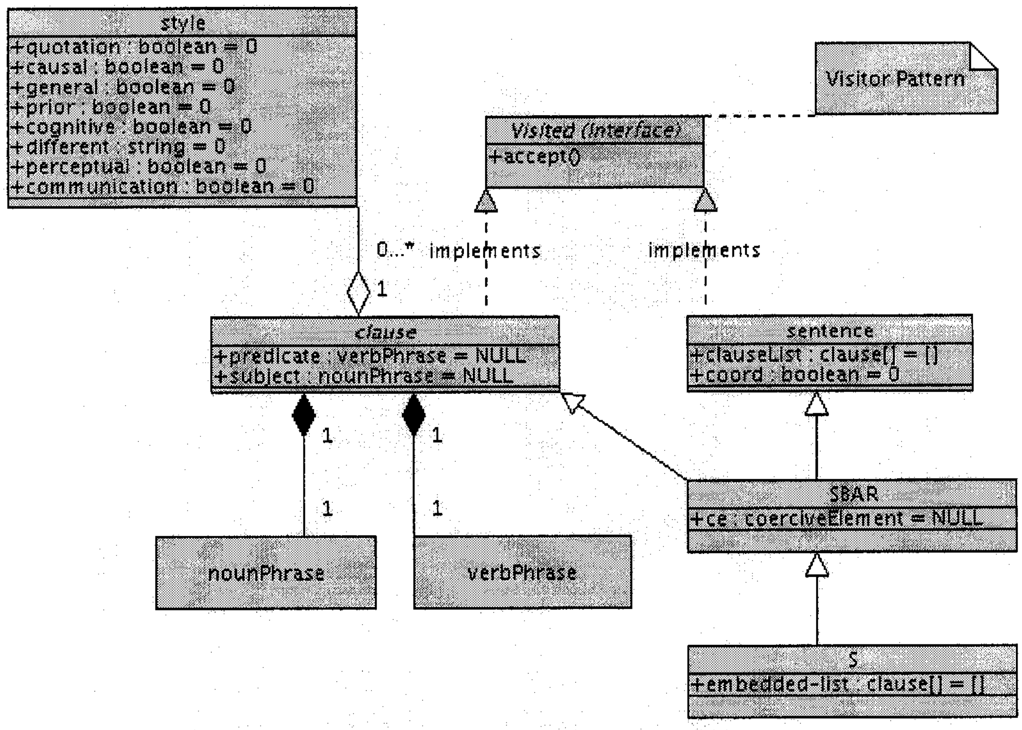


Illustration 25. Clausal Data Structures

APPENDIX C. Program architecture

The following diagrams are class diagrams produced in standard Unified Modeling Language (UML) notation. Class member functions are not shown for simplicity.

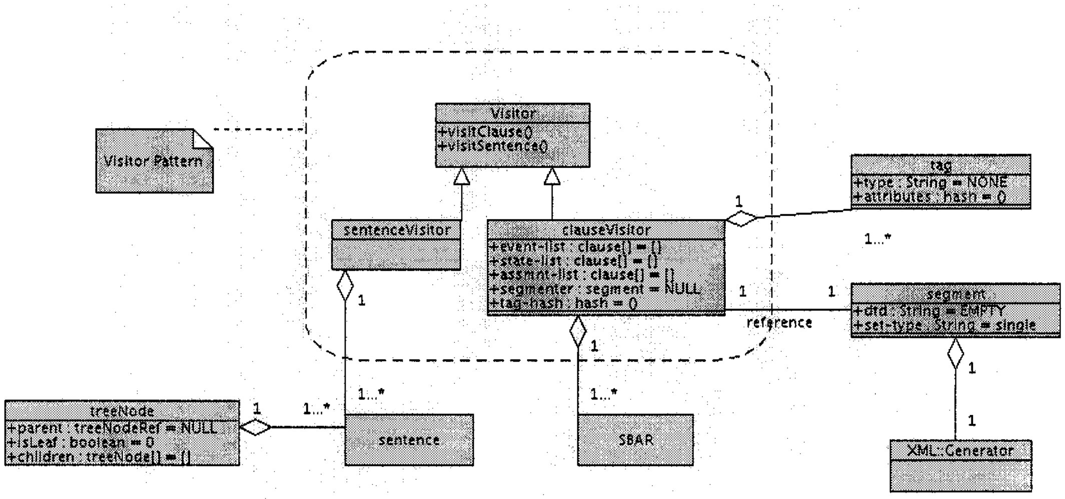


Illustration 26. Matching Algorithm Classes

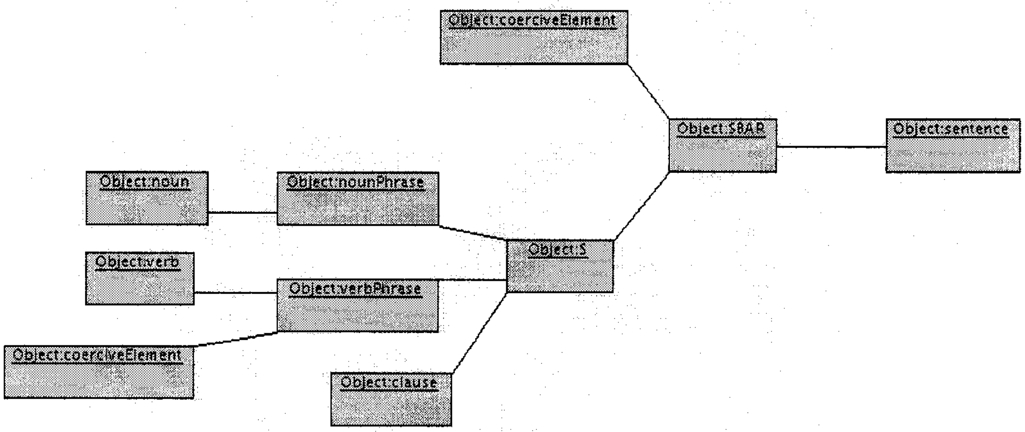


Illustration 27. Live Populated Object Tree

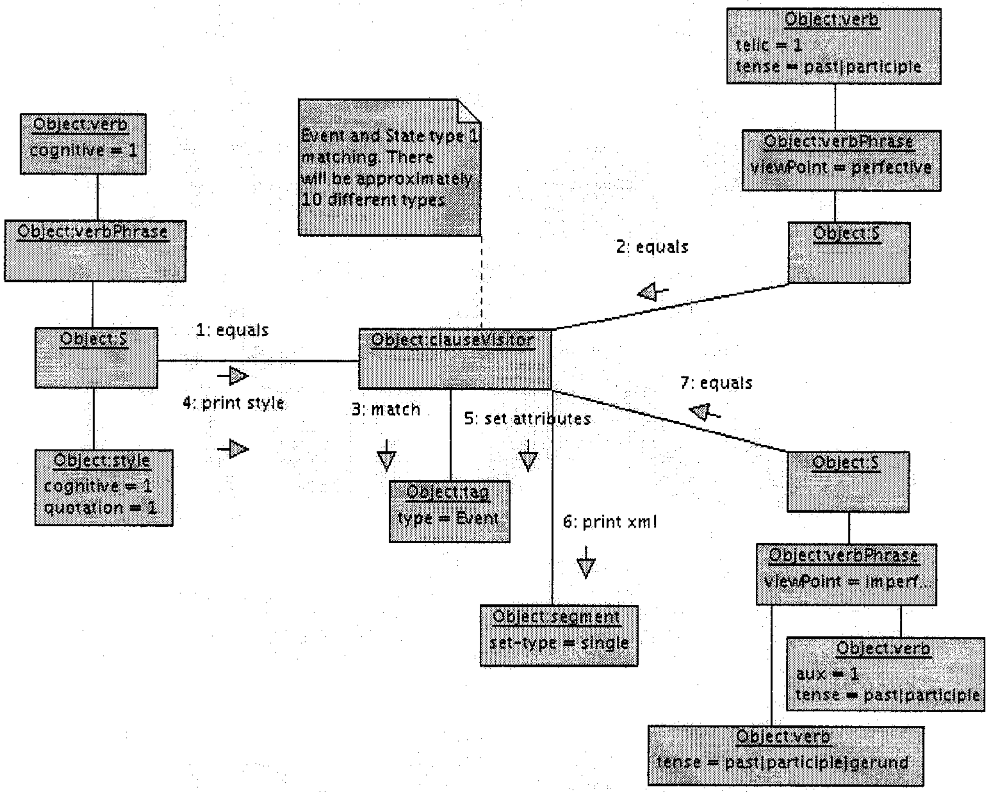


Illustration 28. Matching Algorithm

APPENDIX D. Software

- **Eric Brill's Part-of-Speech Tagger:** a state-of-the-art hybrid machine learning and statistical token annotator for natural language. Open Source. Available at http://research.microsoft.com/%7EBrill/. This tagger is the same employed for the widely used GATE architecture (Genera Architecture for Text Engineering), which is the software platform employed by Battelle Laboratories, the contractor for NASA under the AvSP project.

- **Michael Collins' Head-Driven Statistical Parser:** a world-class statistical syntax parser for natural language. Open Source. Available at: http://www.ai.mit.edu/people/mcollins/

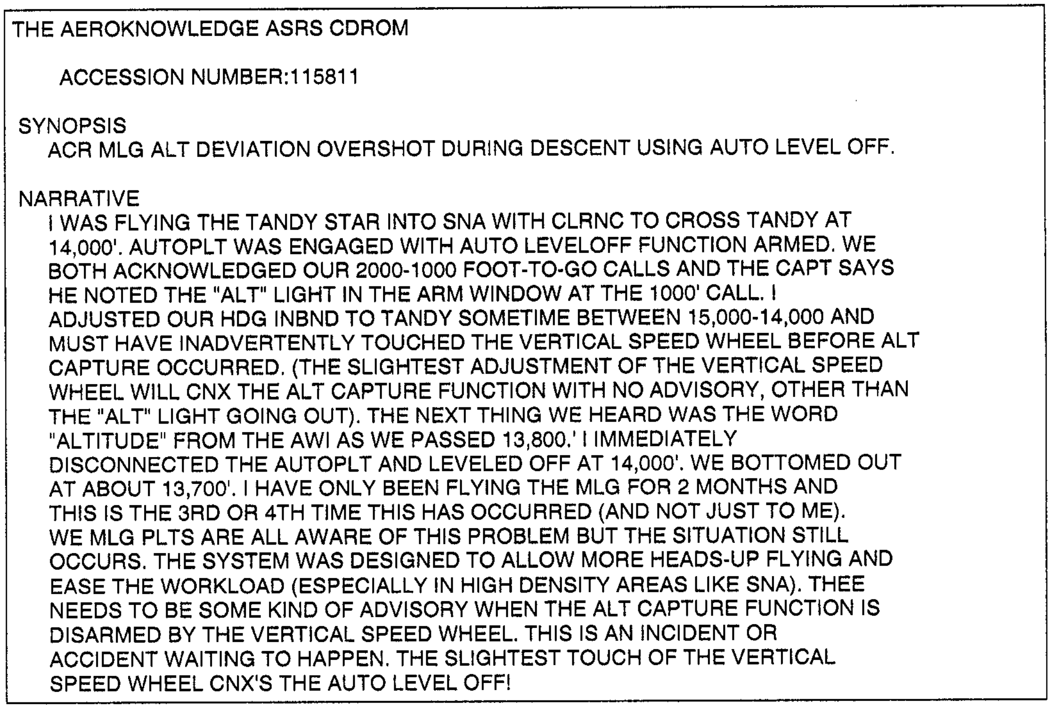
- **WordNet:** an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. Available at: http://www.cogsci.princeton.edu/~wn/

- **GNU Aspell:** a Open Source spell checker that does a much better job of coming up with possible suggestions than just about any other spell checker out there for the English language, including Ispell and Microsoft Word. Available at: http://aspell.sourceforge.net/

- **R:** developed by The R Foundation for Statistical Computing. A state-of-the-art statistical computing environment, freely distributed under a GNU General Public License. It is developed by a community of scientists users who contribute with free libraries of statistical methods. The program is an implementation of the 'S' language, developed by AT&T Bell labs.

APPENDIX E. Samples

Following is a sample report in the various stages of preprocessing.



THE AEROKNOWLEDGE ASRS CDROM

ACCESSION NUMBER:115811

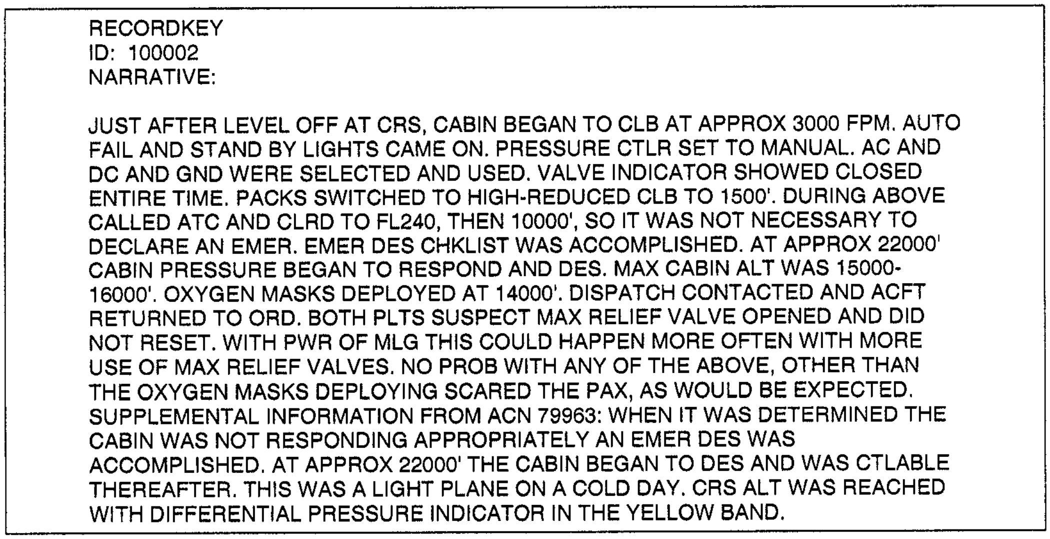
SYNOPSIS

ACR MLG ALT DEVIATION OVERSHOT DURING DESCENT USING AUTO LEVEL OFF.

NARRATIVE

I WAS FLYING THE TANDY STAR INTO SNA WITH CLRNC TO CROSS TANDY AT 14,000'. AUTOPLT WAS ENGAGED WITH AUTO LEVELOFF FUNCTION ARMED. WE BOTH ACKNOWLEDGED OUR 2000-1000 FOOT-TO-GO CALLS AND THE CAPT SAYS HE NOTED THE "ALT" LIGHT IN THE ARM WINDOW AT THE 1000' CALL. I ADJUSTED OUR HDG INBND TO TANDY SOMETIME BETWEEN 15,000-14,000 AND MUST HAVE INADVERTENTLY TOUCHED THE VERTICAL SPEED WHEEL BEFORE ALT CAPTURE OCCURRED. (THE SLIGHTEST ADJUSTMENT OF THE VERTICAL SPEED WHEEL WILL CNX THE ALT CAPTURE FUNCTION WITH NO ADVISORY, OTHER THAN THE "ALT" LIGHT GOING OUT). THE NEXT THING WE HEARD WAS THE WORD "ALTITUDE" FROM THE AWI AS WE PASSED 13,800.' I IMMEDIATELY DISCONNECTED THE AUTOPLT AND LEVELED OFF AT 14,000'. WE BOTTOMED OUT AT ABOUT 13,700'. I HAVE ONLY BEEN FLYING THE MLG FOR 2 MONTHS AND THIS IS THE 3RD OR 4TH TIME THIS HAS OCCURRED (AND NOT JUST TO ME). WE MLG PLTS ARE ALL AWARE OF THIS PROBLEM BUT THE SITUATION STILL OCCURS. THE SYSTEM WAS DESIGNED TO ALLOW MORE HEADS-UP FLYING AND EASE THE WORKLOAD (ESPECIALLY IN HIGH DENSITY AREAS LIKE SNA). THEE NEEDS TO BE SOME KIND OF ADVISORY WHEN THE ALT CAPTURE FUNCTION IS DISARMED BY THE VERTICAL SPEED WHEEL. THIS IS AN INCIDENT OR ACCIDENT WAITING TO HAPPEN. THE SLIGHTEST TOUCH OF THE VERTICAL SPEED WHEEL CNX'S THE AUTO LEVEL OFF!

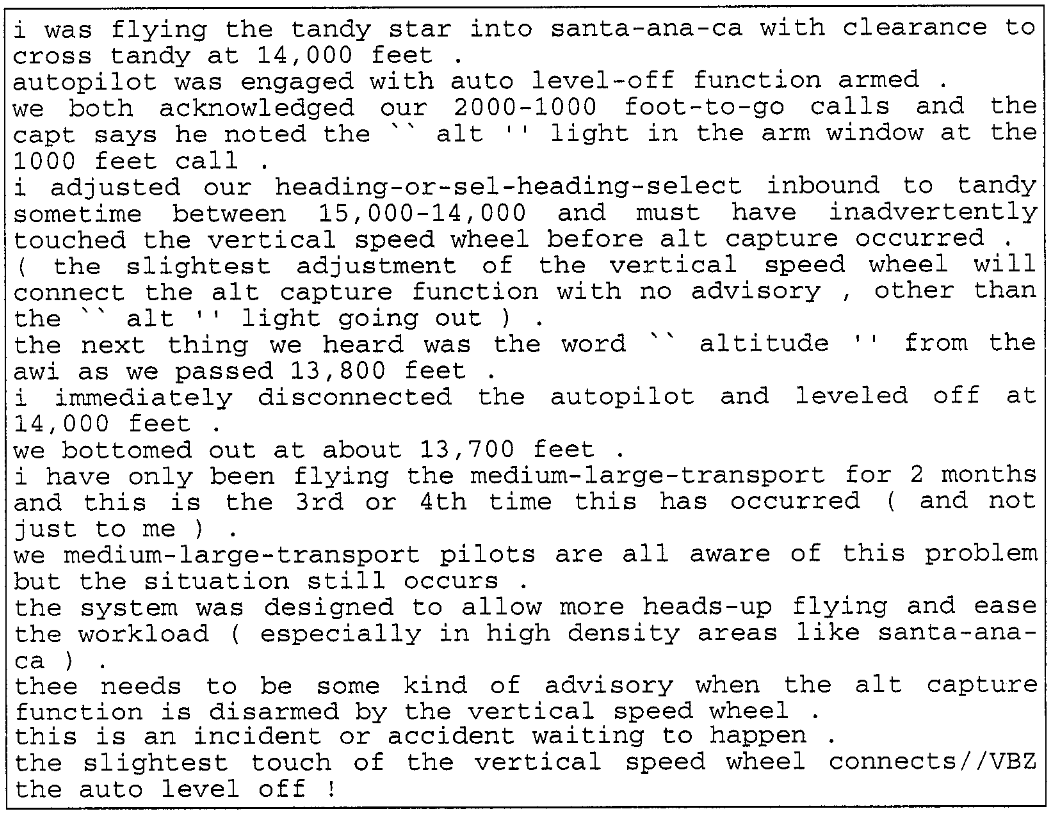
Table 13. Original Data from Aeroknowledge CDROM



RECORDKEY
ID: 100002
NARRATIVE:

JUST AFTER LEVEL OFF AT CRS, CABIN BEGAN TO CLB AT APPROX 3000 FPM. AUTO FAIL AND STAND BY LIGHTS CAME ON. PRESSURE CTLR SET TO MANUAL. AC AND DC AND GND WERE SELECTED AND USED. VALVE INDICATOR SHOWED CLOSED ENTIRE TIME. PACKS SWITCHED TO HIGH-REDUCED CLB TO 1500'. DURING ABOVE CALLED ATC AND CLRD TO FL240, THEN 10000', SO IT WAS NOT NECESSARY TO DECLARE AN EMER. EMER DES CHKLIST WAS ACCOMPLISHED. AT APPROX 22000' CABIN PRESSURE BEGAN TO RESPOND AND DES. MAX CABIN ALT WAS 15000-16000'. OXYGEN MASKS DEPLOYED AT 14000'. DISPATCH CONTACTED AND ACFT RETURNED TO ORD. BOTH PLTS SUSPECT MAX RELIEF VALVE OPENED AND DID NOT RESET. WITH PWR OF MLG THIS COULD HAPPEN MORE OFTEN WITH MORE USE OF MAX RELIEF VALVES. NO PROB WITH ANY OF THE ABOVE, OTHER THAN THE OXYGEN MASKS DEPLOYING SCARED THE PAX, AS WOULD BE EXPECTED. SUPPLEMENTAL INFORMATION FROM ACN 79963: WHEN IT WAS DETERMINED THE CABIN WAS NOT RESPONDING APPROPRIATELY AN EMER DES WAS ACCOMPLISHED. AT APPROX 22000' THE CABIN BEGAN TO DES AND WAS CTLABLE THEREAFTER. THIS WAS A LIGHT PLANE ON A COLD DAY. CRS ALT WAS REACHED WITH DIFFERENTIAL PRESSURE INDICATOR IN THE YELLOW BAND.

Table 14. Original Data from Battelle-PNWD



```
i was flying the tandy star into santa-ana-ca with clearance to
cross tandy at 14,000 feet .
autopilot was engaged with auto level-off function armed .
we both acknowledged our 2000-1000 foot-to-go calls and the
capt says he noted the `` alt '' light in the arm window at the
1000 feet call .
i adjusted our heading-or-sel-heading-select inbound to tandy
sometime between 15,000-14,000 and must have inadvertently
touched the vertical speed wheel before alt capture occurred .
( the slightest adjustment of the vertical speed wheel will
connect the alt capture function with no advisory , other than
the `` alt '' light going out ) .
the next thing we heard was the word `` altitude '' from the
awi as we passed 13,800 feet .
i immediately disconnected the autopilot and leveled off at
14,000 feet .
we bottomed out at about 13,700 feet .
i have only been flying the medium-large-transport for 2 months
and this is the 3rd or 4th time this has occurred ( and not
just to me ) .
we medium-large-transport pilots are all aware of this problem
but the situation still occurs .
the system was designed to allow more heads-up flying and ease
the workload ( especially in high density areas like santa-ana-
ca ) .
thee needs to be some kind of advisory when the alt capture
function is disarmed by the vertical speed wheel .
this is an incident or accident waiting to happen .
the slightest touch of the vertical speed wheel connects//VBZ
the auto level off !
```

Table 15. Sample Tokenized and Dis-abbreviated Data

```
i/PRP  was/VBD  flying/VBG  the/DT  tandy/NNP  star/NN  into/IN
santa-ana-ca/NNP with/IN clearance/NN to/TO cross/VB tandy/NNP
at/IN 14,000/CD feet/NNS ./.
autopilot/NN was/VBD engaged/VBN with/IN auto/NN level-off/JJ
function/NN armed/VBN ./.
we/PRP both/DT acknowledged/VBD our/PRP$ 2000-1000/JJ foot-to-
go/JJ  calls/NNS  and/CC  the/DT  captain/NN  says/VBZ  he/PRP
noted/VBD  the/DT  ``/:  altitude-or-alternate-or-hold-altitude-
hold-mode/NN ''/'' light/NN in/IN the/DT arm/NN window/NN at/IN
the/DT 1000/CD feet/NNS call/VBP ./.
i/PRP  adjusted/VBD  our/PRP$  heading-or-sel-heading-select/NN
inbound/JJ  to/TO  tandy/NNP  sometime/RB  between/IN  15,000-
14,000/JJ and/CC must/MD have/VB inadvertently/RB touched/VBN
the/DT  vertical/JJ  speed/NN  wheel/NN  before/IN  altitude-or-
alternate-or-hold-altitude-hold-mode/NN capture/NN occurred/VBD
./.
(/( the/DT slightest/JJS adjustment/NN of/IN the/DT vertical/JJ
speed/NN  wheel/NN  will/MD  connect/VB  the/DT  altitude-or-
alternate-or-hold-altitude-hold-mode/NN capture/NN function/NN
with/IN  no/DT  advisory/NN  ,/,  other/JJ  than/IN  the/DT  ``/:
altitude-or-alternate-or-hold-altitude-hold-mode/NN       ''/''
light/NN going/VBG out/IN )/)
the/DT next/JJ thing/NN we/PRP heard/VBD was/VBD the/DT word/NN
``/:  altitude/NN  ''/''  from/IN  the/DT  awi/RB  as/IN  we/PRP
passed/VBD     13,800/CD     feet/NNS     i/PRP     immediately/RB
disconnected/VBN the/DT autopilot/NN and/CC leveled/VBD off/IN
at/IN 14,000/CD feet/NNS ./.
```

Table 16. Sample Tagged Data

```
(TOP (S (NP (NP (PRP i))) (VP (VBD was) (VP (VBG flying) (NP
(NP (DT the) (NNP tandy) (NN star))) (PP (IN into) (NP (NP (NNP
santa-ana-ca)))) (PP (IN with) (NP (NP (NN clearance)) (S (VP
(TO to) (VP (VB cross) (NP (NP (NNP tandy))) (PP (IN at) (NP
(NP (CD 14,000) (NNS feet)))))))))))))
(TOP  (S  (NP  (NP  (NN  autopilot)))  (VP  (VBD  was)  (VP  (VBN
engaged) (PP (IN with) (NP (NP (NN auto) (JJ level-off) (NN
function)) (VP (VBN armed))))))))
(TOP  (S  (S  (NP  (NP  (PRP  we))  (NP  (NP  (DT  both))))  (VP  (VBD
acknowledged) (NP (NP (PRP$ our) (JJ 2000-1000) (JJ foot-to-go)
(NNS calls))))) (CC and) (S (NP (NP (DT the) (NN captain))) (VP
(VBZ says) (SBAR (S (NP (NP (PRP he))) (VP (VBD noted) (SBAR (S
(NP (NP (DT the) (: ``))) (NP (NP (NN altitude-or-alternate-or-
hold-altitude-hold-mode) ('' '') (NN light)) (PP (IN in) (NP
(NP (DT the) (NN arm) (NN window)) (PP (IN at) (NP (NP (DT the)
(CD 1000) (NNS feet))))))) (VP (VBP call))))))))))))
(TOP (S (NP (NP (PRP i))) (VP (VP (VBD adjusted) (NP (NP (PRP$
our)  (NN  heading-or-sel-heading-select)))  (ADJP  (JJ  inbound)
(PP (TO to) (NP (NP (NNP tandy) (RB sometime) (IN between))))
(JJ 15,000-14,000))) (CC and) (VP (MD must) (VP (VB have) (VP
(ADVP (RB inadvertently)) (VBN touched) (NP (NP (DT the) (JJ
vertical) (NN speed) (NN wheel)))) (SBAR (IN before) (S (NP (NP
(NN     altitude-or-alternate-or-hold-altitude-hold-mode)     (NN
capture))) (VP (VBD occurred)))))))))
```

Table 17. Sample Parsed Data

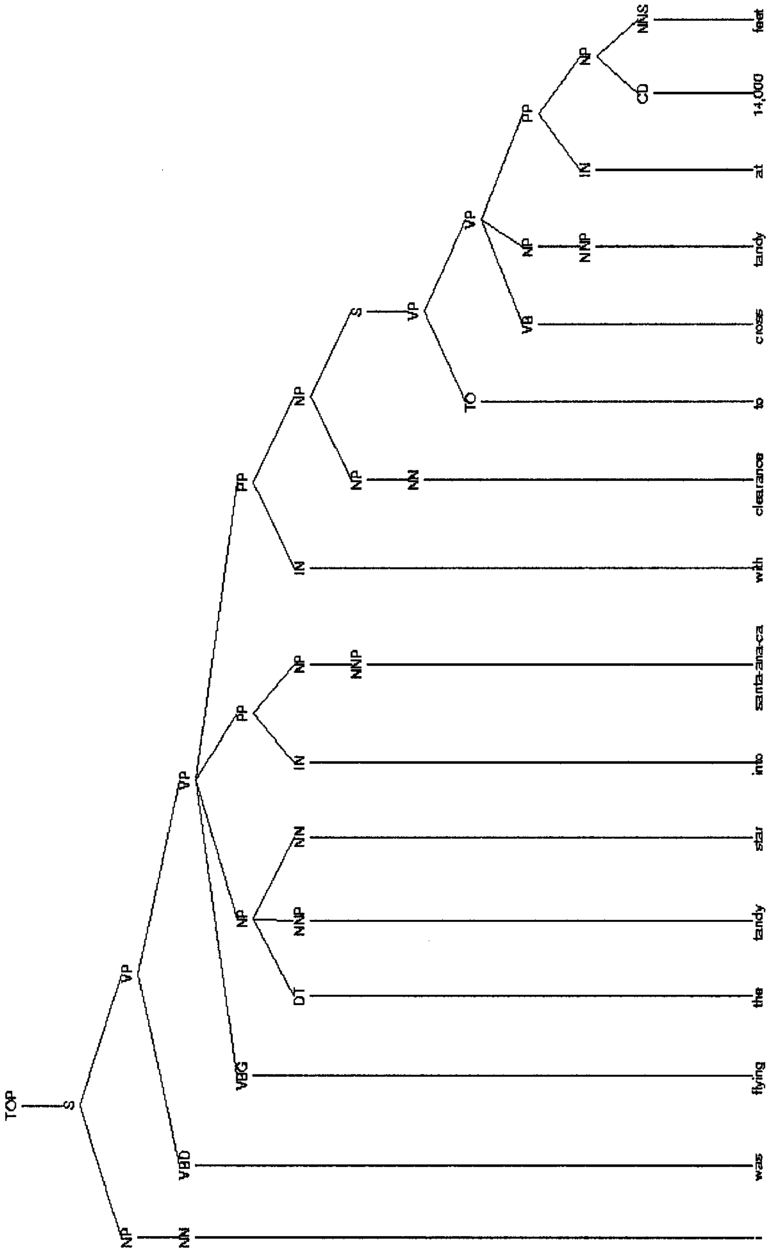


Illustration 29. Sample Syntactic Tree

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE ReportSet [
<!ELEMENT ReportSet (ReportDescription, Report+)>
 <!ATTLIST ReportSet Name CDATA #REQUIRED>
   <!ELEMENT ReportDescription (SectionDescription+)>
     <!ELEMENT SectionDescription EMPTY>
      <!ATTLIST SectionDescription Name CDATA #REQUIRED
                                 Type
(unique_id|category|data|text) #REQUIRED>
  <!ELEMENT Report (Section+)>
    <!ELEMENT Section (#PCDATA|St|S|E|A|X)*>
     <!ATTLIST Section Name CDATA #REQUIRED>
      <!ELEMENT St (#PCDATA)>
      <!ELEMENT S (#PCDATA)>
      <!ELEMENT E (#PCDATA)>
      <!ELEMENT A (#PCDATA)>
      <!ELEMENT X (#PCDATA)>
]>
<ReportSet Name="Development">
<ReportDescription>
<SectionDescription Type="unique_id" Name="ID" />
<SectionDescription Type="text" Name="NARRATIVE" />
</ReportDescription>
<Report>
  <Section Name="ID">115811</Section>
  <Section Name="NARRATIVE">
  <St>i was flying the tandy star into santa-ana-ca with
clearance to cross tandy at 14,000 feet </St>
  <St>autopilot was engaged with auto level-off function armed
</St>
  <A>the captain says he noted the `` altitude-or-alternate-
or-hold-altitude-hold-mode '' light in the arm window at the
1000 feet call </A>
  <E>i adjusted our heading-or-sel-heading-select  inbound to
tandy sometime between 15,000-14,000 </E>
  and
  <E>must have inadvertently touched the vertical speed wheel
before altitude-or-alternate-or-hold-altitude-hold-mode
capture occurred </E>
  <S>the next thing we heard was the word `` altitude '' from
the awi as we passed 13,800 feet i immediately disconnected
the autopilot and leveled  off at 14,000 feet </S>
  <E>we bottomed  out at  about 13,700 feet </E>
  <S>the system was designed to allow  more heads-up flying
and ease the workload -LRB- especially in high density areas
like santa-ana-ca </S>
  <A>there needs to be some kind of advisory when the
altitude-or-alternate-or-hold-altitude-hold-mode capture
function is  disarmed by the vertical speed wheel </A>
  <A>the slightest touch of the vertical speed wheel connect
's the auto level off </A>
  </Section>
</Report>
<Report>
</ReportSet>
```

Table 18. Sample XML-Tagged Output

APPENDIX F. Penn Treebank Tag Set

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb