

**FACILITATING READING THROUGH A THEME-DRIVEN  
APPROACH**

A Dissertation

by

JIE DENG

Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2008

Major Subject: Computer Science

**FACILITATING READING THROUGH A THEME-DRIVEN  
APPROACH**

A Dissertation

by

JIE DENG

Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved:

Chair of Committee,	Richard Furuta
Committee Members,	John J. Leggett
	Frank M. Shipman, III
	Eduardo Urbina
Head of Department,	Valerie E. Taylor

December 2008

Major Subject: Computer Science

## ABSTRACT

Facilitating Reading through a Theme-Driven Approach. (December 2008)

Jie Deng, B.S., Tsinghua University, China;

M.S., University of Nebraska at Omaha

Chair of Advisory Committee: Dr. Richard Furuta

Readers often encounter the need to explore a document only for a specific point of interest. We call the phenomena of approaching a narrative not for its entirety, but for a thread of a particular topic, thematic reading. Present reading tools and information retrieval techniques provide only limited assistance to readers in such a situation. Our research centers on this phenomenon. We conducted investigations on both human behavior and machine automation, with a goal of better meeting the requirements of thematic reading.

To observe readers' behavior and understand their expectations, we implemented a reader's interface with designs targeting the predicted needs of thematic readers. We conducted user studies using both the system and Microsoft Word. We proved that thematic reading is capable of achieving the goal of understanding a specific topic, at least to a degree that succeeds in topic-wise tasks. We also reached guidelines for designing future reading platforms in major aspects such as view, navigation, and contextual awareness.

As for machine automation, we investigated the potential to automatically locate thematically relevant excerpts. This investigation was inspired by the editorial compilation of a textbook index. To increase the search performance, we proposed a two-step methodology which first expands the query with expansion and then filters the intermediate results by checking the term-occurrence proximity. For query expansion, we compared the query expansion with WordNet, morphological inflections, and both processes together. Our results show that in the context of our study, WordNet made almost no contribution to the enhancement of recall, while expansion with the inflectional variants turned out to be a successful and essential scheme. For the refinement section, the results show that the proximity check on the alternative phrases formed after inflectional expansion can effectively increase the precision of the previously acquired return results.

We further tested a different scheme – using sliding window – of defining target and verification units in the methodology. Our findings show that the structural delimitations (sentences and chapters) outperformed sliding windows. The first scheme was able to achieve consistently desirable results, while the results from the second were inconclusive.

## **DEDICATION**

To my family

## ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Richard Furuta, for all his help and input that led to the completion of this dissertation. Without his guidance and encouragement, it would not have been possible for me to travel this far on the journey of pursuing the PhD. With a passion for knowledge and an open mind that shows extensive interest in a variety of different fields, he has set a wonderful example of appreciating the joy of learning and of having the courage to attempt things new and unfamiliar. Most noticeably, his disposition of organizing and relating information systematically also has shown me the value of thinking and expressing ideas both logically and efficiently. Beyond his academic merit, his considerateness and flexibility towards the needs of his individual students is another great quality to emulate. I would like to thank him sincerely for all his support throughout the years of my PhD program.

I would also like to recognize my committee. I would like to thank Dr. Eduardo Urbina for his valuable input and discussions that have helped clarify many issues in my dissertation. I am grateful for his time and effort. I am also grateful for the opportunity he has given me to work on the Cervantes Project. The project offered me the stage I needed to explore, attempt, and apply the technologies developed, in their various and varied aspects. It was also with this project that I formed the idea for this dissertation and gained numerous essential insights. I would like to thank Dr. Shipman and Dr. Leggett for giving me valuable suggestions during the preliminary examination. More

importantly, I would also like to thank them for being forces behind the wonderful collaborative atmosphere in the lab, from which I have certainly benefited a great deal.

I would like to thank the people from the Cervantes Project. It was a great experience working with them. I would also like to thank all the subjects that have participated in my user study. Their input provides the essential base of this dissertation.

I would like to thank all the graduate students from the Center for the Study of Digital Libraries. I feel extremely lucky to be pursuing my PhD with this group, and not anywhere else. It is their warm support and company that had brought so much joy to this lengthy journey. My sincere appreciation goes to each every single one of them. In particular, I would like to thank Carlos and Unmil for mentoring me not only with respect to academia, but also in views of life. I would like to thank Michael for his valuable suggestions in designing the user studies.

Lastly and most importantly, I would like to thank my dear family: my parents, my husband, my daughter, and my sisters. I cannot express enough appreciation for their unending love.

## TABLE OF CONTENTS

	Page
ABSTRACT .....	iii
DEDICATION .....	v
ACKNOWLEDGEMENTS .....	vi
TABLE OF CONTENTS.....	viii
LIST OF FIGURES .....	x
LIST OF TABLES.....	xi
1. INTRODUCTION .....	1
1.1 The Material.....	2
1.2 The Reader .....	3
1.3 The Behavior.....	3
1.4 The Challenge .....	6
1.5 The Value.....	7
2. RELATED WORK .....	10
2.1 Related Research Areas .....	10
2.1.1 Relevant Passage Retrieval .....	10
2.1.2 Index Manipulation.....	13
2.1.3 Highlighting.....	15
2.1.4 Others .....	16
2.2 Pertinent Concrete Techniques .....	17
2.2.1 WordNet.....	18
2.2.2 Term Variation .....	18
2.2.3 Combining Evidence at Different Levels.....	19
3. THE READER'S INTERFACE.....	20



	Page
3.1 Basic Functionalities .....	23
3.2 Interactive Functionalities .....	26
4. USER STUDY AND EVALUATION .....	33
4.1 Subject Pool .....	34
4.2 Experiment on Single Reading .....	34
4.2.1 Experimental Design.....	34
4.2.2 Discussion of the Phenomena of Thematic Reading .....	40
4.2.3 Discussion of the Guidelines for Immersive Design .....	47
4.3 Experiment on Multiple Readings .....	57
4.3.1 Experimental Design.....	57
4.3.2 Results and Discussion.....	59
5. AUTOMATIC PINPOINTING.....	64
5.1 Methodology .....	65
5.2 Tests under Structural Segmentation.....	69
5.2.1 Implementation of the System Components .....	70
5.2.2 Experimental Design.....	75
5.2.3 Results and Discussion.....	77
5.3 Tests on Segmentation and Proximity Tuning.....	87
5.3.1 Implementation of the System Components .....	88
5.3.2 Experimental Design.....	91
5.3.3 Results and Discussion.....	93
6. OPEN ISSUES AND FUTURE WORK .....	99
7. CONCLUSIONS .....	103
REFERENCES .....	108
VITA.....	114

## LIST OF FIGURES

		Page
Figure 1.	Example of using bookmarks in Microsoft Word .....	21
Figure 2.	The main window of the Reader's Interface .....	23
Figure 3.	An Open-File Dialogue for opening a saved copy of the book.....	24
Figure 4.	Peripheral information regarding the current copy.....	25
Figure 5.	An example of the highlighted view .....	27
Figure 6.	An example of the search operation.....	28
Figure 7.	An example list of suggested readings.....	29
Figure 8.	An individual suggested reading in Single View .....	30
Figure 9.	Two individual suggested readings in Shared View.....	31
Figure 10.	An example of the list of highlights.....	32
Figure 11.	Increase in the knowledge of the topics .....	46
Figure 12.	The overview of the methodology .....	67
Figure 13.	Some examples from the taxonomy.....	76
Figure 14.	F <sub>2</sub> measure of the different expansion methods.....	82
Figure 15.	F <sub>1</sub> measure of the refined search results from S4 after different refinement methods .....	84
Figure 16.	F <sub>2</sub> measure comparison of the suggested methods .....	87
Figure 17.	Sliding window demonstration when verification unit encloses 6 target units.....	89
Figure 18.	The two aspects involved in thematic reading .....	104

## LIST OF TABLES

		Page
Table I.	The subject pool of the user study.....	35
Table II.	Information about the topics .....	38
Table III.	Survey questions.....	39
Table IV.	Choice of reading when the subtask was conducted in the Reader's Interface .....	42
Table V.	Expert ratings on the highlighted excerpts for representing the topics.....	43
Table VI.	Subjects' understanding of the topics .....	46
Table VII.	Observations related to the view .....	49
Table VIII.	Rating the importance of a compact view of the selected text.....	49
Table IX.	Observations related to navigation when the subtask was conducted in the Reader's Interface .....	51
Table X.	Ratings about the importance of the bidirectional navigation .....	51
Table XI.	Observations related to the use of the search function .....	55
Table XII.	Responses to system preference.....	57
Table XIII.	Survey questions for the second scenario of the user study.....	58
Table XIV.	Behaviors exhibited when the collation view was not given to the subjects.....	60
Table XV.	Behaviors exhibited when the collation view was given to the subjects.....	62
Table XVI.	Results of the study .....	62

	Page
Table XVII. Search results of the different expansion methods.....	81
Table XVIII. Refined search results from S4 after different refinement methods.....	83
Table XIX. Final result comparison of the suggested methods.....	86
Table XX. Search results at different proximity values in R1 .....	94
Table XXI. Search results at different proximity values in R2 .....	95
Table XXII. Final results comparison between R1 and R2 .....	97

## 1. INTRODUCTION

Because people indulge in reading on a daily basis for various reasons such as recreation, study, or work, it is often unrecognized that reading is a complex process, more complex than the oft-assumed line-by-line scanning of a piece of text. If we take a few examples of reading under different circumstances, we can see that it is important first to clarify the “type” of reading within a given set of defining parameters. A museum attendee, standing in front of the famous portrait of Mona Lisa, reads the deliberation of the light and shading, and even the nuance of an emotion from the enigmatic smile. A psychological therapist, professionally directing the conversation with a patient, reads the experiences and feelings of the patient and a possible causal chain that leads to a troubled mindset. A college student, perusing the book of *The Old Man and the Sea*, reads about the perseverance of the old man on the lonely sea to bring back a big fish.

As we can see, reading is a very broad term, and hence is vague without certain efforts at confinement. Our study falls into the general scope of studies about “reading.” To identify the nature of reading that is under our investigation, we define the following three parameters: the material, the reader, and the behavior. Following that, we further discuss the observations that have motivated our work and the difficulties that readers face if no assistance is provided.

## 1.1 The Material

Even from the few examples given above, it is apparent that the object being read could appear in many forms, such as pictorial, textual, or even as as transcendental as thoughts. The material used in our study is uniform: text documents. Furthermore, since all our studies are carried out on a particular work of literature, the book of *Don Quixote*, for simplicity we refer in the context of this dissertation that the reading material is a self-contained text entity. Nevertheless, the ideas we propose here are not confined to a single book, and may easily extend to include multiple and heterogeneous text documents of different sources. The media chosen to present the text material is electronic. Texts in electronic form, in comparison to texts in the traditional paper form, are dynamic and fluid, which gives us the desired flexibility to regroup and reorganize its content easily.

It is worth mentioning that a substantial amount of research has been dedicated to the examination of the physical comparison of reading with regards to the two presentation media – the paper and the screen – inspecting factors like page size and resolution of the screen [Dillon 1994]. Our studies, however, don't fall into the general scope of investigating the physical affordance of the media and the effects introduced by them. Our attention, by comparison, is focused on the content that is presented through the media. The media, as the vehicle for carrying content, admittedly and inevitably has some effect on the reading process, but we try our best to focus not on the physical elements of the platform, but instead on the information layer on top of it. We emphasize the ways to explore the information, as opposed to the media, and conduct experiments

to see what applications of information technology may influence the exploration of the content at the user's will.

## **1.2 The Reader**

Reading of a text requires two levels of recognition: the level of word recognition, which deals with the transformation of visual images into perceived words or the extraction of meaning from structured prose [Rumelhart 1997]; and the level of content recognition, in which readers grasp the content of the text to fulfill the purposes of specific tasks. We target the second level reading, i.e., reading at a behavioral or task level, in our investigation. The readers are assumed to be advanced readers whose purpose in reading is to exploit the content of the text. Although the level of word recognition is an inevitable factor in all phases of text reading, the concentration of the reading in our studies is placed on the exploration of the content, instead of the decoding of the letter symbols.

## **1.3 The Behavior**

There are many ways and reasons that people read a document. O'Hara [O'Hara 1996] summarized the common reading goals referred to in the literature into a set of categories, including reading to learn, reading to search/answer a question, proof-reading, reading for enjoyment, and the like. In discussing how a text can be read, Lunzer [Lunzer 1979] distinguished between four different ways: receptive reading, reflective reading, skim reading, and scanning. One single reading task may involve many different ways of accessing the text. It would be impossible to create an exhaustive list of all the permutations of reading goals with the various ways to approach a text.

Among the many reading behaviors, our studies aim at one specific type, a type that Chorney termed “interactive reading” [Chorney 2005]: a process in which readers have control over the texts they read. In her vision, if there is a magic tool, which she refers to as a “reading wheel”, available to assist the reader to navigate the text at will, readers are encouraged to read not for a “linear narrative” but for points of interest, and are empowered to shape and control the reading process by selecting and reading only those parts of the text that are memorable or relevant to them. Interactive reading is similar to scanning, as defined by Lunzer, referring “more specifically...to locat[ing] a piece of information” [Lunzer 1979]. Interactive reading, or scanning, has been applied to many reading purposes, as described by O’Hara. For instance, reading to learn may require “a need to support...quick skim-based reviews of topics;” reading to search/answer questions involves “the location of relevant bits of information” [O’Hara 1996].

Using a specific focal point, which we generalize as a “theme,” as the pivot to direct reading is, in fact, prevalent and observable in the reading process. For instance, in literary reading, Iser [Iser 1972] pointed out, “the active interweaving of anticipation and retrospection” is one of the basic elements of the reading process. The reader constantly searches for the connections between the fragments. “With all literary text, then we may say that the reading process is selective, and the potential text is infinitely richer than any of its individual realizations.”

The concept of interactive reading or reading thematically is also well recognized by many other scholars, who have addressed the topic using various metaphors. In discussing the pragmatic and cognitive dimensions of literary reading, Vipond and Hunt



[Vipond and Hunt 1984] proposed three types: point-driven, where the reader is concerned to find out the overarching value or belief of the text; story-driven, where the readers emphasize plot, character, and events; and information-driven, where the text is surveyed for its information content. Both in story- and information-driven reading, readers “process discourse in units smaller than the entire text: most likely, in narrative episodes.” Given a “cut” text, the readers “tend to seek closure, and to reject disparate and seemingly unrelated text elements.” Bolter [Bolter 1991] stated that knowledge can be transmitted as “collections of ideas that can arrange themselves into a kaleidoscope of hierarchical and associative patterns – each pattern meeting the needs of one class of readers on one occasion.” Murray [Murray 1997] pointed out that with electronic text the “author” is procedural, like a choreographer “who supplies the rhythms, the context, and the set of steps that will be performed.” The reader, whom she called the “interactor,” is a “navigator, protagonist, explorer, or builder, [who] makes use of a repertoire of possible steps and rhythms to improvise a particular dance among the many, many possible dances the author has enabled.” From a more theoretical point of view, Lesgold and Perfetti [Lesgold and Perfetti 1981] referred to the cognition process during reading as a manifestation of episodic memory which can be thought of as a content-addressable trace of ongoing cognitive experience.

Consider, as an example, casual reading, what happens when a reader while perusing *Don Quixote* encounters Don Quixote promising an island to his squire. He starts to wonder what the outcome of the promise will be, yet the book, with multiple storylines carefully and craftily interwoven, diverts the reader’s attention to tell them

about something not directly related to the promise. The reader's urge to know the follow-on actions arouses a sense of impatience. He might skip through the text, searching for points where the suspense of the promise is readdressed. Likewise, scholarly reading also observes many instances of this scenario. A historian, interested in knowing more about a famous figure, studies the historical events associated with her in a chronologically written American history book. An architect, eager to form her own theory on the evolution of baroque design, scrutinizes all the instances where the topic appears in a book that describes well-known European architecture.

We summarize here that our target type of reading is a non-linear task level reading of electronic texts. The reading behavior deals specifically with exploring, selecting, and reading only those parts that are relevant to or coherent with a particular point of interest. The specific point of interest, which we generalize as a "theme," could have numerous possibilities: an episode, a character, an object, a relationship, and the like. The reading that is carried out surrounding the defined theme and the goal is an effort to select, cover, and connect the many discrete elements of the theme.

#### **1.4 The Challenge**

To fulfill the specific goals of thematic reading, readers actively search for the particular segments that are coherent to the driving focus. Some of the literary characteristics, though crucial to the beauty of the literariness, make it difficult to conduct such a search. Narrative, and in particular a literary narrative where art is dependent upon establishing distance, is full of "unexpected twists and turns, and frustration of expectations" [Iser 1972]. Throughout the story the narrator "directs the

reader's focus of attention to a changing array of topics, characters, and locations" [Bower and Morrow 1990]. Subjects under the examination of interactive readers are hence usually not readily available, and are likely to be scattered throughout the narrative. Readers would like to pay little attention to irrelevant information, but yet they constantly encounter it.

Locating the discontinuous pieces, with the support currently available, is somewhat tedious. The searching operation, commonly available in electronic texts, is helpful, but only to a limited extent. Such searching operations are typically implemented as string matching, and thus only able to find consecutive letter-by-letter matches. In the case of a textbook, the index in the back of the book is a resource precompiled to assist navigations of this sort. If a representative entry is available, the designated page numbers can direct a reader to a closer context. In some sense, such an index entry could be regarded as a pre-defined theme condensed into a short phrase or a word, while its page numbers, although imprecise and sometimes incomplete, are the information nuggets that are threaded together to form a composition specifically of that theme.

## **1.5 The Value**

Ideally, we can hope for a reading tool that automatically extracts for readers the text they desire, after they offer an initial input expressing their interests. If this magic tool exists, the product of the printing press will be radically expanded to provide infinite opportunities for nonlinear access to written ideas. The visions of many scholars would become a reality in our lives. The bare text would finally become a play script that the

reader uses, like a theater director, to construct in their imagination a full stage production [Bower and Morrow 1990]. The literary text would become a real collection of “stars.” The reader would use a magic pencil to draw a plough or a dipper. The innumerable variables would come right to our eyes [Iser 1972]. We would be granted the means for “disorderly reading, a reading practice of depth, rather than superficiality” [Brown 2007]. Readers, as imaginative weavers of textual fragments, would be able to conduct collative reading in terms of the equation  $1+1=3$  and come up with creative, unknown categories of thought. The interactors with the electronic narrative would be availed of a sea of dancing rhythms and would be able to improvise a variety of creative choreography [Murray 1997].

Nevertheless, a closer look at the presumption of this future tool makes it difficult to realize. The realization of the tool depends upon its capability of interpreting impeccably the reader’s intention. Different people might enter the same input to represent different ideas. Even when the ideas are largely the same, people may have different readings of the same subject. The final achievement of the tool requires continuous efforts that include, and are not limited to, the following research disciplines: cognitive science, literary theory and literature critics, information retrieval, and computer and human interaction. With the studies conducted in our work and the results reported here, we can only hope to cast a bit more light on the road forward.

This dissertation is roughly divided into two major parts, with foot-holds on the human-side and the machine-side, respectively. The first part is presented in Sections 3 and 4, the second in Section 5. The first part describes a demo reading system with

features designed specifically for testing the hypothesis related to the needs of thematic readers. Following that, we conduct user studies using our system and discuss our general observations on issues such as whether or not the readers are susceptible to the idea of thematic reading, what the key requirements are for an interface that caters this specific action, and how the tool might influence the readers' conception, if designed with features that favor a certain interpretation.

The thematic contents used in the studies of the first part are manually pre-marked. This, in turn, leads to the work of the second part, which heads toward the direction of automatically locating the relevant excerpts through the application of information retrieval. The work is inspired by the manual compilation of the index of a textbook, and investigates the potential of automatically finding the segments relevant to a search query. We hope to drive forward research in both automation and representation. The ultimate goal is an integration of the two sides, in which computation is able to extract a thread of theme effectively and the interface is conducive to the use and exploration of discrete themes. We report in this dissertation our endeavor toward such an ideal vision.

## 2. RELATED WORK

In Information Retrieval (IR), passage retrieval has long attracted a substantial amount of research effort, and remains to date an active research area. The goal of passage retrieval is to find those passages that are semantically similar to a given query from a collection of documents. Passage retrieval, document retrieval, and our excerpt retrieval indeed tackle the same extraction problem. Their distinction lies in how the “passage” is defined. If there were an extraction technique that worked ideally and were able to dynamically disintegrate a document into “passages” of proper size, the “reading wheel” as envisioned by Chorney [Chorney 2005] would have been part of our reality. Unfortunately, such a technique, as of now, does not exist. Nor is the general passage retrieval mature enough to handle support for reading thematically. Consequently, solutions are still under exploration in fine-grained areas, taking into consideration their specialized needs and conditions.

Section 2.1 broadly discusses the related work in various research areas. Section 2.2 describes prior work related to several concrete techniques involved in our study.

### 2.1 Related Research Areas

#### 2.1.1 *Relevant Passage Retrieval*

##### 2.1.1.1 Question-Answering Systems

Among the many IR specialty sub-areas, the research area closest to ours is passage retrieval in Question-Answering (QA) systems [Roberts and Gaizauskas 2004; Tellex et al. 2003]. QA systems take a question as an input, and return parts of the documents that may render a potential answer to the question. The current practice of

QA, as presented in the publications, typically tests the systems on standard test collections (for instance, TREC) and the results are presented as a ranked list of passages. Driven by distinctive needs, the differences between QA systems and our theme-driven reading system can be seen in the following three major aspects:

1) The test collection used. We prefer to use real books, as opposed to test collections, in our application, due to the fact that interactive reading is not merely a fact-searching task. Many occasions require it to go further, to the level of perceptual comprehension. It will be our ultimate interest to address questions like how reading thematically affects people's understanding of narrative, and whether new readings of the narrative may be initiated by approaching discrete themes instead of the more linear conventions. The human factor is core to addressing these questions. It is, therefore, recommended to have the tests conducted using real books with feedback from real subjects from the beginning of the study.

2) Ordering in the result set. Returns in QA extraction are ordered by their similarity rankings between the passages and the query. In thematic reading, another rule of ordering should also be available: the sequence of the excerpts that appears in the original text. If the needs of an interactive reader go beyond locating a simple fact (as is analogous to QA), to forming a complete sub-story, then weighing the excerpts according to similarity measures is no longer appropriate. All relevant excerpts contribute, evenly, to the forming of a theme. Furthermore, it is preferable to preserve the original order of their appearance such that the story flow is not destroyed and the author's view is accurately presented.

3) Evaluation emphasis. Performance evaluation for a QA system is primarily determined by the system's capability of including at least one correct answer in the top few returns. On the contrary, we are looking for a finite list which, ideally, contains all, and only, the relevant excerpts; the overall exhaustiveness and relevance of the result set defines our evaluation metrics.

#### 2.1.1.2 Passage Similarity Detection

Salton et al. [Salton et al. 1994; Salton et al. 1996] introduced the technique of the text relation map to automatically decompose texts into themes by merging triangles on a map. Its test on encyclopedia searching yielded high recall and good precision after supplementing searches through the text passages. The automatic detection of themes of this sort, although sharing the same sense that the text passages grouped are cohesive units, targets a different research question. The goal of the automatic detection of themes is to group texts into potential themes. In other words, the themes are unknown, and the process prepares texts for potential theme identification. In our research, the theme is pre-defined by the query, and very likely this theme scatters too sparsely in the text to be detectable by the automatic theme recognition. The automatic detection of themes focuses on the cohesiveness of the passage clusters, while our research emphasizes more the exhaustiveness of the coverage of a given theme. The difference between the natures of the two research areas makes it difficult to apply in our work the techniques for automatic theme detection.

Lyon et al. [Lyon et al. 2001] exploited the characteristic distribution of word trigrams (three words in succession) and used it to determine similarities between



passages. The technique has proven to be especially useful for plagiarism detection, where minor editing is conducted but similar phrases remain in the duplicates. As the technique relies heavily on the idiosyncrasy that is only available in almost identical passages, its usefulness to our work is limited. A thematic thread is unlikely to be found in identical pieces, but rather in evolving or related ones, which may be hard to detect using word trigrams.

### 2.1.1.3 Manual Authoring of Thematic Connections

Manually edited thematic connections are seen in the hypertextual view embedded in books like *The Thompson Chain-Reference Study Bible* [Thompson 2005]. Manually adding cross-references has the advantage of including more historical, interpretative elements into the authoring of relevance. On the other hand, the resulting link is somewhat subjective and is constrained by the interest, knowledge, and preference of the editor. Such constraints may in turn lead to a lack of comprehensiveness. Manual authoring of thematic connections is often seen to particularly emphasize one-to-one pair-wise linkages. Gathering the components associated with a certain subject still requires an extra step of congregating the related pairs. Our system, although not as powerful as a conceptual interpreter, aims to deliver a compact package of relevant chunks to one place.

### 2.1.2 *Index Manipulation*

Taking a Natural Language Processing approach, LinkIT [Evans 1998] is one of the tools implemented for automatic identification of significant topics in domain-independent full text. Its effectiveness for identifying representative phrases has been

recognized in research studies [Wacholder et al. 2001]. Two issues need to be considered if we incorporate automatic indexing to our system as a middle layer between user input and target locations. First of all, an effective methodology has to be in place to map the user input to the entries in the extracted list. Secondly, it is possible that such a mapping may not be established if the extracted list does not contain an entry to reflect the user input. When this happens, there is no benefit gained from the automatic indexing process. The system still requires a mechanism to go directly from the user query to the target destinations in the document. We thus have chosen not to adopt automatic indexing. In cases where index entries are available, as in a textbook, they could be incorporated as bonus knowledge. Such incorporation then involves the handling of the first issue above.

ScentIndex [Chi et al. 2004] contained experiments exactly in the domain of the first issue. Chi et al. studied the possibility of narrowing down a large index to conceptually match a user's interest. They proposed a function, which was computed based on word co-occurrence and information scent, to calculate the Degree of Interest for each entry in the original index. They then used the interest values to reduce and reorganize the index entries down to a single page for users to easily peruse.

The Hyper-TextBook project [Crestani and Melucci 2003] is another example of a study centered on the exploitation of an available index. Crestani and Melucci investigated the conversion of a textbook into a non-linear, hypertextual version by ranking and linking index terms and their associated pages. It shows some level of similarity to our project in the sense that it also aims to assist reading by inter-linking

thematically connected artifacts. The restructured textbook has outperformed the original as a better self-reference source. Nevertheless, due to the fact that it wholly relies on the resources available through its index entries, two problems arise: 1) the thematic linkage by the term-page authoring is not precise since the context in use—a page—is rather loose; 2) the technique is not applicable to narrative texts, such as novels, which do not typically have a precompiled index. Given these observations, we first break the physical constraints of page limitation and examine instead the similarity of small text units to a specified theme. More importantly, we do not rely on index entries to define a theme. Our work on the machine side presented in Section 5 is specifically aimed at lifting the requirement of an existing index.

### *2.1.3 Highlighting*

#### *2.1.3.1 ScentHighlights*

The ScentHighlights project [Chi et al. 2005] bears some similarity to our project in that both aim to assist reading by visually emphasizing semantically related text groups. ScentHighlights identifies the conceptual keywords that are highly relevant to the search terms, and colors those sentences that contain the keywords. Similar techniques are adopted in our methodology. Nevertheless, such a technique accounts only for the first step. A Boolean search for all sentences that contain one keyword doesn't necessarily guarantee a defining relationship to the query. We further propose a second step to shrink the returns of the Boolean search to a more precise pool. Furthermore, our research is designed to facilitate reading in a wider scope than ScentHighlights by considering additional features such as theme suggestion.

### 2.1.3.2 XLibris

XLibris [Schilit et al. 1998; Golovchinsky and Marshall 2000] bears a strong resemblance to our Reader's Interface in terms of adding highlights and listing those highlights. XLibris is designed to be a reading machine situated on a tablet meant to resemble a paper-like interface that allows, in particular, free-form ink annotations. In fact, XLibris's functionality is more extensive in terms of adding and displaying different kinds of annotations. Nevertheless, XLibris doesn't offer a way for the reader to enter arbitrary search terms. Thus, it is missing the first link that connects the following three steps: enter a query to express an interest, use the initial search results to mark and record the contents of interest, and recombine the elements to form a particular path of reading. Besides, there is no direct way in XLibris to compare the annotations – even in the same format – made to the same document by different people, or by the same person in different rounds of reading. The Reader's Interface contains a shared view of multiple copies, which is needed to carry out primitive studies of multiple readings.

### 2.1.4 *Others*

#### 2.1.4.1 Text Summarization

It is important to differentiate the theme-driven approach of reading from the research areas that use a hierarchical structure to provide non-linear hypertextual reading. The hierarchical methods focus on vending information with varied levels of granularity. For example, information presentation on small devices [Yang and Wang 2003; Patel and Marsden 2004; Björk et al. 1999; Yin and Lee 2004] has used a variety

of text summarization models to create different levels of information presentation. Zellweger et al. [Zellweger et al. 2002] generated abstracted views through carefully designing the tree leaves on a treetable with fine-grained information pieces. Regardless of the techniques they used, their hierarchical methods all tried to provide a global view of the underlying document. Our approach, by contrast, concentrates on the exhaustiveness and cohesion of the information delivery of one individual subtopic, which is a partial view. There is no text reduction involved once the text coherent to the theme is identified.

#### 2.1.4.2 Subtopic Boundary Detection

Another related research area is subtopic boundary detection. Recent research in this field, particularly the TextTiling algorithm [Hearst 1997; Hearst and Plaunt 1993], seems to show that a document's topic boundaries can be identified with a fair amount of success. The TextTiling algorithm tiles texts up into chunks of subtopics by using patterns of lexical connectivity to find coherent sub-discussions. Other boundary detection algorithms vary from TextTiling by their similarity indicators, but the basic idea stays the same: a topic change is indicated and identified by the valley on the similarity map. For example, Kozima [Kozima 1993] used the lexical cohesion profile as the similarity measure to detect topic switching points.

## 2.2 Pertinent Concrete Techniques

In summary, the methodology we propose in our system automation consists of two parts: query expansion and result refinement. In query expansion, we tested the expansion effectiveness with both WordNet synsets [Miller 1995] and morphological

variants. The refinement element filters the sentence-level returns with the evidence collected from the chapters. The methodology will be covered in more detail in Section 5. Below we briefly discuss the previous work relative to the three techniques involved: Wordnet, morphological variant expansion, and combining different levels of evidence in document retrieval.

### 2.2.1 *WordNet*

As a widely-adopted semantic thesaurus in linguistics computing and natural language processing, WordNet has attracted much interest in the past with attempts to incorporate semantic knowledge into IR systems. Its effectiveness, however, has received mixed reviews. There are cases where its adoption is encouraged; most noticeably, it has proven to be useful in disambiguating word senses in query manipulation [Cañas et al. 2003; Mihalcea and Moldovan 2000; Liu et al. 2004]. Gonzalo [Gonzalo et al. 1998] also achieved an increase in document retrieval performance by applying it at the document indexing phase. Nonetheless, when it comes to automatic query expansion using its synsets, experiments [Voorhees 1994] have shown that there is no higher retrieval effectiveness for long queries, though there is the potential to improve an initially short query.

### 2.2.2 *Term Variation*

Bilotti et al. [Bilotti et al. 2004] quantitatively compared two different approaches to handling term variation: applying a stemming algorithm at the indexing time, and performing a morphological query expansion at the retrieval time. Their results showed that stemming resulted in decreased recall, while retrieval-time query expansion

increased recall. Their tests were conducted using a question-answering test collection that they meticulously constructed by hand. Their experimental results give us an indication that inflectional expansion will also benefit our searching needs.

### *2.2.3 Combining Evidence at Different Levels*

Callan [Callan 1994] concluded in his study that with regards to document retrieval, it is always best to combine document-level evidence and passage-level evidence. His attempt aimed at filtering the document retrievals by zooming in to finer evidence at the passage level. In our study, the direction of the refinement is reversed. It is the returns of the smaller unit that need to be confirmed by the evidence from a larger unit. Verification is needed to see whether Callan's assertion still holds true for the reversed direction.

Salton et al. [Salton et al. 1993] has also used the approach of combining global and local processing, but in a different area of text analysis: clustering and connecting relevant text excerpts. They took a top-down approach that began with comparing pairs of full documents. If the similarity of a pair exceeded a given threshold, the documents were successively broken down into smaller sections. Section level similarities were further examined. Documents that didn't show a satisfactory global similarity between each other were immediately considered irrelevant, without conducting further text excerpting. Similar to ours, the global evidences were used as base verifications to qualify partial or finer local evidences.

### 3. THE READER'S INTERFACE

At present, there are a variety of tools available to operate electronic texts. Microsoft Word, web browsers, and Adobe Reader are all very popular. However, despite their popularity, they lack some features that we hope to include in our study of thematic reading: the functionality to indulge solely in a focused reading, and the support to navigate back and forth between the focused reading and the regular linear reading. As we mentioned earlier, thematic reading is marked by a tendency to constantly go off the trail of linear movement and hop through discrete locales that concern a certain subject. From time to time, readers choose to retreat back to the regular flow when their interest is satisfied, or when they simply want to look for information that is initially deemed irrelevant but has induced interest as the reading carries on. Therefore, to observe and investigate the needs of thematic readers, it is crucial to have direct support for rendering both linear and nonlinear reading patterns. At the same time, it is also important to have the means to switch between these two.

Currently there is no direct support in the tools mentioned above for the idea of focused view, together with transitional navigation. Though there are ways to eventually establish such an effect, the workarounds are usually less than optimal. The most commonly seen method is to create a list of anchors and use it as a side map for directing thematic reading. For instance, in Microsoft Word, people may add bookmarks (Figure 1) to locales of interest. With this method, the effort of traversing individual pieces is reduced because the platforms usually provide a compact view of all the anchors placed together. The anchors can direct you to the exact location. However, this view of anchors



tends to show only a brief description of the bookmarks. For instance, Microsoft Word won't even accept bookmark names that include spaces. Thus it is difficult to remember what the destination is really about. To see the more meaningful context, you have to select and physically go to each individual bookmark. Even if the title of the anchor is carefully composed to reflect what it represents, the list always exists as an entity apart from the narrative. The reading experience is thus not immersive, as you constantly break the action of reading to find the next piece of text, in order to continue. There is simply no easy way with the tools popular at present to indulge in a focused reading and, in the meanwhile, maintain the capability of returning to the linear flow of the narrative.

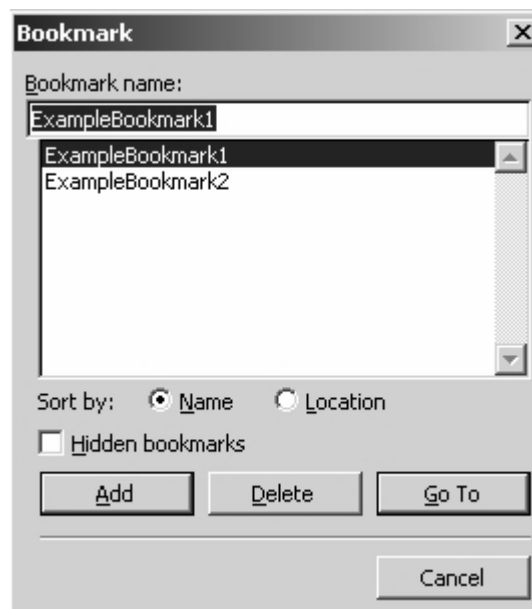


Figure 1. Example of using bookmarks in Microsoft Word

With this observation in mind, we implemented a reader's interface in our own study. Our goal for this reading system was to include the following essential features: 1) marking of the content relevant to a theme; 2) viewing of the thematically related content only; 3) navigating between the theme and the original. This interface, though preliminary, helped us to verify and test some of our hypotheses regarding the various scenarios that thematic readers encounter. We devote the present section to the description of this interface. We separate the features into two groups: the basic functionalities and the interactive ones. We first cover very briefly the basic functionalities as they are implemented to satisfy the peripheral and non-essential needs of the studies. We then move on to present the more interesting interactive functionalities, which are designed more specifically for the purpose of thematic studies.

The reader interface was developed as a web application. It is loaded with a full English translation of the book *Don Quixote*. Figure 2 shows the main window of the tool. It consists of three panels: the Tool Bar Panel on the top, the Anchor and Information Panel on the left, and the Book Representation Panel on the right. The Tool Bar Panel contains the general action buttons that apply to the whole application, such as save and open. The Anchor and Information Panel also contains a set of action buttons. These buttons are either navigation anchors or they involve additional input or output. We use the Anchor and Information Panel to host the extra information. Lastly, we display the content of the book in the Book Representation Panel.

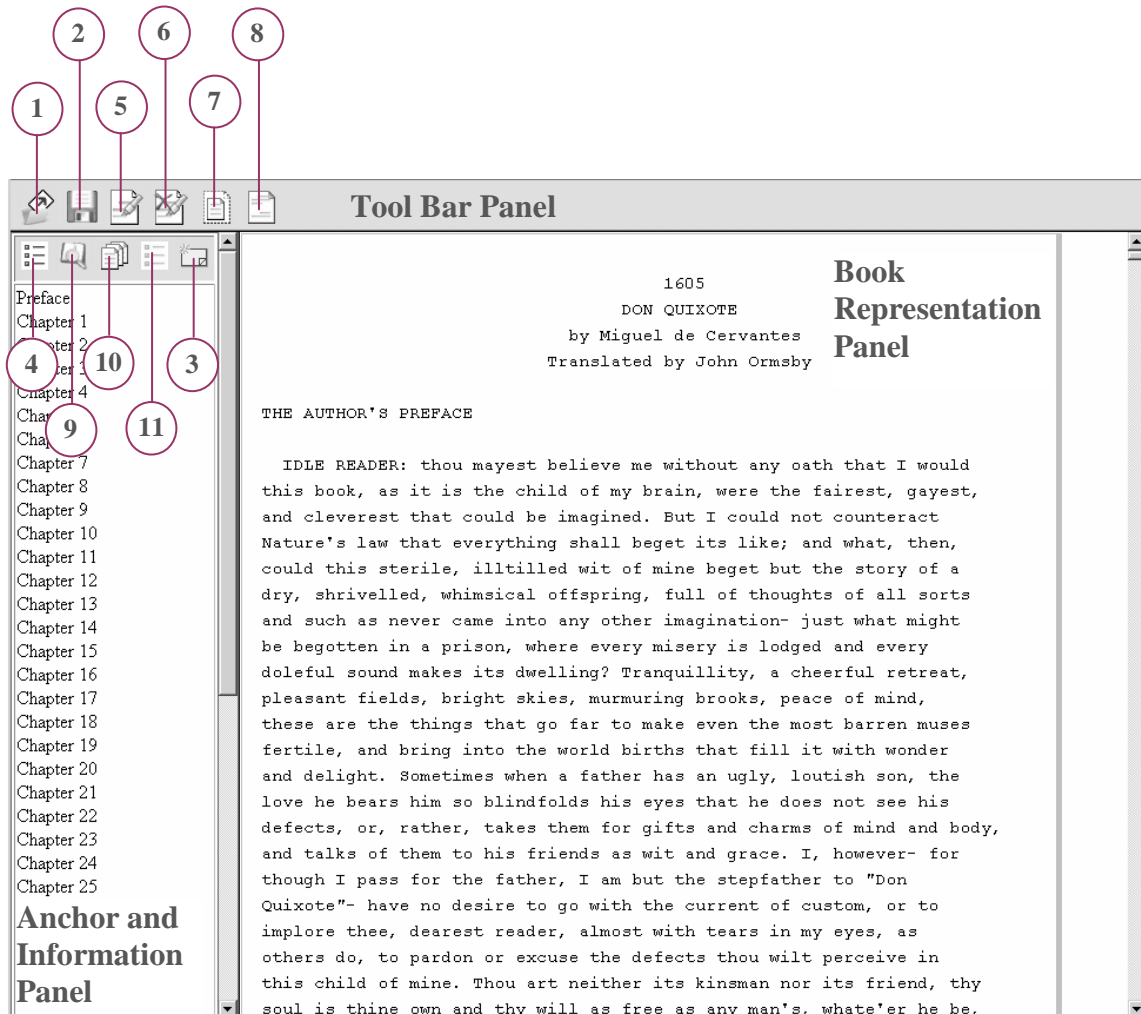


Figure 2. The main window of the Reader's Interface

### 3.1 Basic Functionalities

1. Open. This button is to open a copy of the book that has been saved earlier. A standard File-Open Dialogue (Figure 3) pops up and users select the file to open by traversing through the file system. Each opened application of the interface hosts one copy of the book at a time. Therefore, when a selection is made and the file is loaded, the

interface is refreshed to reflect the information from the newly opened file, such as the highlight status, the author, the title of the topic, and the like.

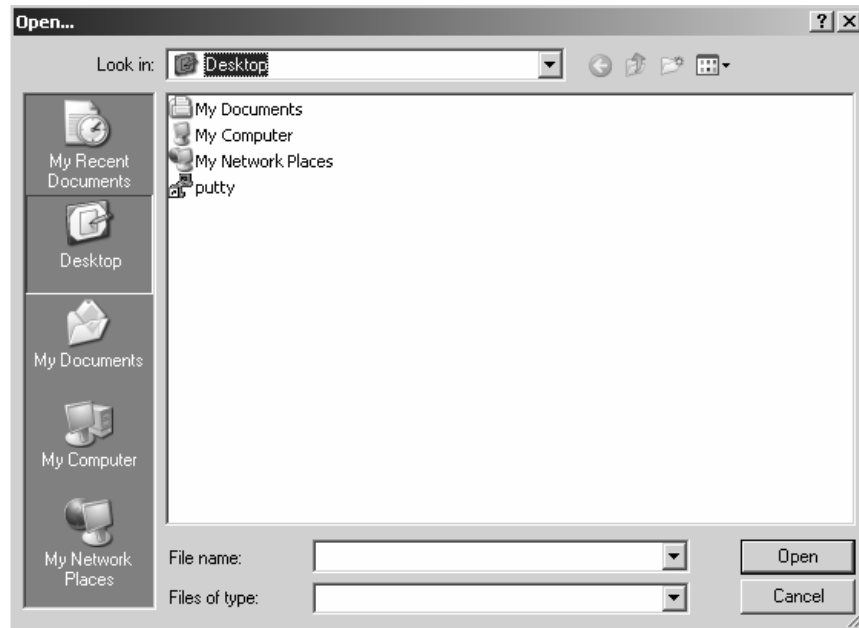
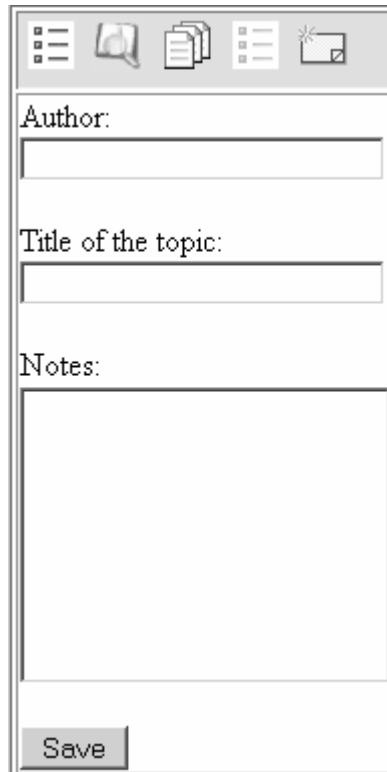


Figure 3. An Open-File Dialogue for opening a saved copy of the book

2. Save. This button saves the working copy of the book, as seen in the interface. Clicking on the button activates the Add Information Button in the Anchor and Information Panel (Figure 4). Users enter or verify the information and then press the Save button on the bottom to save. A File-Save Dialogue similar to Figure 3 prompts users to assign a location to save. The book is saved in XML format with information preserved regarding the highlight status, the author, the title of the topic, and the extra notes.

3. Add Peripheral Information. This button prompts for the entering of the author, the title of the topic, and some notes regarding the current copy in display.



The image shows a vertical window with a toolbar at the top containing five icons: a list, a speech bubble, a document, another list, and a window with a checkmark. Below the toolbar are three input fields. The first is labeled 'Author:' and is a single-line text box. The second is labeled 'Title of the topic:' and is also a single-line text box. The third is labeled 'Notes:' and is a larger multi-line text area. At the bottom left of the window is a button labeled 'Save'.

Figure 4. Peripheral information regarding the current copy

4. Chapter Anchors. This button displays a list of chapter anchors in the Anchor and Information panel (Figure 2). Each anchor points to the start of a chapter and scrolls the display between chapters.

### 3.2 Interactive Functionalities

5. Add Highlight. This button is to add highlight to the user-selected texts in the Book Representation Panel. The highlighting effect is switched on for the parts of the selection that are not yet highlighted. Highlighted texts in the selection are not affected.

6. Remove Highlight. This button erases the presence of a highlight from the current user selection in the Book Representation Panel. Texts in the selection that don't initially have the highlight turned on are ignored.

7. Selection View. In the context of the dissertation, we use the term "selection view" to refer to the display in the Book Representation Panel that contains solely the contents that are highlighted (Figure 5). The Selection View button and the Whole Text View button are used to toggle back and forth between the display with the highlights only and the overall layout of the book. The purpose of making the selection view available is to test the hypothesis that readers can better focus on the information under examination with the distractions being hidden. Whenever they prefer, however, they still have the freedom to bring back the initial context as needed. Therefore, each chunk of the text in the selection view is rendered clickable. With a single click, readers are brought back to the whole text view with the particular highlights focused at the top.

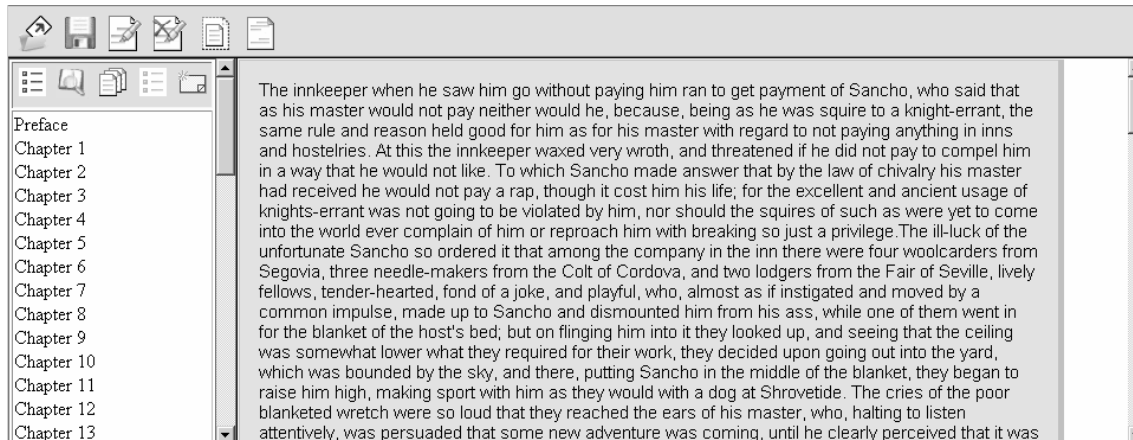


Figure 5. An example of the highlighted view

8. Whole Text View. This button switches the Book Representation Panel back to the whole text view in which the book is displayed as a complete entity with the highlights scattered and embedded throughout the document (Figure 8).

9. Search. This button opens a query input box in the Anchor and Information Panel. Users type a search query. Clicking on the search button fills the panel with a list of search returns (Figure 6). The search engine is built on top of the open source full-text search API Apache Lucene [Apache Lucene 2008]. Before a search command is issued, the initial user input goes through two steps. 1) The query input is parsed using Lucene's StandardAnalyzer to remove the stop words. 2) Inflectional variants for the individual terms are added to the parsed query to form an expanded query. More details about the expansion are given in Section 5.2.1. With the query prepared, we conducted a *Boolean* search to look for the sentences that contained any instance of the terms in the expanded query. When displaying the results, the matching sentences are grouped by their matching chapters and are shown with a certain number of characters (the first eighty) to

provide a quick glance of the entry. The chapter number and the characters are clickable in an effort to bring to focus, respectively, the beginning of the chapter and the particular sentence in the Book Representation Panel.

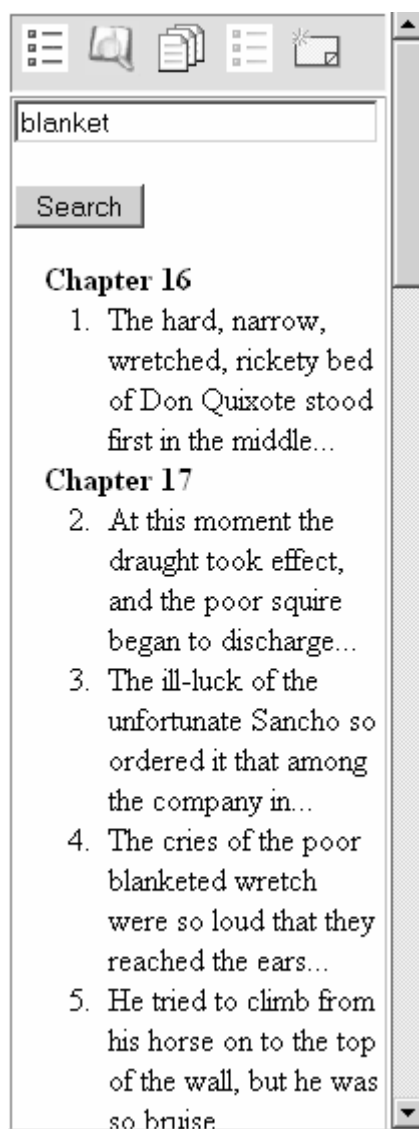


Figure 6. An example of the search operation



10. List Suggestions. This button displays a list of copies of the book with highlights that are available for suggested reading. Figure 7 shows an example with three readings offered as examples. The availability of the suggested readings is determined by two factors: the presence of a document at a designated data folder and a property setting that configures the document as a recognized suggestion in the system. This sanction procedure is a simulation of the situation where an authoritative suggestion is screened and justified by authorities before it is made public.

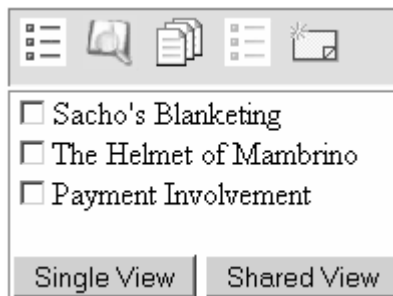


Figure 7. An example list of suggested readings

When loading the suggested readings, users have the choice of selecting individual or multiple copies by pressing the Single View button or the Shared View button. Single View (Figure 8) loads the first selection and displays it, which is the same action as that of regularly opening the selection except that the file name is implicitly integrated and thus doesn't require an explicit specification of the location of the file.

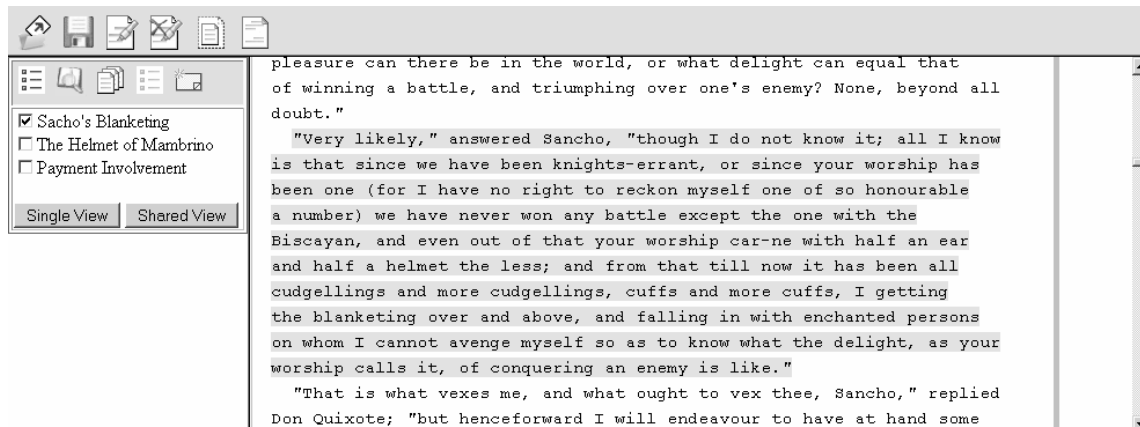


Figure 8. An individual suggested reading in Single View

Shared View, on the other hand, loads all of the selections and combines them in one display (Figure 9). Common highlights from all the selected copies remain as highlights. Non-overlapping highlights from each copy are visualized with color brackets. Each pair of brackets marks the beginning and end of the highlighted block in an individual copy. To make the block more apparent, users may press down the brackets, which will add an underline effect to the enclosing block. This view is designed for thematic reading that involves multiple threads.

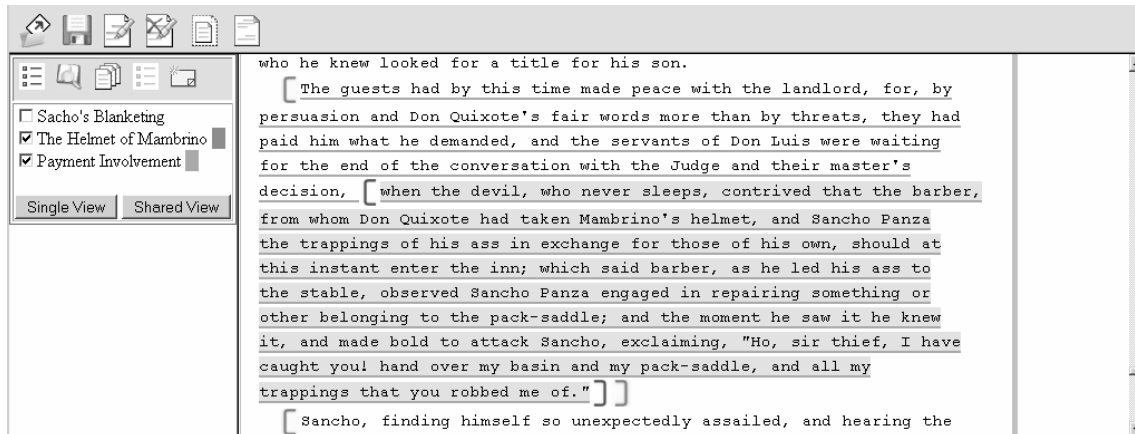


Figure 9. Two individual suggested readings in Shared View

11. List Highlight. This button generates a list of highlights currently marked in the interface (Figure 10). Continuous portions of the highlights appear in the list as one entry displayed with a certain number of characters (the first eighty). Clicking on the individual entry moves the Whole Text View to the beginning of the highlight. This button is provided such that people may easily scroll up and down the highlights while staying within the whole text view.

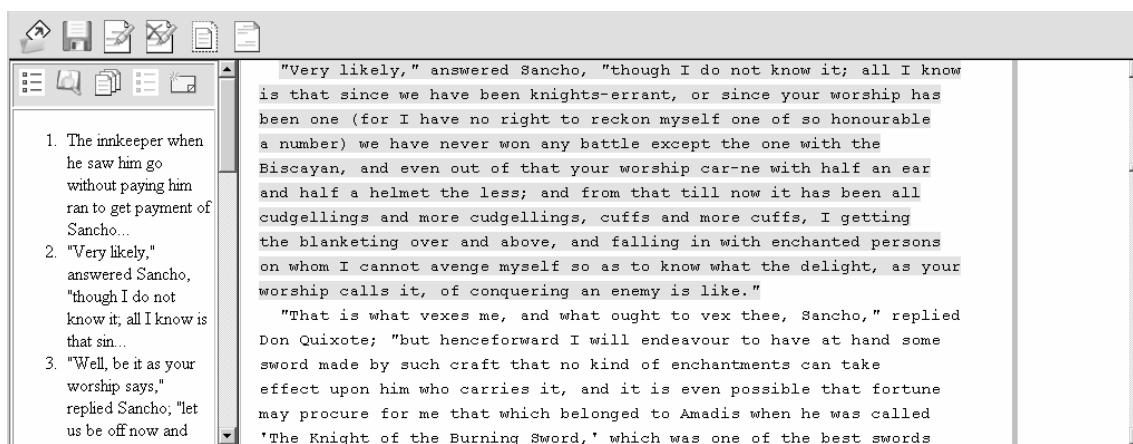


Figure 10. An example of the list of highlights

#### 4. USER STUDY AND EVALUATION

There are scenarios of thematic reading that involve simply one person and one purpose of reading the book. For instance, a student taking a literary criticism class might be given an assignment to write an essay on the topic of madness in *Don Quixote*. He/She would read the book with a focus on the topic and closely examine those excerpts that contain such elements. There are also scenarios of thematic reading that involve multiple readings of the book. These could include readings of one single topic by different people or of different topics. The example scenario we give above could easily be extended to reflect the case of multiple readings. The student, while working on the assignment, has a discussion with another student who happens to pick instead the topic of dialogue for the assignment. While discussing their progress on their assignments and their findings regarding the two topics, they discover that the two themes overlap from time to time. The intertwining of the two readings may well arouse a vivid discussion of their understanding and comprehension of the book.

To observe people's behavior under the different situations, we experimented with two different scenarios in our user study. One scenario deals with the case of a single reading, and the other with multiple readings. Through observing the subjects' behavior, we hoped to discover some guidelines for future interface designs that target the needs of thematic reading. We first give a brief description of the subject pool of the user study. We then go on to present the experimental designs and the results for the two scenarios, respectively.

## 4.1 Subject Pool

A total of 12 subjects (Table I) participated in our study. The subject pool consists mainly of graduate students majoring in Computer Science and Hispanic Studies. The test book used is the English translation of *Don Quixote* by John Ormsby [Cervantes Saavedra 1885]. As the test book is a translation from its native copy, we tried to include in the pool users who have experience with foreign languages, and thus are relatively used to reading writings in languages other than their mother tongue or translations from a foreign language. The subject pool is also diversified to include people who have everywhere from none to excellent prior knowledge of *Don Quixote*. Subject 4, 8, and 12 are considered expert representatives. All three have read the book more than 3 times and have recently taken a literature course that focuses on the critical study of the book and its influence in the literary community.

## 4.2 Experiment on Single Reading

### 4.2.1 Experimental Design

When people conduct thematic reading, a few assumptions seem plausible. First, they may prefer to have the relevant content grouped together and work more closely with that content. Second, in the mean time, it is important to be able to resume the normal view at any time. The present experiment is designed to examine whether these assumptions indeed hold. At the same time, it also gives us the opportunity to conduct behavioral observations related to thematic reading in general.

Table I. The subject pool of the user study

<b>Subject ID</b>	<b>Native language</b>	<b>Language skills other than the native</b>	<b>Number of times having read <i>Don Quixote</i></b>
1	Italian	English - fluent; German – beginner	0
2	Arabic	English - fluent	0
3	Greek	English - good	1
4	Spanish	English - proficient; Italian - intermediate; French - intermediate	5
5	Spanish	English - fluent; Italian - fluent; French - fluent; German - intermediate; Catalan - intermediate	2
6	English	Japanese - low-level fluent; Latin - very low-level	0
7	Turkish	English - good	0
8	Spanish	English - 80%; Italian - 70%	4
9	Spanish	English - excellent	1
10	English	Spanish - very low-level	0
11	Chinese	English – fluent	1
12	Spanish	English - excellent; Portugese - fair; German – fair	3

The experiment is composed of two subtasks that follow the same procedure but are carried out in two different systems: Microsoft Word and our own implemented Reader's Interface. We abbreviate them as Word and Interface in the Tables. In each subtask, users are assigned a specific topic for which to read. The excerpts regarding the topic are pre-highlighted in the text. In Microsoft Word, there is currently no direct way to display only those parts with highlights. In the Reader's Interface, users have the choice of switching back and forth between the selection-only view and the normal view.

At the beginning of each subtask, we gave each participant a copy of *Don Quixote* with highlights concerning one assigned topic. The participants had the option of taking up to 15 minutes to read about the topic and get themselves familiar with it. After that, they took a quiz about the topic. Nevertheless, there was no need to remember any specifics of the storyline. The quiz was completely open book and had no time limitation. The purpose of the reading is mainly to familiarize the subjects with the topic so that they would have a rough idea of where to refer back to for the answers to the quiz. As there was no time limitation imposed on the quiz, the participants could choose to further scrutinize the text about topic, at any point. Their goal with this task was to try to answer all the questions correctly. The quiz was not handed to the participants until: 1) they requested the quiz when they finished the reading, before they used up the 15-minute slot, or 2) the end of the 15-minute slot. Users were informed that the highlights present in the text were about the topic given. Nevertheless, they were free to read anything they liked. Before and after each task, they were asked to grade a self-



evaluation of their knowledge about the topic. During the course of the user study, the investigators conducted over-the-shoulder observations and took notes. Additionally, surveys containing questions regarding the general experience were also used to collect responses directly from the participants.

The two topics used in the experiment are Sancho's Blanketing and the Helmet of Mambrino (Table II). We abbreviate them as Blanketing and Helmet when discussing the experimental results in this section. We alternated the combination of the topic and the system during the experiment. We also alternated the sequence of the system to minimize the influence of prior impressions introduced by the first system used.

#### 4.2.1.1 Experimental Procedure

1. Participants provided demographics information.
2. Participants were given a brief introduction about the procedure of the task.
3. Participants were assigned a topic and a system. If the system used was the Reader's Interface, participants were given a quick tutorial on the tool.
4. Participants responded to the question regarding their prior knowledge of the topic (Table III).
5. Participants were given up to 15 minutes to carry out the reading about the topic.
6. Participants took the quiz.
7. Participants responded to a follow up survey (Table III) about their experience conducting the task in the designated system.
8. Participants repeated steps 3 – 7 on the other topic in the other system.

Table II. Information about the topics

Topic	Information about the highlighted excerpts	Quiz questions	Acceptable answers to the questions
<b>Sancho's Blanketing</b>	1766 words; 122 lines when rendered in size 12 Times New Roman; extracted from 10 chapters	1. Right after Sancho is blanketed, what is the compassionate gesture Maritornes showed to him?	Offer him a jar of water.
		2. What does Don Quixote think about the blanketing incident?	He believes that it is all enchantment.
		3. How is Sancho's behavior later on affected by the blanketing incident?	The blanketing is a painful and embarrassing memory in Sancho's mind. He complains several times to Don Quixote. He is also hesitant to go into the same inn the second time they arrive there because of the incident.
<b>The Helmet of Mambrino</b>	1842 words; 128 lines when rendered in size 12 Times New Roman; extracted from 6 chapters	1. Please describe briefly the episode of Don Quixote's acquiring the helmet.	A barber is wearing a basin to avoid spoiling his hat by the rain. Don Quixote sees the basin and believes that it is a golden helmet. He attacks the barber with his lance and extorts the basin.
		2. Right after Don Quixote acquires the helmet, he calls it a head piece; what is Sancho's reaction when hearing that? Does Sancho believe that the basin is indeed a helmet?	Sancho laughs at Don Quixote's mistaking a basin as a helmet. No, Sancho thinks it is just a barber's basin.
		3. When the barber tries to reclaim his basin later in the book, why does the other barber agree with Don Quixote and say that the basin is indeed a helmet?	To carry out a joke for the general amusement.

Table III. Survey questions

<b>Pre-reading question</b>	Your knowledge about the topic is: 0 1 2 3 4 5 6 7 8 9 10 None Neutral Expert
<b>Survey after finishing each quiz</b>	1. Your knowledge about the topic is NOW: 0 1 2 3 4 5 6 7 8 9 10 None Neutral Expert
	2. Please describe the procedures that you took to answer the questions in the quiz.
	3. What features of the tool are useful for doing the task?
	4. What features of the tool are not effective for doing the task?
	5. Do you have any other suggestions for better doing the task?
	6. (Expert only) The highlighted part represents the story: 0 1 2 3 4 5 6 7 8 9 10 Poorly Neutral Perfectly
<b>Survey after finishing both tasks</b>	1. It is important to have the capability to view the highlights only, instead of a whole book. 0 1 2 3 4 5 6 7 8 9 10 Strongly Disagree Neutral Strongly Agree
	2. It is important to have the capability to traverse easily back and forth between the highlights and the original text. 0 1 2 3 4 5 6 7 8 9 10 Strongly Disagree Neutral Strongly Agree
	3. In order to fulfill the task, you prefer to use ○ The Reader Interface ○ Word
	4. In order to fulfill the task, the tool chosen above is clearly better than the other. 0 1 2 3 4 5 6 7 8 9 10 Strongly Disagree Neutral Strongly Agree

9. Participants responded to a system comparison survey after finishing both tasks (Table III).

10. Participants were debriefed.

We discuss the results of our experiments from two different aspects. We first look into the general phenomena of thematic reading. From practical observations, we would like to determine how acceptable the idea of thematic reading is to the subjects and how possible the comprehension of a topic is by following the thread of a single storyline. We then analyze the behavioral responses of the subjects regarding the features of the interface. We hope to find out for future designs what the essential desirable features are and what might inhibit a reader's experience.

#### *4.2.2 Discussion of the Phenomena of Thematic Reading*

One open question that has driven the design of our study is how practical thematic reading is for advanced readers. Our results show that it is indeed a recognized reading strategy and advanced readers adopt it often. The study has revealed the following: 1) the majority of the subjects immediately indulged in the view, when available, of the extracted storyline; 2) the experts' ratings on how well the excerpts represent the chosen topics were favorable; 3) the majority of the subjects were able to understand the extracted storyline to a degree good enough to answer the quiz questions correctly.

##### *4.2.2.1 Choice of Reading*

Table IV shows the choice of reading when the subtask was carried in the reader's interface. Half of the subjects stayed exclusively in the selection view during

the reading. Before each study, we presented to the participants the choice between the selection view, which contains only the content regarding the assigned topic, and the whole text view, in which the topic content is embedded. Half of the subjects immediately chose the selection view, without the study investigator hinting or suggesting one way or the other. The other half of the group spent more than 90% of their time in the highlighted view, as well. Only three of the subjects evidently preferred to carry out the reading with the whole text present and spent more than 90% of their time in the regular view. Nevertheless, for those who didn't stick with the view of the selection during the reading, they all used it as an anchor to traverse back and forth when locating the highlights. This tells us that although they preferred the feeling of the overall book, the way they carried out the action was still by following the thematic thread. All the same, while taking the quiz, they all stayed mainly in the selection view, with only an occasional or brief switching to the whole text.

When using Microsoft Word, although there isn't a selection view available, we witnessed similar patterns of preferring reading only about the relevant information. We will cover this in detail in Section 4.2.3.1.

Table IV. Choice of reading when the subtask was conducted in the Reader's Interface

	Subject ID											
	1	2	3	4	5	6	7	8	9	10	11	12
During reading, the percentage of total time that the subject stayed in the highlighted view	10	100	100	50	100	70	10	100	90	100	10	100
The subject used the highlighted view as the anchor for locating the text for reading in the normal view instead of scrolling directly from highlights to highlights	√			√		√	√				√	
When taking the quiz, the subject started with the highlighted view and stayed there most of the time looking for the answers	√	√	√	√	√	√	√	√	√	√	√	√

*Note: Grayed area is data collected through investigators' observations. Clear area is data collected from the survey responses.*

In the study, no one ever questioned the idea of reading a small portion of a lengthy document to gain the understanding of a certain topic. The idea seems to be readily assumed. When the subjects were told that someone has marked the book about the subject, they showed a strong preference for using the information given. All these examples suggest that the idea of thematic reading is not new to advanced readers. They seem to acquiesce that it is possible to convey the idea of a certain topic, at least for the purpose of fulfilling a special activity, by delivering only the relevant parts - however scattered they might be throughout a lengthy document. The behavior of thematic reading, at least in advanced readers, seems to be both predictable and natural.

#### 4.2.2.2 Comprehension of the Topics

Three expert participants were asked to give their opinions about how well the highlighted excerpts covered the topic. Their rankings are shown in Table V. Both topics received grades of 8 and above, out of 10. This is an acknowledgement from the experts regarding the potential effectiveness of using cuts of the text to represent a certain topic. This also shows that, in the judgment of the experts, our selections were able to present the individual topics, to some degree, successfully.

Table V. Expert ratings on the highlighted excerpts for representing the topics

Subject ID	Topic	Tool used	Rating of the selection (out of 10)	Topic	Tool used	Rating of the selection (out of 10)
4	Blanketing	Word	8	Helmet	Interface	10
8	Blanketing	Interface	10	Helmet	Word	10
12	Blanketing	Word	8	Helmet	Interface	8

The compositions of the three questions in the quizzes are designed to test different levels of understanding in the subjects. The first question is the easiest. Reaching the answer simply requires a search for the fact in the storyline. The last one is the most difficult. There is no explicit answer in the text. A correct response relies on an understanding and interpretation of the overall storyline. The second question falls in between. It is not as explicit as the first. A straightforward fact-searching won't yield the answer. Compared to the third, the answer to the second question was repeated multiple times under different contexts. Therefore, even without a good understanding of the whole storyline, the repetitions may help to draw a good prediction of what to reply.

Table VI records the participants' number of correct answers in the quizzes and their self-evaluations regarding their knowledge about the topic before and after conducting the tasks. Ten out of twelve users answered the three questions in both quizzes correctly. Subject 2 missed the third question in the first quiz, but achieved a perfect score on the second quiz. Subject 9 missed the third question in the first quiz and the second question in the second quiz. In total, we received only three incorrect answers. Two were the third question about the Helmet of Mambrino, and the other was the second question about Sancho's Blanketing.

Figure 11 charts the increase in knowledge of the topics as self-rated by the users. The mosaic bars are the tasks that began with a relatively high pre-knowledge, yielding ratings of 4 and above. Starting with a high pre-knowledge understandably leaves less space to grow. It is reasonable that the increase would, consequently, stay in the lower range. The solid bars are the tasks in which the participants had limited pre-



knowledge, yielding ratings of 2 or less. The subject comes in as a novice in these tasks. Among them, only two – the second task for subjects 9 and 10 - received a rank of increase less than 5. The three tasks that contained incorrect responses in the quizzes had a score of 5, 5 and 4. All were in the relatively low range as compared to the rest, which tells us that the self-ranking shows a correlation to the knowledge that the user gained.

The fact that the subjects who started with a very limited understanding of the topics were able to accomplish a level of high accuracy of the quizzes, even those questions that required a significant understanding of the overall storyline, together with the fact that they all expressed a fairly positive increase of their confidence level on the topics, is practical proof that thematic reading is capable of achieving the goal of understanding a specific topic, at least to a degree that succeeds in topic-wise tasks. It is worth noting that this can be achieved without knowing the general idea of the rest of the book.

Table VI. Subjects' understanding of the topics

Subject ID	Tool used	Topic	Correct answers in the quiz	Self rating of the knowledge of the topic (out of 10)		Tool used	Topic	Correct answers in the quiz	Self rating of the knowledge of the topic (out of 10)	
				Pre	Post				Pre	Post
1	Interface	Blanketing	3	0	9	Word	Helmet	3	0	7
2	Word	Helmet	2	0	5	Interface	Blanketing	3	0	6
3	Interface	Helmet	3	1	8	Word	Blanketing	3	1	6
4	Word	Blanketing	3	5	9	Interface	Helmet	3	7	9
5	Word	Helmet	3	2	8	Interface	Blanketing	3	6	6
6	Word	Blanketing	3	0	7	Interface	Helmet	3	4	7
7	Interface	Helmet	3	0	6	Word	Blanketing	3	0	5
8	Interface	Blanketing	3	8	9	Word	Helmet	3	8	9
9	Interface	Helmet	2	0	5	Word	Blanketing	2	2	5
10	Interface	Blanketing	3	0	5	Word	Helmet	3	0	4
11	Word	Blanketing	3	0	8	Interface	Helmet	3	0	7
12	Word	Blanketing	3	6	8	Interface	Helmet	3	5	8

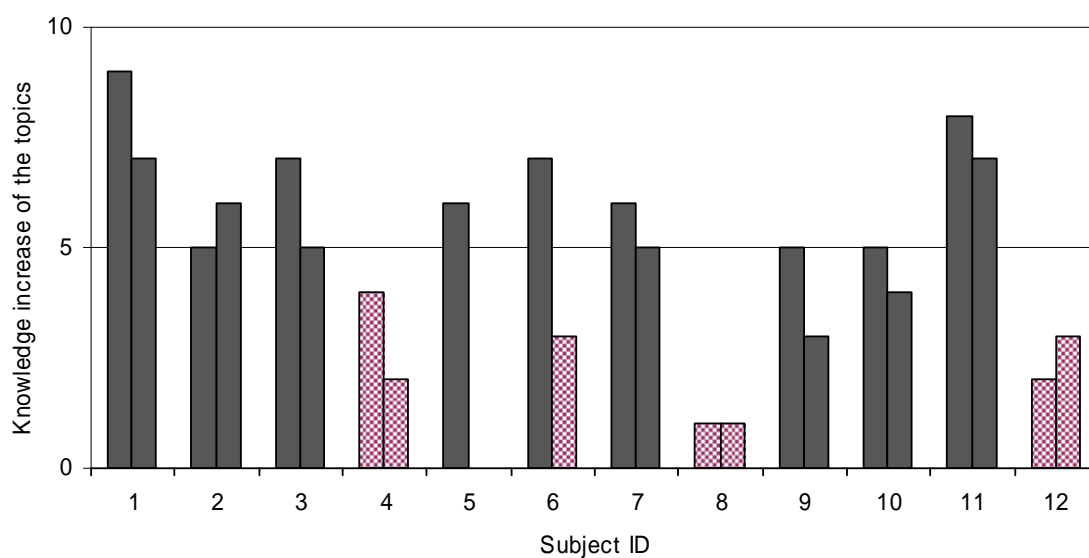


Figure 11. Increase in the knowledge of the topics

### 4.2.3 *Discussion of the Guidelines for Immersive Design*

Thematic reading is a behavior often mixed with regular navigation of the narrative. Regular traversal is comparatively passive, as it follows simply the predetermined flow presented by the author. Its presence exists without the user's input. The thematic way of reading is participatory and interactive, as it takes and targets the particular interests of the reader. In discussion of the aesthetics of the participatory medium, Murray [Murray 1997] identified three aspects of assessment: immersion, agency, and transformation. Immersion evaluates the experience itself. It studies how much people are able to indulge in and absorb things freely within the media. Agency emphasizes the capability of personalizing and taking part in the media. Transformation studies the possibility of customizing the media to suit the desire of the participant with a different type of experience. As we can see, agency and transformation examines especially the gaming effect of the media. For our study of thematic reading, we are interested in immersion. In fact, immersion is exactly our key direction in looking for guidelines for the design of a reader interface that supports thematic reading. We would like to see that future designs are designs that can make the reading experience interactive, immersive, and indulgent.

Below we organize the discussion of the feedback and observations that we gathered about the immersive aspect of the experience into three main sections: the view, the navigation, and the contextual awareness. Such a division, however, is only meant to facilitate simplicity in the organized presentation of the results. They are, by no means, mutually exclusive to each other. On the contrary, they are highly interrelated.

There are also a few minor observations that are not directly related to the three main topics of inquiry. We briefly discuss them in the last section.

#### 4.2.3.1 Views

We have already seen in Table IV that the subjects show a strong preference for using the view that displays the thematically related information together, as compared to bouncing from place to place in the original book in the Reader's Interface. Additionally, eight subjects (Table VII) stated in the survey that the highlighted view was helpful. After conducting the study in Microsoft Word, nine participants expressed their frustration with locating the highlighted portions to read. Of these nine, six explicitly stated that they wished there was a view that contained only the highlights. It is important to note that half of those who made this comment were given Microsoft Word as the first system, and thus were not influenced by the availability of such a view in the Reader's Interface. Subject 3, in particular, created a brand-new document before conducting the reading, manually copying each highlighted area into the document, and completely carrying out the reading in the separate document.

Table VII. Observations related to the view

	Subject ID											
	1	2	3	4	5	6	7	8	9	10	11	12
<b>In the Reader's Interface</b>												
The subject has expressed in the survey that the highlight view is helpful	√	√	√			√	√		√	√	√	
<b>In Microsoft Word</b>												
The subject complained in the survey that locating the highlights was unpleasant	√	√	√	√	√		√	√	√	√		
The subject explicitly stated in the survey that they wished there was a view with the highlights grouped together		√	√	√	√		√		√			

Table VIII. Rating the importance of a compact view of the selected text

	Subject ID											
	1	2	3	4	5	6	7	8	9	10	11	12
It is important to have the capability to view the highlights only, instead of seeing them embedded in the whole book (out of 10)	8	10	9	10	10	7	10	10	10	10	10	9

After the users finished the two parallel tasks, users ranked the importance of having a view with the selected text only (Table VIII). Eight participants strongly agreed. The ubiquity of using the selection view in the Reader's Interface, the expressed desire for a similar view in Microsoft Word, and the high rating of the significance of such a feature, together clearly show that it is important to have the capability of viewing the excerpts related to a target topic only.

#### 4.2.3.2 Navigation

Another important aspect of the experience is navigation. When the study was carried out in the Reader's Interface (Table IX), eleven users used the single click switch to return to the original book from the highlighted view. Moreover, most noticed this specific action and remarked on it in the survey, saying that it was both important and useful. Even without any hint given by researchers, commenting on the bidirectional navigation between the two distinct views, five participants addressed this issue explicitly in the survey, stating that a smooth implementation is crucial. After the survey, we asked the subjects to rank their feelings regarding the bidirectional navigation. The responses (Table X) further confirmed its importance.

Table IX. Observations related to navigation when the subtask was conducted in the Reader's Interface

	Subject ID											
	1	2	3	4	5	6	7	8	9	10	11	12
When taking the quiz, the subject did click on the highlights to switch from the selection view to the whole text view for more contexts	√	√		√	√	√	√	√	√	√	√	√
The subject expressed in the survey that it is important to have bidirectional movement between the highlighted view and the normal view	√	√				√				√		√
The subject tried to select the text in the highlighted view, but was then was taken to the whole text view		√					√		√			
The subject expressed specifically the inconsistency in the way that the bidirectional navigation is designed		√								√		√

*Note: Grayed area is data collected through investigators' observations. Clear area is data collected from the survey responses.*

Table X. Ratings about the importance of the bidirectional navigation

	Subject ID											
	1	2	3	4	5	6	7	8	9	10	11	12
It is important to have the capability to traverse easily back and forth between the highlights and the original text (out of 10)	10	9	10	10	10	9	10	10	7	10	10	10

The consistency of the bidirectional traversal also attracted a lot of attention in the study. The design of the interface uses one click to go from the selection view to the whole text view, while the other way around is achieved by explicitly clicking on a button on the tool bar. Subjects 2, 7, and 9 complained that the system should follow the same design and the user should be taken to the selection view by clicking on the highlighted area in the book layout. At the same time, due to the inconsistency of the navigational mechanism, subjects 2, 10, and 12 noted that they tried once to select certain parts of the text while in the selection view, but were taken to the whole text, which was contradictory to their expectation and, consequently, interruptive. Subject 1 and 7 also noted in the survey that they should be directed back to the previous viewing location each time they switch back to the selection view. The present implementation of the interface doesn't record the final browsing status. The fact that the cursor is redirected to the end makes the reading of the theme intermittent. Most of the time the users switched to the book view simply to get a quick glance of the surrounding context. Then they had to scroll to be where they were in the selection view when they came back.

When there was no compact view of the selections one click away, as in Microsoft Word, some participants initiated ways to speed up navigation. The method most often observed was use of the thumbnail view. Subjects 6, 7, and 9 used the "Zoom" function to assist their navigation. Subjects 6 and 7 constantly switched between two sizes of the zoom: one extremely small, such that thumbnails of a big chunk of the document were visible on the screen, and the other the normal size. The small size was



used to get a glance of the locations of the yellow areas, and the normal size was used to read the actual text. Subject 10 opened two copies of the document in different windows, with one window containing the thumbnails as a navigation guide and the other the regular display to conduct the reading. Subject 2 created bookmarks for each chunk of the highlights and used the bookmark to assist the navigation.

These creative efforts in Microsoft Word bear a great similarity to the zigzag pattern seen in the use of the selection view in the Reader's Interface. Users all used a view where a glance of the selections could be obtained. From there they travelled to the specific locales in the book. They stepped back to the compact view to get another glance, and then set out to examine the book once again.

Bidirectional navigation apparently is one major element of user need, when it comes to thematic reading. When there is a condensed view available for the theme, most participants automatically choose it over the normal layout. At the same time, they consider navigation between the two views to be a key element that should be designed consistently and self-evidently.

#### 4.2.3.3 Contextual Awareness

Subjects 1, 5, and 6 commented that although they preferred to use the selection view, they would have liked to have some sort of context indication to keep them aware of where they are. Subject 1 suggested having a narrow panel on the right-hand side. Along with the selections, the panel would list the number of the chapter from which the selections were extracted. Subject 5 suggested providing an index of the highlighted sections, as well as adding a bit more of the snippets of text that surround the highlights,

even in the selection view. Subject 6 thought that a number of words/lines/pages in between the highlighted sections would be helpful. Subjects 1 and 9 further proposed one possible solution: providing a document thumbnail map, like the PDF navigation panel with the page thumbnails on the side. The map would be concurrent to the reading action in order to provide a synchronized notion of where the readers are. We have already seen subjects using a zoom view in Microsoft Word to assist their traversal of the disjointed selections. It looks like the readers of electronic texts are quite used to adopting a thumbnail view as an aerial view of their exploration.

Subjects 3, 5, and 6 mentioned that the selection view implemented in the Reader's Interface shouldn't be compressed into a big chunk of text. It would be better if they maintained the original look as in the book, or had a clearer space between the different sections.

All of these factors indicate that although thematic readers like the convenience of a compact view of their interested excerpts, the preservation of a notion of contextual surroundings of the book might give them a sense of not being lost, and thus may make them feel more comfortable or freer to indulge in views other than the original. The awareness and indication of the hidden parts does not impede, as we originally expected, the experience of thematic reading, but rather encourages readers to put aside the rest of the narrative. As a result, it promotes immersion in a flow of the readers' own creation.

#### 4.2.3.4 Other

Another interesting observation is the use of the search function (Table XI). When the study was taken in the Reader's Interface, Subjects 8 and 12 used the search

function in the Mozilla browser. The Reader's Interface showed its own search that was implemented, but the feature wasn't explicitly explained to the user. At the same time, no one raised the question of whether there was a search available while conducting the study. Neither was the search mentioned in the survey to be included as a future add-on. While in Microsoft Word, a total of nine subjects used the search function, with eight explicitly mentioning their use in the survey. The reason such a difference might have been triggered is that users are less impeded by conducting some manual processes when the information they need is right in front of them. Then, when the manual work is tolerable, they are less likely to think about technical improvements. Otherwise, without the information being readily available, manual work begins to look overwhelming. People then tend to resort to and rely on the other means of assistance on hand. Of course, the participants' familiarity with Microsoft Word might also have played a part. People are more likely to use the functions with which they have prior experience.

Table XI. Observations related to the use of the search function

	Subject ID											
	1	2	3	4	5	6	7	8	9	10	11	12
<b>In the Reader's Interface</b>												
The subject has used FireFox search to help find the answers or locate a certain context								√				√
<b>In Microsoft Word</b>												
The subject has used search to help find the answers or locate a certain context	√		√		√	√		√	√	√	√	√

*Note: Grayed area is data collected through investigators' observations. Clear area is data collected from the survey responses.*

Subjects 2 and 9 made the comment that they would like to be able to add comments or impose additional tags over the highlights. Subject 3 liked that fact that he was able to rearrange the content freely in Word, yet in the Reader's Interface, the sequence had to follow what appeared in the book. Annotations and content rearrangement both constitute research areas of their own, but are beyond the focus of our studies.

After the users finished the two parallel tasks, users responded to the question comparing the two systems (Table XII). Users all preferred to use the Reader's Interface, rather than Microsoft Word, for fulfilling the task. Furthermore, the Reader's Interface was regarded as considerably better suited to the task. Contrasting with what the Reader's Interface provides but is not supported by Microsoft Word, we draw the following guidelines for designing future reading platforms that target the behavior of thematic reading. 1) It is critical to be able to view solely the relevant contents with the irrelevant part temporarily hidden. 2) The key to the experience of thematic reading lies in the navigation between the thematic and the normal views. It is critical to provide a natural and smooth traversal between the two. 3) A notion of contextual awareness should be provided as a peripheral feature in the background, so that the reader is confident about where they are.

Table XII. Responses to system preference

Subject ID	In order to fulfill the task, you prefer to use	In order to fulfill the task, the tool chosen above is clearly better than the other (out of 10)
1	Interface	10
2	Interface	9
3	Interface	9
4	Interface	10
5	Interface	10
6	Interface	8
7	Interface	10
8	Interface	10
9	Interface	10
10	Interface	10
11	Interface	10
12	Interface	10

### 4.3 Experiment on Multiple Readings

#### 4.3.1 Experimental Design

After finishing the experiments on single readings, the same set of subjects moved on to the second part of the user study. Here, the study stretched into the arena where thematic reading involves multiple threads of reading. We mimicked a scenario where people study different topics and try to figure out how the topics relate to each other. This study was conducted using solely the Reader's Interface. The participants were given three copies of the book *Don Quixote* through the Reader's Interface. Each copy contained highlights pre-marked for a specific topic. Two of the three topics were already addressed in the previous part of the study. The topic of payment involvement, which is about payment actions that involve the two major characters, was new here.

1. Sancho's Blanketing
2. The Helmet of Mambrino
3. Payment Involvement

The task was to find those places where two or all three of the topics crossed each other, and briefly describe what the interactions were about. Out of the 12 subjects, half were asked to use only the individual copies; the other half were allowed to use the collation view available in the Reader's Interface. After they finished the assignment, they were debriefed with a survey (Table XIII) that contained the following questions.

Table XIII. Survey questions for the second scenario of the user study

<b>Survey after finishing the task</b>	1. Please describe the procedures that you took to answer the question.
	2. What features of the tool are good for doing the task?
	3. What features of the tool are bad for doing the task?
	4. What features, if available, would make it easier for doing the task?
<b>Extra survey questions to the subjects who were given the collation view</b>	1. The extent that you have relied on the collations between the topics in fulfilling the task. 0 1 2 3 4 5 6 7 8 9 10 Never Some Extensively
	2. It is important to have the collation view. 0 1 2 3 4 5 6 7 8 9 10 Strongly Disagree Neutral Strongly Agree
	3. Do you wish the collation view would be implemented in another way? If yes, how?

#### 4.3.2 *Results and Discussion*

The interactions between topics 1 and 3 and between topics 2 and 3 are well defined. Don Quixote doesn't pay for their stay in the inn and the innkeeper, who grabs Sancho at that moment, makes him suffer by tossing him a blanket. The barber from whom Don Quixote extorts the basin - his helmet - encounters them later in the inn. He requests and, in the end, receives the return of his property. There isn't a definite relational interaction between topic 1 and 2. Although they appear close to each other during the altercations of the two characters, it is highly subjective whether or not to regard the altercation as an interaction between the topics.

Table XIV shows the major observation gathered from the six subjects who didn't have the permission to use the collation view. When describing the procedure that they took to fulfill the task, both Subjects 1 and 10 stated that they alternatively switched between topics to see if any of the text matched. Then four participants, in responding to what features would make it easier to perform the task, all wished to have the capability of showing all the highlights in one common place, with the overlapping part possibly in another color.

Table XIV. Behaviors exhibited when the collation view was not given to the subjects

	Subject ID					
	1	4	5	7	10	11
The subject wished that the topics could be combined in one view	√		√	√	√	
The subject stated that the strategy of finishing the task was to find the overlapping text	√				√	
The subject stated that they stayed mainly in the topic of payment and tried to find the references to the other two from there	√	√				

When the subjects were given the option to use the collation view, the observation was unanimous (Table XV). They all jumped right into the collation view and relied on it exclusively for conducting the task. The availability of the collation was also highly appreciated by them. On average, the group given the collation view spent much less time than the group without the collation view (Table XVI).

With regards to the interactions found between the topics, those who didn't use the collation were all able to find the interactions between the first two topics with the third one. Only two subjects reported an interaction between topics 1 and 2. Two factors could have contributed to this phenomenon. One, the topic of payment is brand-new. The participants tended to focus more on the new topic while conducting the task. This is seen in the comment given by Subjects 1 and 4. Both mentioned that they stuck mainly with the topic of payment, and tried to find references from there to the other two. Two, as we have mentioned earlier, topics 1 and 2 don't have a relational



interaction as strong and well-defined as that which is between the other pairs. The participants might have made the judgment that they simply didn't interact.

Those who used the collation all reported an interaction between each pair, even between topics 1 and 2. Although the two don't have a strong relational interaction, there was indeed one common excerpt highlighted in both copies. The fact that the subjects saw the overlap explicitly might have induced an impression that the topics must relate. This impression, consequently, might have made the subjects more susceptible to acknowledging the overlap as an interaction than in the previous case.

One unexpected result was given by subject 1, who found a subtle interaction between all three topics: when the dispute of the Helmet of Mambrino, after a long, later turned fierce, practical joke carried out by a group of people in the inn, is finally settled. The barber is paid back for his basin. The innkeeper immediately demands compensation for his loss and services and his request is also fulfilled. Not paying the innkeeper is the trigger of the blanketing. All three threads do connect at this exact locale, although there isn't an explicit area in the highlights overlapping them. Reaching to the interaction requires a looking beyond the text matching. It is interesting to see such serendipity being unveiled only when the participant took one step further than what the technology offered.

Table XV. Behaviors exhibited when the collation view was given to the subjects

	Subject ID					
	2	3	6	8	9	12
The subject has used the collation view and located the overlapping text	√	√	√	√	√	√
The extent that the subject has relied on the collations in fulfilling the task (out of 10)	10	9	10	10	10	10
The importance of having the collation view (out of 10)	10	9	9	10	10	10

Table XVI. Results of the study

Subject ID	Being given the Shared View	Time spent (min)	Time average (min)	Interactions found
1	without	19	17.8	1+3; 2+3; 1+2+3
4	without	9		1+3; 2+3
5	without	16		1+3; 2+3; 1+2
7	without	27		1+3; 2+3
10	without	11		1+3; 2+3; 1+2
11	without	25		1+3; 2+3
2	with	7	10.7	1+2; 1+3; 2+3
3	with	11		1+2; 1+3, 2+3
6	with	9		1+3; 2+3; 1+2
8	with	17		1+3; 2+3; 1+2
9	with	10		1+3; 2+3; 1+2
12	with	10		1+3; 1+2; 2+3

Adoption of technology increases the efficiency of certain types of work, as it reduces the amount of time that would otherwise be spent on manual work. In our study, people with the collation view spent 40% less time than those without it. With the aid of

technical support, the results people reached generally had a high correlation to the affordance of the support itself, as seen in the uniform responses from the participants who used the collation view. People tend to trust and be led by the implementation. If the technology has a biased direction toward the procedures of this research, people would have been seen to be biased in their judgments as well. When the participants were given the option to collate the text, the impression they had from the introduction of the collation function led them to assume that the task was simply a process of matching the highlights. Although some participants who didn't receive such an introduction also had similar impressions, the results show that they were less likely to equal the overlaps with these interactions. When technological assistance is limited and when it depends mainly on manual exploration, as was the case without the collation, the procedure might be slower. At the same time, it is usually under this situation that the serendipities are discovered and appreciated, as demonstrated by the findings of Subject 1.

## 5. AUTOMATIC PINPOINTING

This section of the dissertation is dedicated to reporting the investigations that were conducted, primarily involving the technical or machine side of reading thematically, or in other words, the potential that computers may be programmed to direct the reader to places of their interests. Needless to say, it is a huge topic of its own. As it is unlikely that one technique will suffice to explain it all, research in this area understandably consists of many fine-grained investigations targeting certain focuses. Our focus is inspired by the manual compilation of indices in textbooks and is overlaid on top of the investigation of locating places in a narrative book relevant to a given phrase. This section reports on the approaches we have taken and the results of the investigation [Deng et al. 2007].

The currently available operation of searching for a phrase in digital texts is helpful, but does not satisfy our purposes since it is only able to find consecutive letter-by-letter matches. The resulting set, consequently, is too restrictive to include the majority of the potential pinpoints. On the other hand, searching by matching the individual terms is too imprecise, since it is the alliance of and interaction between all the terms that defines the meaning of the phrase. To overcome the shortcomings of both approaches, we propose a two-step methodology, which basically examines the terms in the phrase first as individuals, then as a group, and combines the evidence collected from both search operations to form a result set that can represent more accurately the potential pinpoints in the target document.

We report our investigation and experiments as follows. Section 5.1 is dedicated to providing a description of the proposed methodology. We present it with an outline of its overall flow and then describe in detail the interactions between the three functional components contained. Following that, we present in Section 5.2 the tests conducted when the segmentation of the text is acquired by following the structural accommodations. The purpose of these experiments is to examine the effectiveness of the overall mechanism of the methodology and to test the effects of several techniques that had the potential to enhance the performance of the general system. Section 5.3 compares the performance of the system when the structural accommodations are taken or not taken into consideration, when segmenting the text. These experiments aim to justify whether or not structural segmentation is beneficial for our application. In the mean time, we tie together the tuning of an important system parameter, the proximity value, into the experiments under both segmentation techniques and report the results. Both Sections 5.2 and 5.3 begin with a description of the three system components implemented under the specific tests, continue to discuss their experimental design, and analyze the test results in the end.

## **5.1 Methodology**

In summary, the goal is to find information pieces in a book relevant to a given phrase. A phrase usually contains several terms with each contributing partially to the overall thematic implication. Here we propose a two-step methodology that uses a relax-and-tighten strategy. In the first step, the methodology relaxes the search by 1) treating the terms as individual entities and temporarily neglecting the interaction between

them, and 2) incorporating expansions to these entities. In the mean time, a parallel search is performed by examining the interaction between the terms. Feedback collected from this search is then used to tighten the results achieved in the first step by filtering out the ones representing no interactions. To efficiently explain the methodology, we first give a landscape description of the overall mechanism. We then break it into three logical components and discuss in detail the individual functionalities and the interactions between the components.

Figure 12 is an overview of the methodology as reflected in the workflow of the experimental system. Here, rectangles represent data and ellipses represent actions. The inputs of the system include a book, which we term as a document, and a phrase. The notion of a document presently refers to a narrative book. Nevertheless, the input is not confined to just one book. It can be easily generalized, with slight customizations, to allow for multiple documents. Within the scope of the current investigation, however, only one book is used because the emphasis of the study is placed on the effectiveness of the methodology rather than on the scalability of the system. The output of the system is the list of target units retrieved. Each unit represents a location whose context is potentially relevant to the topic under request. From the input to the output, the process procedure may be logically decomposed into three components, as suggested in the graph, based on the following functional distinctions: document preparation, query preparation, and search and refinement.

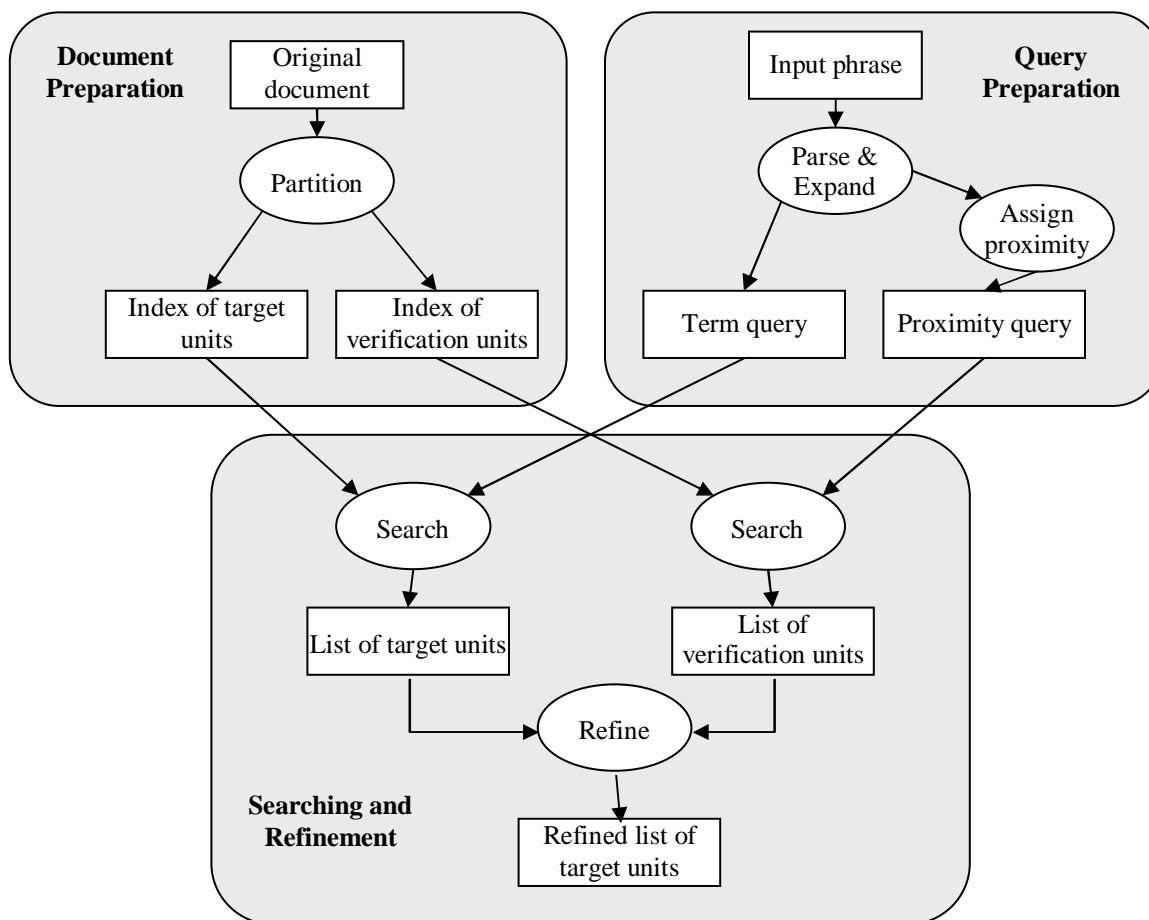


Figure 12. The overview of the methodology

In document preparation, the document goes through a partitioning and indexing process which produces two full-text indexes: an index of the target units and an index of the verification units. The target units are the finer segmentations of the two. They serve as the target locations near where a potential excerpt, coherent to the given theme,

is likely to be found. Compared to verification units, they are smaller chunks of the original text. Verification units are larger segmentations, each of which contains a group of consecutive target units. Taking the segmentation mechanism applied in Section 5.2 as an example, target units are the individual sentences, while verification units the chapters.

In the query preparation, the input phrase is first parsed and expanded to form a term query. We apply a Boolean search of this term query to the index of target units and find all the target units that contain at least one occurrence of any of the terms. This is a relaxed search since the terms are treated as independent entities and their interrelationship is at the present stage disregarded. The goal of the first step is to enhance recall to the fullest extent by loosening the restriction for searching, and by including alternative variants. The second step is a step to tighten the search results by bringing the interrelationship between the terms back into consideration. To do that, the query preparation first uses the terms after the parsing and expansion to form alternative phrases, then imposes a proximity value (the number of terms apart) on each one. This procedure produces a series of proximity queries which are finally conjugated with a Boolean *or*. The final query is searched in the index of verification units. As a result, a list of satisfied verification units is acquired.

The third component is the search and refinement stage. There the retrieved target entries are filtered through the qualified verification instances. If a target unit does not belong to any verification unit that passes the proximity check, it is proof that the target only sees partial thematic clues and does not provide strong enough evidence to



qualify itself as an eligible return. It is, therefore, eliminated from the final list of retrieval.

Consider as a metaphor, the case of a detective trying to figure out whether a squad has had a secret meeting. The detective first sets out to question each member regarding their specific activities. Then he checks to see if, one day, the squad happened to appear in the same building. Although during that day they might have done things alone, like going to different lunches, it is highly possible that during that same day a secret meeting occurred. All the activities in that day by each individual deserve to be more closely investigated. On the contrary, if during the other days, the members didn't even appear close, it might be that a secret meeting is very unlikely and any evidence from those days can be easily ignored. In our case, the list of target units contains the specific activities by each individual, and the list of verification units gives the days when the members came close to one another. The evidence from the list of verification units is used to filter out those activities that happened on the days other than the ones on the list.

All the experiments presented in the Sections 5.2 and 5.3 are implemented in Java, with the search engine built on top of the open source full-text search API Apache Lucene [Apache Lucene 2008].

## **5.2 Tests under Structural Segmentation**

In this section, the original text is segmented following certain structural features. Here we test the effectiveness of our query expansion by using WordNet synonyms and inflectional variants; we also test the effectiveness of the overall methodology.

## 5.2.1 *Implementation of the System Components*

### 5.2.1.1 Document Preparation

We chose sentences and chapters to be our two units of information. The small unit is the basic unit for judging the relationship with a given phrase and, consequently, it pinpoints places in the book where the working theme discretely emerges. Sentences are a feasible choice for the small unit as they provide meaningful descriptions and are autonomous entities. Chapters are large information chunks that are pre-delimited by the author for logical or structural purposes. We chose them to be the large unit. The decisions regarding the definitions of the units, however, are not claimed to be the most authoritative or effective. As a matter of fact, Section 5.3 contains a further exploration into the effects of using different unit definitions, with or without considering the structural confinement.

Rudimentary heuristics were used to decide sentence delimitations in our text. A sentence was simply considered to be a string with a terminal character that is one of the typical end-of-sentence delimiters, for instance, ‘?’, ‘!’, or ‘.’. Chapter boundaries are encoded in the text with the word “chapter,” followed by a chapter number, and thus can be easily detected. Using Lucene, we constructed the two separate unstemmed indexes with the stop words removed.

### 5.2.1.2 Query Preparation

The basic query for the Lucene document retriever is a conjunction of clauses. The input phrase first goes through a query parser which does two things: removes the stop words and detaches the inflectional endings from a word. We followed the rules of

detachment [WordNet Manual 2008] applied in WordNet, and in the mean time used the test book as a dictionary of possible correct words, in order to find the base form of a word. The query parser generates the baseline query, which contains the base forms of the non-stop words claused together by a Boolean *or*. When forming a query with expansion, we first look for all the possible variants derived from the expansion, then trim off the ones that don't appear in the book, and finally replace each term in the baseline with a clause that joins together the original term and its variants using a Boolean *or*. We experimented with the expansion using WordNet, inflectional morphology, and a combination of both. As a result four queries, represented as S1, S2, S3, and S4, were tested on the sentence-wise retrieval.

**S1. Baseline:** The baseline query contains no morphological or synonym expansions of any sort. Query terms simply consist of the base forms of the non-stop words from the natural language input.

**S2. Inflection expanded:** The inflectional variants of each term are added to the baseline query. Inflectional variants include all possible plural forms and verb tense conjugations.

**S3. WordNet expanded:** Terms in the baseline query are expanded by the synonyms found in WordNet.

**S4. WordNet & Inflection expanded:** The synonyms from Wordnet and the inflectional variants are added together to the baseline query.

To form the queries applied in chapter-wise searching, disjunctive phrases were first constructed by selecting one alternative, either a variant or the original form, for

each term. Following that, a proximity value was applied to each individual phrase before they were conjoined together by a Boolean *or* to form the final query. Depending upon what was included in the sets of alternatives, three queries were defined and tested. To differentiate them from the ones above, we represent them as C1, C2, and C3.

**C1. Proximity imposed but with no variants included:** The proximity value is applied to the baseline query.

**C2. Proximity imposed after inflection expansion:** Alternative phrases are formed from alternative terms composed of the originals and the inflectional variants.

**C3. Proximity imposed after WordNet & Inflection expansion:** Alternative phrases are formed from alternative terms composed of the originals, the WordNet synonyms, and the inflectional variants.

After we examined the results from the sentence-level retrieval, the expansion by WordNet achieved no improvement in the number of relevant documents. Therefore, WordNet synonyms were not carried out as a single test in proximity checking.

The proximity value applied in the current study is 300 words apart. As the use of the proximity verification is based on the theoretical assumption that if all the terms in a given phrase co-occur within a particular vicinity, then the location of the occurrence is considered to be of significance. This is analogous to a passage retrieval that defines the passage by a sliding window. The determination of the proximity value is essentially a determination of an effective passage size. Studies conducted by Callan [Callan 1994] have previously shown that if a sliding window of a fixed length is chosen for passage retrieval, sizes of 150-300 words yield the best results. We, therefore, chose 300 as the

proximity value for our study. It worked well in the present experiment; more systematic tuning on this parameter is conducted in Section 5.3.

As a concrete example, consider the phrase “Sancho’s blanketing” as the input.

The following shows the seven queries generated from this phrase.

S1. sancho blanket

S2. sancho blanket blanketing blanketed blankets

S3. sancho blanket mantle cover

S4. sancho blanket mantle cover blanketing blanketed blankets

C1. “sancho blanket”~300

C2. “sancho blanket”~300 or “sancho blanketing”~300 or “sancho blanketed”~300 or “sancho blankets”~300

C3. “sancho blanket”~300 or “sancho mantle”~300 or “sancho cover”~300 or “sancho blanketing”~300 or “sancho blanketed”~300 or “sancho blankets”~300

### 5.2.1.3 Searching and Refinement

Using Lucene, we applied a Boolean search of the four sentence queries to the sentence index and retrieved lists of potentially relevant sentences. The ranking of the returns are ignored due to two reasons: 1) navigating a narrative thematically tends to follow the order in which the excerpts appear in the text, and consequently the ranking sequence becomes undesirable; 2) a typical case for a ranked search is based on the assumption that the relevant documents could and, whenever applicable, should be sorted according to a scale of relevance. Our application, on the other hand, focuses on

forming a complete relay of thematically connected segments, and all relevant excerpts are equally valued.

Recall-enhancing strategies tend to have the adverse effect of reducing precision. In other words, increasing the number of relevant documents is generally achieved at the cost of including more irrelevant returns. There are two factors in our study that aim at enhancing recall. One, a Boolean search is less strict than a ranked search that cuts off at a given number or by a certain criteria, since a Boolean search selects all the documents that contain at least one occurrence of any of the terms. Two, a query expansion, in particular, by predicting the semantic meaning of a term using a general-purpose thesaurus, may add bad guesses and, consequently, produce more unwanted returns.

In order to counteract such an effect, on the one hand, we increase the precision of the Boolean search by maintaining a list of frequently appearing terms and ignoring them when extracting the sentences. For instance, *Don Quixote* and *Sancho Panza*, being the names of the two principal characters in the book, are neglected in the sentence search because they do not add value as stand alone terms. This strategy is analogous to the 50% threshold implemented in a MySQL full-text search [MySQL Documentation 2008] – if a word is present in at least 50% of the entries in a data set, then the word is considered to be a stop word and is subsequently left out of the search. Nevertheless, in checking for proximity at the chapter level, the list is not applied. The frequent terms, though by themselves not informative, could add great value to defining a concept when combined with other artifacts.

More importantly, in order to refine the search results, we introduced a simple dropping strategy that uses the evidence collected at the chapter level. For each chapter, we verify whether the phrase terms appear within a certain distance from each other. If not satisfied, we regard the absence as proof that the content from that chapter is unlikely to be relevant and filter out the sentences from that chapter. Exploiting co-occurring terms in IR systems [Billhardt et al. 2002] has proven to be beneficial in document retrieval. This demonstrates that a term's proximity is a valid clue for judging thematic or semantic coherence.

### 5.2.2 *Experimental Design*

The phrases for the test were selected from the taxonomy [Urbina et al. 2006] composed by the Cervantes Project [Cervantes Project 2008] as a comprehensive summary of the episodic and thematic elements of the narrative. The taxonomy has been used to catalogue a digital archive of *Quixote*-related illustrations. Figure 13 shows a few examples from that taxonomy. When selecting test phrases, short phrases that by themselves could reveal solid episodic clues were preferred. For example, “the helmet of Mambrino” appears in the taxonomy as the entry “Adventure of the helmet of Mambrino,” and thus was chosen as a test phrase because the phrase clearly describes one adventure. Using this strategy, five phrases were selected from the taxonomy by extracting the defining part of an individual adventure or episode.

```

18. Chapter 18
18.1 Adventure of the flock of sheep (rebaños)
18.2 SP counting DQ's teeth

19. Chapter 19
19.1 Adventure of the dead body
19.2 SP names DQ, knight of the sad countenance (triste figura)

20. Chapter 20
20.1 Adventure of the fulling mills (batanes)
20.2 Torralba's tale by SP
20.3 SP's tricks DQ
20.4 SP's mishap (cagada)

21. Chapter 21
21.1 Discourse of DQ about ideal knight
21.2 Encounter with barber
21.3 Adventure of the helmet of Mambrino

```

Figure 13. Some examples from the taxonomy

**P1.** Sancho's blanketing

**P2.** The helmet of Mambrino

**P3.** Sancho's island

**P4.** The fulling mill

**P5.** The galley slave

For each phrase, we collected the search results for all four sentence queries. Since S4 encompasses the other three queries, the results from S1, S2, and S3 are all a subset of the results from S4. A Cervantes scholar then went through the list generated by S4, and judged one by one whether the surrounding context of the sentence was relevant to the given theme. This scholar, who regularly teaches graduate-level courses in Spanish literature, is an experienced editor, the Director of the Cátedra Cervantes in Spain, and has written several monographs and over 100 articles related to Cervantes



and *Quixote*. His expertise and strong background qualifies him for making authoritative or editorial decisions for this study. As people's perception and preference of reading narrative themes differ, it will be interesting, as a future-work evaluation, to have a larger pool of judges and to see how their judgment reflects their perceptions. The current study, however, relies only on this expert's opinion, because the focus is on testing the methodology. The results are given and analyzed in the following section.

### 5.2.3 Results and Discussion

To evaluate the retrieval performance, IR systems typically use the well-established precision and recall metrics. The standard recall measure, however, is hard to apply directly to our application. Unlike the standard IR test collections, e.g., TREC, which gives the total number of relevant documents in the set as a pre-determined parameter, it is extremely difficult, if not impossible, to know exactly how many pinpoints will be considered relevant to a theme in a book. Instead of using absolute recall, which calculates the ratio of relevant documents returned out of the total relevant documents present, we use the largest number of relevant documents known to be retrieved as the base, and compute a recall relative to it. This largest possible number is actually the number of relevant returns retrieved by S4.

$$\text{Relative recall} = \frac{\text{Number of relevant documents retrieved}}{\text{The maximum number of relevant documents ever retrieved}}$$

The total number of relevant documents is the sum of the maximum number of documents ever retrieved and the number of documents relevant but missed in the extraction. Therefore, absolute recall and relative recall share exactly the same tendency and will produce the same judgment if they are applied to a comparison between two mechanisms of retrieval. Since the purpose of our quantitative calculation is to compare across alternative mechanisms, and not to judge the absolute performance of one single mechanism, the use of relative recall suffices.

In terms of the overall performance evaluation, we combine precision and recall with the F-Measure [Makhoul et al. 1999; Van Rijbergen 1979], which is simply a weighted harmonic mean. The weight  $\alpha$  represents the relative importance of precision and recall for a particular application. We use  $\alpha = 2$ , i.e., the  $F_2$  measure, for evaluating the effectiveness of both the retrieval and the procedure as a whole (retrieval together with refinement). The  $F_2$  measure weighs the recall twice as much as precision. Under the context of thematic reading, the effect of encountering pieces of information unrelated to what you are expecting is not as detrimental as not finding what you are looking for. In other words, recall should be valued higher than precision. However, for the evaluation of the refining methods,  $\alpha = 1$ , i.e., the  $F_1$  measure, is used. In the  $F_1$  measure, recall and precision are evenly weighted. The reason for choosing the  $F_1$  measure in that single test is that the test's focus is specifically on evaluating precision-enhancing techniques. It is understandable that if you relax the criteria for filtering, you are more likely to preserve more records, and thus achieve a higher recall. Therefore, in order to have an objective and honest judgment regarding the effectiveness of precision-

enhancement, it is better to treat precision as being at least as important as recall in a comparison of the refinement methods. The F-measure is defined in the following formula, where  $\alpha = 1$  for  $F_1$  and 2 for  $F_2$ .

$$F_{\alpha} = (1 + \alpha) \cdot \textit{precision} \cdot \textit{recall} / (\alpha \cdot \textit{precision} + \textit{recall})$$

### 5.2.3.1 Comparison of the Expansion Methods

Table XVII compares the retrieval effectiveness of the different expansion methods. The table contains the number of relevant and irrelevant sentences returned by all four queries for each phrase.

Figure 14 shows a graph of the performance. As seen in the figure, expanding the baseline by WordNet does not improve the performance. As we discussed in Section 5.2, the effectiveness of using WordNet synsets for query expansion has been controversial. The majority of the research findings lean toward the claim that if WordNet is applied simply for expanding the query, the retrieval performance will be downgraded. Our results confirm the claim made by the majority and prove that in the context of indexing the narrative texts by short thematic phrases, expanding by WordNet provides little benefit. If we examine closely one of the test phrases, *Sancho's island*, it gives us some insight into why WordNet turned out to be unsuccessful in our case. Basically, from this phrase, we expect to see all the incidents that are relevant to Don Quixote's promise of an island, Sancho's eagerness for the island, his constant reminders to his master about the island, and the master's reprisals regarding the issue. The book used both *island* and

*insula* in related occasions. The word *insula*, however, is not included as a synonym in WordNet. In another test case, *ship* is recognized as a synonym for *galley*. Nevertheless, the book exclusively used *galley* in referring to the specific adventure of *the galley slave*. Returns added by *ship* all turned out to be unrelated. In summary, the disconnection caused by the idiosyncratic need of a particular context may be the major reason why a general thesaurus-based query expansion turns out to be unhelpful, as has also been observed by other researchers [Srinivasan et al. 2000]. On the other hand, the fact that the test book used is as a translation from its original language, Spanish, also makes it difficult to find appropriate synonyms (as in the case of *insula*) by a lexicon in canonical English.

For three phrases, inflection expansion significantly improved the performance over the baseline. For the other two, it only added a negligible number of irrelevant returns. It also outperformed its combination with WordNet. We conclude that expansion with inflectional morphological variants is essential for locating candidates for a thematic subject.

Table XVII. Search results of the different expansion methods

Phrase tested	Query used	Number of relevant returns	Number of irrelevant returns	Precision	Relative recall	F <sub>2</sub> measure
P1	S1	10	1	0.9091	0.5000	0.5882
	S2	20	2	0.9091	1.0000	0.9677
	S3	10	15	0.4000	0.5000	0.4615
	S4	20	16	0.5556	1.0000	0.7895
P2	S1	29	19	0.6042	1.0000	0.8208
	S2	29	20	0.5918	1.0000	0.8131
	S3	29	19	0.6042	1.0000	0.8208
	S4	29	20	0.5918	1.0000	0.8131
P3	S1	18	5	0.7826	0.6923	0.7200
	S2	26	5	0.8387	1.0000	0.9398
	S3	18	5	0.7826	0.6923	0.7200
	S4	26	5	0.8387	1.0000	0.9398
P4	S1	11	7	0.6111	1.0000	0.8250
	S2	11	10	0.5238	1.0000	0.7674
	S3	11	8	0.5789	1.0000	0.8049
	S4	11	11	0.5000	1.0000	0.7500
P5	S1	22	15	0.5946	0.6667	0.6408
	S2	33	22	0.6000	1.0000	0.8182
	S3	22	23	0.4889	0.6667	0.5946
	S4	33	29	0.5323	1.0000	0.7734

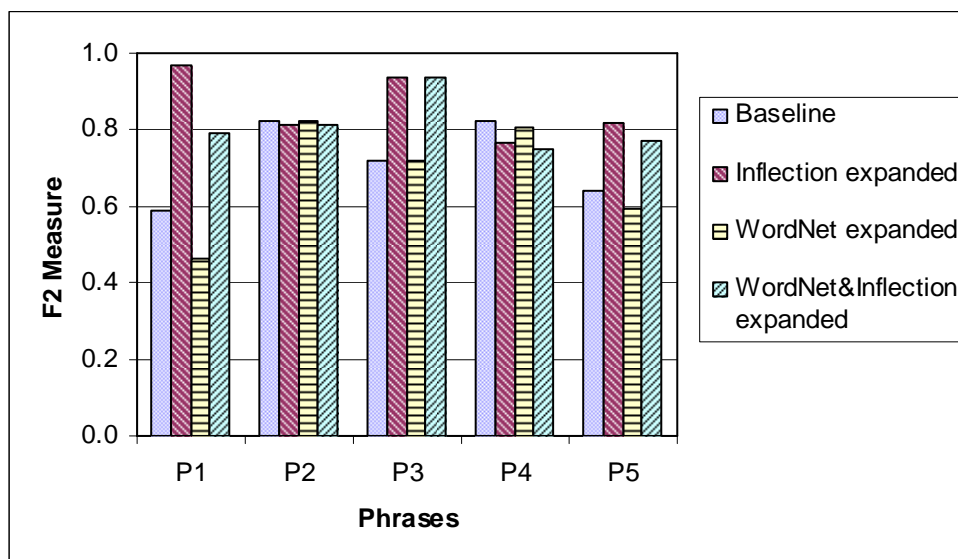


Figure 14.  $F_2$  measure of the different expansion methods

### 5.2.3.2 Comparison of the Refining Methods

To examine the effectiveness of the refining methods, we applied the three proximity checks to the result sets returned by S4. The fact that the results from S4 contained the highest possible number of noises made it the best sample for filter testing, compared to the results returned by S1, S2, and S3. The goal of the refining process is to preserve as many relevant returns and filter out as many of the unneeded ones as possible. The results are given in Table XVIII and Figure 15. As we see in Figure 15, refinement done through the proximity check by C2 stands out as the best.

Table XVIII. Refined search results from S4 after different refinement methods

Phrase tested	Filter used	Largest number of relevant returns	Number of relevant returns	Number of irrelevant returns	Precision	Relative recall	F <sub>1</sub> measure
P1	C1	20	14	4	0.7778	0.7000	0.7368
	C2		20	5	0.8000	1.0000	0.8889
	C3		20	11	0.6452	1.0000	0.7843
P2	C1	29	28	7	0.8000	0.9655	0.8750
	C2		28	7	0.8000	0.9655	0.8750
	C3		28	7	0.8000	0.9655	0.8750
P3	C1	26	23	1	0.9583	0.8846	0.9200
	C2		26	1	0.9630	1.0000	0.9811
	C3		26	1	0.9630	1.0000	0.9811
P4	C1	11	4	0	1.0000	0.3636	0.5333
	C2		11	0	1.0000	1.0000	1.0000
	C3		11	0	1.0000	1.0000	1.0000
P5	C1	33	26	6	0.8125	0.7879	0.8000
	C2		33	8	0.8049	1.0000	0.8919
	C3		33	17	0.6600	1.0000	0.7952

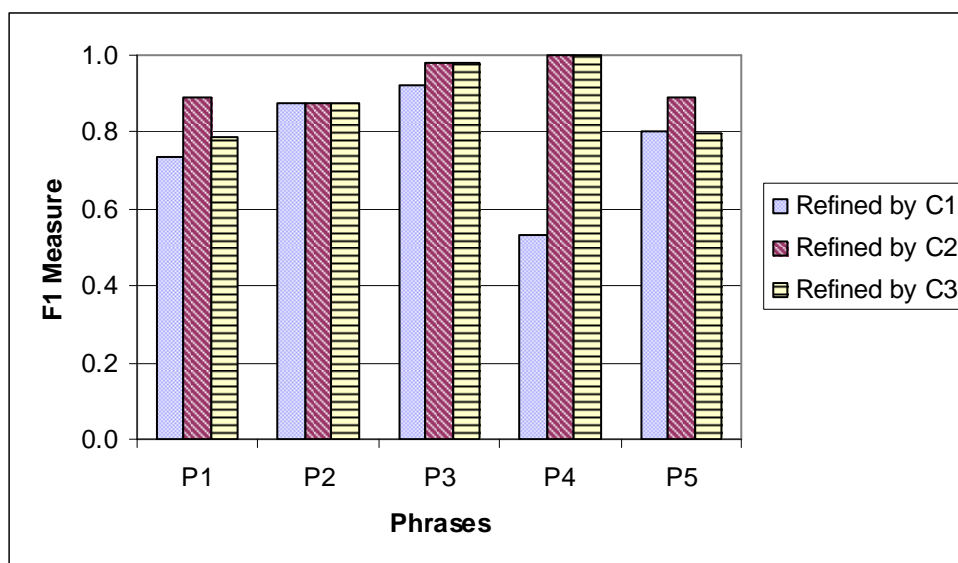


Figure 15.  $F_1$  measure of the refined search results from S4 after different refinement methods

One issue that should draw our attention is the filtering out of relevant returns, as in the case of P2. The one return that was incorrectly recognized as irrelevant by the refinement module says:

“...and all the rest of it that your worship swore to observe until you had taken that helmet of Malandrino’s, or whatever the Moor is called, for I do not very well remember...” [Cervantes Saavedra 1885].

The theme at issue is *the helmet of Mambrino*, and it is interesting that the reason the above pinpoint was incorrectly dropped is due to our character Sancho’s absurd habit of creating and using incorrect words. To cope with situations like this, word co-occurrence might be helpful, and conducting experiments on co-occurrence is part of our future work investigation.



Another interesting phenomenon is that if we apply the Lucene-ranked search (which uses the standard  $tf*idf$  model) on C1 for four of the phrases, the chapter number given in the taxonomy that indicates the place where the major incident of the adventure happens was ranked as number one. For the other phrase, the chapter number was ranked number two. C2 and C3 were able to extract those chapter numbers within the top three returns. This indicates that the vocabulary used in the taxonomy was very carefully chosen and is representative of the contents.

### 5.2.3.3 Examination of the Overall Methodology

The comparisons from the above sections suggest two things. 1) Expanding the input using inflectional morphological variants enhances the retrieval performance. 2) If the input contains multiple terms, the results will be further improved by going through a filtering process that uses the chapter-level evidence gathered by a proximity check on the possible phrases after the inflection expansion. Guided by these suggestions, we then compare the results of the two steps, together with the baseline, in Table XIX and Figure 16. The comparison between S1 and S2 has already been discussed in Section 5.2.3.1. From Figure 16, for all five phrases, the results gained by S2 and filtered by C2 (represented as S2 and C2) shows a performance better than both the baseline and the S2 alone. In particular, the performance is significantly higher than the baseline.

In summary, our study shows that it is recommended to expand the input phrase by inflection, and if there are multiple terms present, the result should be further refined by going through the filtering procedure.

Table XIX. Final result comparison of the suggested methods

Phrase tested	Result set	Largest number of relevant returns	Number of relevant returns	Number of irrelevant returns	Precision	Relative recall	F <sub>2</sub> measure
P1	S1	20	10	1	0.9091	0.5000	0.5882
	S2		20	2	0.9091	1.0000	0.9677
	S2 & C2		20	2	0.9091	1.0000	0.9677
P2	S1	29	29	19	0.6042	1.0000	0.8208
	S2		29	20	0.5918	1.0000	0.8131
	S2 & C2		28	7	0.8000	0.9655	0.9032
P3	S1	26	18	5	0.7826	0.6923	0.7200
	S2		26	5	0.8387	1.0000	0.9398
	S2 & C2		26	1	0.9630	1.0000	0.9873
P4	S1	11	11	7	0.6111	1.0000	0.8250
	S2		11	10	0.5238	1.0000	0.7674
	S2 & C2		11	0	1.0000	1.0000	1.0000
P5	S1	33	22	15	0.5946	0.6667	0.6408
	S2		33	22	0.6000	1.0000	0.8182
	S2 & C2		33	9	0.7857	1.0000	0.9167

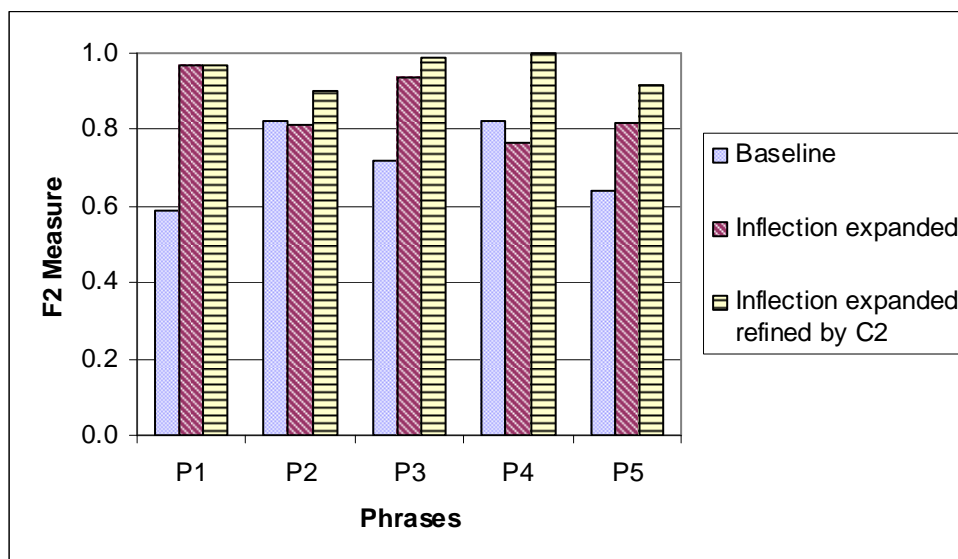


Figure 16.  $F_2$  measure comparison of the suggested methods

### 5.3 Tests on Segmentation and Proximity Tuning

In this section, we compare the performance difference when the text is segmented using two different techniques: one follows the structural features of the original book, as presented in Section 5.2; the other uses a sliding window and completely ignores the structural boundaries. Sliding windows [Kaszkiel and Zobel 1997] have proven to be applicable to passage retrieval in the experiments using standard IR test collections. Verifications are needed to prove that it is also applicable in our study where documents could be as small as sentence size. We also explore further into the effect of adopting different proximity values and hope to find a premium value that applies in general.

### 5.3.1 *Implementation of the System Components*

#### 5.3.1.1 Document Preparation

In our earlier experiments, we used one scheme of unit definition to conduct a preliminary study on the feasibility and effectiveness of our methodology. Here, we reuse this same scheme. Additionally, we introduce a new scheme and issue a parallel comparison between them. These two rules of unit definition are as follows, and are represented as R1 and R2. Under both schemes, a variety of the proximity values are tested. The increase in this parameter has the tendency to improve the performance in the beginning, as it moves from highly to rightly restrictive. Afterwards, a decrease in the performance may be seen as it may become completely unrestrictive when the range is too wide to function as an effective proximity check. We aim to find a roughly premium value for our methodology.

**R1. By structural accommodations.** The target units and the verification units are defined as sentences and chapters, respectively. Detailed information on the delimitation characteristics of the sentences and chapters is given in Section 5.2.1.1.

Under R1, if each term of an alternative phrase appears once within a specific distance in a chapter, all the target instances from that chapter are deemed to be relevant. It is reasonable to assume that the target units close to the locale of the occurrence are likely to be relevant. But if the chapter is lengthy, and there is only one combination of the term occurrence that meets the proximity check, it could be too generous to grant the passport to every extracted target entry of that chapter. Driven by this reasoning, we contrived another test scheme. This scheme, represented as R2, is designed with the goal

of letting through only those target instances that are within a certain distance of a qualified verification.

**R2. By sliding windows.** This rule completely ignores the punctuation and semantic structure of the original text. It uses a sliding window of term counts to shape both units. The target unit is determined by 50 consecutive non-stop terms, and there is no overlapping in adjacent units. The verification unit is also decided by a certain number of terms, but the sliding window only forwards half its size each time. In other words, each verification unit shares half of its window with the previous unit, and half with the successive unit. Figure 17 demonstrates the scheme when the verification unit is defined as 300 terms, and thus encloses six target units.

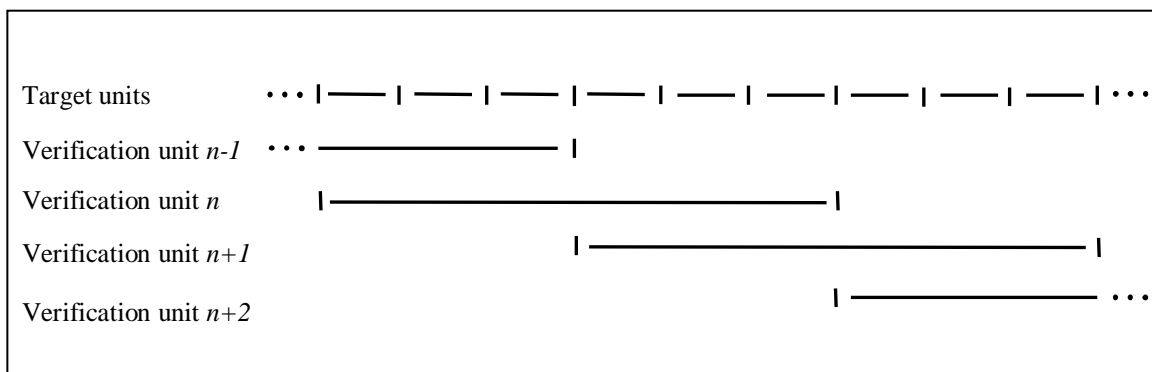


Figure 17. Sliding window demonstration when verification unit encloses 6 target units

In total, four indices are generated; two for each rule. All indices are unstemmed, with stop words removed, and constructed using the Lucene index writer.

### 5.3.1.2 Query Preparation

In query preparation, the query parsing, the same procedure as described in Section 5.2.1.2, yields a set of base terms. Afterwards, for each term in the set, we form its inflectional variants and add to the set those variants that appear in the book. The inflectional variants include all possible plural forms and verb tense conjugations. With the base terms and their variants ready, we construct three queries, namely Q1, Q2, and Q3. Q1 is a query that contains merely disjointed terms that will be used to search through the indices of target units under both R1 and R2. Q2 and Q3 are two queries formed to be searched on the verification units, one for each scheme. To form these queries, disjunctive phrases are first constructed by selecting one alternative, either a variant or the original form, for each term. Then for R1, a proximity value is explicitly assigned to each of the disjunctive phrases. For R2, however, no explicit proximity value is needed since it automatically inherits one from the term count that forms the verification units. When verification units are defined by 300 terms, the proximity boundary that determines how far away the terms are allowed to be from each other is implicitly 300. Finally, the disjunctive phrases are conjoined together by a Boolean *or*.

**Q1. Query with disjointed terms.** The query contains all of the non-stop terms of the original phrase and their present morphological variants. All of the terms are conjugated with a Boolean *or*.

**Q2. Query with proximity explicitly imposed.** The query has proximity values explicitly imposed on each of the alternative phrases and will be applied to the index of the verification units in R1.

**Q3. Query with proximity implicitly derived.** This query doesn't have an explicit proximity value assigned to the alternative phrases and is used to acquire a list of verification entries in R2. As R2 uses a certain number of terms to decide the size of the verification units, the proximity value is implicitly equivalent to that specific size.

Consider as a concrete example that if the phrase "the galley slaves" is entered as the input, the following three queries are constructed.

Q1. *galley or slave or galleys or slaves*

Q2. *"galley slave"~300 or "galleys slaves"~300 or "galley slaves"~300 or "galleys slave"~300*

Q3. *"galley slave" or "galleys slaves" or "galley slaves" or "galleys slave"*

### 5.3.1.3 Searching and Refinement

Using Lucene, we apply the Boolean search of Q1 to the index of target units and retrieved lists of potentially relevant target entries. We then search Q2 on the verification units in R1 and Q3, for those in R2. The general idea of using the list of qualified verification units to filter the initial target units is the same as that which is described in Section 5.2.1.3. Other relevant operations applied here are also the same as Section 5.2.1.3, which can be referred to for details such as maintaining a list of frequently appearing terms only in the extraction of target units.

### 5.3.2 *Experimental Design*

The present study uses the same five phrases as in Section 5.2. For each phrase, we collected the search results by performing the following tests.

- Applied Q1 and Q2 on R1's indices of target units and verification units, respectively. The forming of Q2 used four different proximity values: 5, 50, 100, and 300.
- Applied Q1 and Q3 on R2's indices of target units and verification units, respectively. Four proximity values, i.e., four different sizes of verification units, were tested: 300, 400, 500, and 600.

The unfiltered lists returned by Q1 were already judged by an experienced Cervantes scholar, as described in Section 5.2.3, and were reused here. The purpose of the target units is to present the occurrences of the terms within a context. Since the total number of occurrences of the terms, both the original and the expanded, is constant, the choice of their contexts may slightly change the number of returns. Therefore, under the two rules, it is likely that the total number of target returns differs. If one long sentence contains multiple occurrences of the terms, it is possible that that sentence may show up as multiple entries in the target list returned in R2. There might also be cases where multiple term occurrences are close to each other but are separated by a sentence delimiter. It is possible that these multiple sentences are enclosed by only one return in R2. Nevertheless, the way the methodology uses the target unit makes its definition uncritical. Since we judge the relevance of the target based on its adjacent context, and not precisely on the target itself, the differences in the number of returns under the two schemes can be normalized to one scale by using the number of contexts, as opposed to the number of targets. We perform this mapping in Section 5.3.3.3, at a point when the



comparison was between the two schemes. When the schemes are examined as standalone applications and the analysis is within each scheme, mapping is unnecessary.

### 5.3.3 *Results and Discussion*

#### 5.3.3.1 Proximity Tuning in R1

The effectiveness of a filter is generally affected by the strictness of the checking criteria it uses. Intuitively, in the case of proximity verification, the further apart the terms can be, the more likely it is that the irrelevant returns are prevented. On the other hand, being too restrictive may, in turn, eliminate useful retrievals and only return those that are the most tightly coherent. There is always a balance between finding all of the desirable returns, while trying to restrict the number of noises that may be included. In order to get close to a premium balance, we tested four values of the proximity parameter: 5, 50, 100, and 300. We first obtained the intermediary list of targets by applying Q1 to the target units in R1. Then, by applying Q2 with different proximity values to the index of the verification units, we received four lists of verification retrievals. The list of targets then went through separate elimination processes, judging them against each of the four verification lists. The results are presented in Table XX.

For three of the five phrases, P2, P4, and P5, the proximity value didn't matter. Apparently, the terms must have had one instance that appeared less than five words apart in the relevant chapters. For P1 and P3, a proximity value of 100 stood out the best. Although it still requires further and larger scale testing to be more conclusive, our test results do show that when sentences and chapters are used as the target and verification units, a proximity value of 100 to 300 has the potential to achieve a good performance.

Dropping below 100 is likely to eliminate useful returns, and above 300 only adds more irrelevant returns. This result coincides with Callan’s findings [Callan 1994] when he experimented using a sliding window to determine a premium passage size in passage retrieval. He concluded that passages of 150-300 words provide consistently positive results across a variety of test collections. Finding a proper passage size in passage retrieval is similar to the proximity check in our study, in the sense that they both target premium retrieval performance by adjusting the size of the document in the search.

Table XX. Search results at different proximity values in R1

Proximity value applied in filtration	Number of relevant returns					Number of irrelevant returns				
	P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
<b>5</b>	12	28	9	11	33	0	7	0	0	9
<b>50</b>	19	28	25	11	33	2	7	0	0	9
<b>100</b>	20	28	26	11	33	2	7	0	0	9
<b>300</b>	20	28	26	11	33	2	7	1	0	9

### 5.3.3.2 Proximity Tuning in R2

In the previous refinement method, we merely mandated that the occurrence of the phrase terms appear at least once within a specified distance in a chapter. If the condition is met, all the target units that are extracted and belong to this chapter are considered relevant. This raises the issue that the strategy might be too generous to let pass those target instances that appear far away from the location where the verification is fulfilled. R2 was designed with a goal to let pass only those targets that are near a

satisfactory verification. We began the testing of R2 with a proximity value of 300. With four phrases, our result showed that eligible entries were eliminated incorrectly. We then enlarged the proximity up to 400, 500, and 600 in order to retain more relevant returns. The results are given in Table XXI.

Contrary to our prediction, the filtration performance was inconsistent among the test phrases used in R2. The proximity value didn't have a significant impact for P1 and P3. For P4, a value of 500 was required in order to preserve all the relevant returns. For P5, the value had to go up to 600 to achieve consistent relevant returns. For P2, even with a proximity as high as 600, there were still seven desirable returns wrongly filtered out.

Table XXI. Search results at different proximity values in R2

Proximity value applied in filtration	Number of relevant returns					Number of irrelevant returns				
	P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
<b>300</b>	23	20	27	11	29	2	7	0	0	8
<b>400</b>	23	20	27	11	30	2	8	1	0	8
<b>500</b>	23	24	27	12	31	2	8	1	0	9
<b>600</b>	23	25	27	12	32	2	8	1	0	10

### 5.3.3.3 Comparison between R1 and R2

The effectiveness of the two test schemes was contradictory. In R1, P1 and P3 were the phrases that displayed a rise-and-fall pattern, while the proximity value was gradually increased. At the low end, relevant entries were eliminated incorrectly; at the

high end, more irrelevant entries were returned. P2, P4, and P5 achieved consistently high performances given a wide range of proximity assignments. R2, on the contrary, worked well on P1 and P3, but the performance was inconsistent with the other three phrases.

A closer examination on the phrases explains why. P2, P4, and P5 are, as we may call them, idiomatic phrases. The binding of the terms is crucial in the determination of a clear and precise connotation. In order to convey to the readers a definitive thematic meaning, the terms have to appear at least once, and close to one another, within a certain length of context. In the case of *Quixote*, the context, as demonstrated by our findings, is predictably in the form of a chapter. Within that context, the terms might appear by themselves given the assumption that the reader has established a sense of the thematic hint underneath. The other two phrases are, however, less idiomatic. They are more like enumerations of the keywords of an episode or event. The terms are most likely not seen right next to each other. Nevertheless, it is necessary for them to appear close enough to enable a thematic connection. They are phrases that are not critically dependent upon hinted thematic clues, that will therefore achieve consistent performance even when the document's structural features are not taken into consideration, as in R2. Our choice of phrases successfully presented the two extremes of the possible types of phrase inputs: one type requires the terms to stick together like a single entity in order to represent an idiosyncratic meaning; the other is simply an enumeration of key terms.

We further normalize the results, and juxtapose the numbers from the two schemes side by side, in order to compare the two schemes in parallel. The results are

given in Table XXII. This time, we cross-referenced the target units between the two and normalized the total number of target retrievals according to the number of contexts. For instance, if one long sentence showed up as two retrievals in R2, the two retrievals were merged into one. After normalization, the total numbers displayed were slightly different from the ones we gave in the previous sections. A more detailed explanation of the normalization appears in Section 5.3.2.

Table XXII. Final results comparison between R1 and R2

Test method	Number of relevant returns					Number of irrelevant returns				
	P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
<b>R1: 100</b>	20	28	26	11	32	2	7	0	0	9
<b>R2: 600</b>	20	22	26	11	32	2	7	1	0	9
<b>Without filtration</b>	20	29	26	11	32	2	20	5	10	22

Our findings suggest that using a sliding window is not an effective way to verify and locate thematic pinpoints, in particular when the phrases under examination are instances of idiomatic phrases. The filtration performance was not definitive and it is difficult to find a proper proximity value to reach a justifiably consistent performance. R1 was able to achieve good results with all five phrases within a reasonable range of proximity. R2's performance in P2 and P3 are still not as good as in R1, even when a much higher proximity is allowed. Most critically, R2 was not able to preserve all the

relevant returns in P2. Under the context of thematic reading, it is crucial to find all of the thematic instances so that a full-bodied sub-story may form.

## 6. OPEN ISSUES AND FUTURE WORK

This dissertation is no more than a first step on a journey of a thousand miles. First of all, there is much to be collected regarding the behavior of thematic readers. We have used in our studies mock-up scenarios and prototype systems. The use of an experimental setting to collect user's behavior, no matter how much we try to represent the real situation, is never going to be the same as what happens in the real world. For thematic reading, the most desirable way of conducting observations would be by looking over the shoulders of real readers, even without their conscious awareness of being in a study, while they read and while the thematic reading spontaneously takes place. Conducting reading in a thematic pattern often happens in an informal or casual setting, for instance during leisure reading. Free form approaches can better tell us what is indeed expected by the user, what truly has aroused frustration, and how the initial interaction with the document is triggered. It might be hard to conduct a study with hidden cameras to watch how people read, but approaches such as free form survey without the impact of any system or designed task will help us further approach the real issue, and understand the prevalence of the phenomena.

There are also significant improvements that can be added to the Reader's Interface. The participants have already offered a few concrete recommendations in this regard, in their survey feedback. We have included them in the following list of future work.

1. The current design of the bidirectional traversal needs to be rendered with better consistency. Both directions should follow one common mechanism, such as clicking on

the text directly, or explicitly pressing a button. Mixing different operations for one common purpose but in different direction tends to puzzle users. The moment that users stop working to wonder why the reaction of the system is not what it is assumed to be, is the moment they encounter the fourth wall [Murray 1997] that breaks the immersive experience.

2. The participants have also expressed a preference towards the presence of contextual awareness. The Reader's Interface should provide the means to let users know where they are and how far they have been in terms of both the thematic and the linear flow. Some participants mentioned that one way of achieving this effect would be by presenting a sideline of document thumbnails synchronized with the reader's movement.

3. As the current stage of this research is the stage of collecting user requirements and feedback, we only supplied a generic search engine to the Reader's Interface. The direction of the overall research agenda is to come up with a search mechanism that suites, to a satisfactory degree, the needs of thematic navigation. It is our plan to refine the search to have better capabilities as general search technology and our own methodology evolves.

4. Presently our system requires a small amount of preprocessing in order to apply the Reader's Interface to a document other than *Don Quixote*. Down the road, the ultimate goal is to come up with a platform that is plug-and-playable and can be placed right on top of any or many documents.



The machine highlights an even bigger challenge. We might have to wait for an indefinite amount of time for computers to be smart enough to automatically pinpoint and set the boundaries of the thematically related excerpts. At the current stage, extracting such information is still under trial. Experiments in different approaches are being performed by a vast number of researchers, as we have discussed in Section 2. Gradually the improvements are shedding light on the various problems with this sort of system. It is going to take a tremendous effort to reach a major breakthrough. Our study on the machine side is one trial amongst many heading to the final destination of an intelligent “reading wheel.” As it is at a preliminary phase of its own, predictably, many future plans are pending. Here we list and briefly discuss several that are imminent.

1. The proposed methodology has only been tested on *Don Quixote*, but we plan to test it on other books as well. We especially look forward to testing it on a textbook, as a textbook represents a repository text rather than a narrative text. It was interesting to see the differences in the initiatives that drove people to read thematically instead of linearly in the two cases, as well as observing whether and when the action itself might be carried out differently.

2. In the experiments, an expert came in and judged the results. Experts’ judgment is valuable feedback for system justification. Nevertheless, relying on it solely has its drawbacks. In particular, reading is a universal phenomenon covering a whole spectrum of backgrounds. Since we are targeting the general public in designing this tool, it is undoubtedly necessary to include a larger pool of judges and collect feedback from people with various backgrounds.

3. Expansion by inflectional variants stood out in our exploration as a useful method. Other recall-enhancing methods, like word co-occurrence, have long been under investigation by researchers for different applications [Billhardt et al. 2002]. We also plan to test its use in our research context.

4. As the current study is inspired by the editorial procedure of composing the index for a textbook, we focused on the testing of short phrases. One interesting question is how we can find a counterpart methodology for long inputs when proximity checking might be too restrictive to be effective. This is a question that needs to be addressed in the next phase of the follow-up research.

## 7. CONCLUSIONS

The realization of fluent practices of thematic reading counts on the fulfillment of two aspects (see Figure 18): 1) on the interface side, the understanding of the expectations of thematic readers and the availability of dynamic visualizations of the relevant information; 2) on the computation side, the effectiveness of textual analysis that locates and determines the data set of interest. With continuous efforts in both aspects, we can hope to reach the ultimate goal: thematic reading becomes a natural part of interacting with electronic texts. Readers express their interest both dynamically and freely. They are equipped with the capability to explore in ways self-oriented and interactive. The information they would like to see is packaged quickly and presented right in front of them. At any point, if they choose, they may always fall back to passive linear reading.

This dissertation contributes to the realization of the ultimate goal by initiating investigations that push forward both aspects. We have conducted experiments to study both the guidelines for the design of a reader's interface and the automatic searching for excerpts that are relevant to a certain topic.

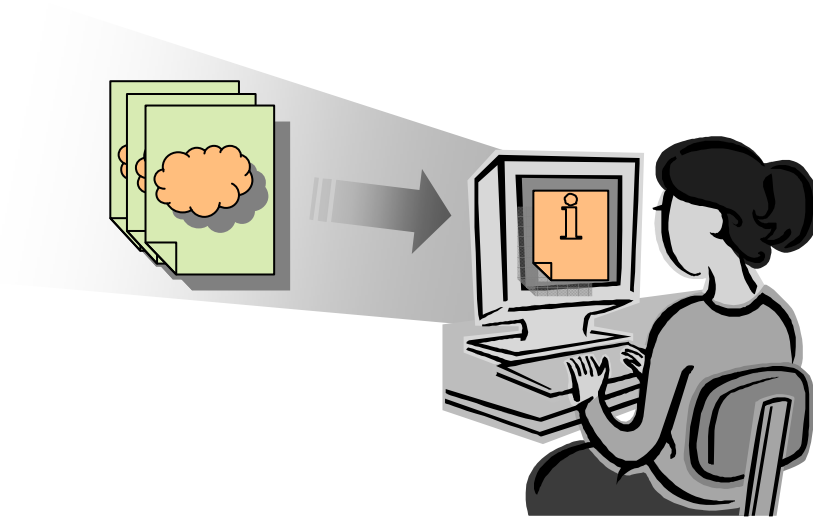


Figure 18. The two aspects involved in thematic reading

When investigating the reader's behavior, we implemented a Reader's Interface with features designed to test the hypothesis in thematic reading. We conducted user studies using both the Reader's Interface and Microsoft Word. The findings suggest that thematic reading is a phenomenon that comes naturally to advanced readers. The idea of gaining an understanding of the topic by carrying out reading solely along an individual topic line is a readily acceptable strategy. During reading, most readers prefer a display of the thread only. From time to time, they do switch to the original text for more contexts. For this purpose, readers have emphasized that it is important to have a consistent two-way navigation between the thread and the linear narrative. It is also important to provide a contextual indication in the margin. Such presence encourages free-form explorations because it provides a sense of certainty of the reader's movement and location when they are off the linear track.

In the testing scenario that mimics multi-thread reading, we discovered that the visualization of the implementation should endeavor not to suggest a biased direction. Otherwise, the reader shows a clear tendency toward the conception that is favored by the biased suggestion. This, consequently, blocks and limits the exploitation of the content itself.

As for automatically locating relevant information, to improve the search performance, we propose a two-step methodology which, in short, is composed of query expansion and search result refinement, directly targeting the enhancements of recall and precision, respectively. In the query expansion section, we experimentally compared the expansion with WordNet, morphological inflections, and both processes together. Our results show that in the context of our study, WordNet made almost no contribution to the enhancement of recall, while expansion with the inflectional tool turned out to be a successful and essential scheme. For the refinement section, we checked the term-occurrence proximity of alternative phrases on the chapter level. The alternative phrases are derived from terms expanded by either inflectional variants or inflectional variants, together with WordNet synonyms. A chapter list that meets the proximity criteria is then used to filter out the sentences that do not belong to the ones in the list. The results show that the proximity check on the alternative phrases formed after inflectional expansion can effectively increase the precision of the previously acquired return results.

We have also tested a different scheme of defining the target and verification units used in the methodology. This scheme ignores the structural delimitations and defines the units by sliding windows of a certain number of terms. Our findings show

that the first scheme of using sentences and chapters as the two units outperformed the second. The first scheme was able to achieve consistently desirable results, while the results from the second were inconclusive. This indicates that the division of the chapters is usually structured in a way that represents a theme-based segmentation. We also conducted proximity tuning on each scheme. A proximity value of 100-300 words apart received consistently desirable results in the first scheme.

In summary, we suggest that if short phrases are used as input for locating thematic pinpoints, then: 1) it is recommended to expand the input with inflectional variants; 2) if the phrase contains more than one term, it is also recommended to use evidence gained by a proximity check on a passage level that is larger than the pinpoints to filter the results; 3) a proximity value of 100-300 is generally recommended in order to provide effective filtration; and 4) following the pre-dictated structural delimitations, such as sentences and chapters, for defining the units of target and verification is a better mechanism than using generic sliding windows.

A narrative text is a remarkable (if it is well written) textual tapestry which is craftily woven together, with a collection of storylines, themes, and leitmotifs taking turns to surface at points exquisitely designed. Nonetheless, it is a compelling prediction that electronic books could let us dictate the course of the plot at our will. It is an important omen that electronic texts offer their readers complete control over the storytelling process. The reality seems to be a bit harsh if we realize how much the interactive readers still yield to the linearity of the texts. Leggett and Shipman [Leggett and Shipman 2004] urged the call for research agendas that support interactive scholarly

communication. Our study falls into this broad agenda, with a particular target of pushing reading practice toward the high end of the scales of interaction and narrative. With more research endeavors joining this stream, let us hope that one day thematic reading is just a button push away. When this happens and when we are not always focused on finding one single path, perhaps we will become more open to the world.

## REFERENCES

- APACHE LUCENE. 2008. *Lucene Java*. <http://lucene.apache.org/java/> [Date accessed: January 2008].
- BILLHARDT, H., BORRAJO, D. AND MAOJO, V. 2002. A context vector model for information retrieval. *J. Am. Soc. Inf. Sci. Technol.* 53, 3, 236-249.
- BILOTTI, M.W., KATZ, B. AND LIN, J. 2004. What works better for question answering: stemming or morphological query expansion? In *Proceedings of the Information Retrieval for Question Answering*. Workshop at SIGIR 2004, Sheffield, England, July.
- BJÖRK, S., BRETAN, I., DANIELSSON, R. AND KARLGREN, J. 1999. WEST: A web browser for small terminals. *CHI Letters* 1, 1, 187-196.
- BOLTER, J.D. 1991. Writing space: The computer, hypertext, and the history of writing. In *Writing Space: The Computer, Hypertext, and The History of Writing*. L. Erlbaum Associates, Hillsdale, NJ, 87-87.
- BOWER, G.H. AND MORROW, D.G. 1990. Mental models in narrative comprehension. *Science* 247, 4938, 44-48.
- BROWN, M.P. 2007. Undisciplined reading: finding surprises in how we read. *Common-Place: Common Reading* 8, 1. Oct. <http://www.common-place.org/vol-08/no-01/reading> [Date accessed: September 2008].
- CALLAN, J.P. 1994. Passage-level evidence in document retrieval. In *Proceedings of the 17<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 302-310.
- CAÑAS, A.J., VALERIO, A., LALINDE-PULIDO, J., CARVALHO, M. AND ARGUEDAS, M. 2003. Using WordNet for word sense disambiguation to support concept map construction. *SPIRE03*. Manaus, Brazil.
- CERVANTES PROJECT. 2008. *Information*. <http://cervantes.tamu.edu> [Date accessed: January 2008].
- CERVANTES SAAVEDRA, M. 1885. *The Ingenious Gentleman Don Quixote of La Mancha*. A translation with introduction and notes by John Ormsby. Smith, Elder & Co., Longdon. Part I (1605), Chapter 18.



[http://www.csd.tamu.edu/cervantes/V2/textos/Ormsby/quijote\\_III.htm](http://www.csd.tamu.edu/cervantes/V2/textos/Ormsby/quijote_III.htm) [Date accessed: January 2008].

CHI, E.H., HONG, L., GUMBRECHT, M. AND CARD, S.K. 2005. ScentHighlights: Highlighting conceptually-related sentences during reading. *IUI'05*. Jan. 272-274.

CHI, E.H., HONG, L., HEISER, J. AND CARD, S.K. 2004. eBooks with indexes that reorganize conceptually. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria, April 24 – 29, 2004). CHI'04. ACM, New York, NY 1223-1226.

CHORNEY, T. 2005. Interactive reading, early modern texts and hypertexts: A lesson from the past. *Academic Commons*.  
<http://www.academiccommons.org/commons/essay/early-modern-texts-and-hypertext> [Date accessed: January 2008].

CRESTANI, F. AND MELUCCI, M. 2003. Automatic construction of hypertexts for self-referencing: The Hyper-Textbook Project. *Information Systems*, 28, 769-790.

DENG, J., RUFURA, R., AND URBINA, E. 2007. Locating thematic pinpoints in narrative texts with short phrases: a test study on *Don Quixote*. In *Proceedings of the 2007 Conference on Digital Libraries* (Vancouver, BC, Canada, June 18-23, 2007). ACM Press, New York, NY 402-410.

DILLON, A. 1994. *Designing Usable Electronic Text*. Taylor & Francis Ltd., London, England.

EVANS, D.K. 1998. LinkIT documentation. Columbia University Department of Computer Science Report.  
<http://www.cs.columbia.edu/~devans/papers/LinkITTechDoc/> [Date accessed: September 2008].

GOLOVCHINSKY, G. AND MARSHALL, C.C. 2000. Hypertext interaction revisited. In *Proceedings of Hypertext'00*. ACM Press. San Antonio, TX 171-179.

GONZALO, J., VERDEJO, F., CHUGUR, I. AND CIGARRÁN, J. 1998. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP*. Montreal, Canada 38-44.

HEARST, M.A. 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23, 1, 33-64.

HEARST, M.A. AND PLAUNT, C. 1993. Subtopic structuring for full-length document access. In *Proceedings of the 16<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pittsburgh, PA, June 27 – July 01, 1993). Korfhage, R., Rasmussen, E. and Willett, P. Eds. SIGIR'93. ACM, New York, NY 59-68.

ISER, W. 1972. The reading process: a phenomenological approach. *New Literary History*. 3, 2, On Interpretation: I, (Winter, 1972). The Johns Hopkins University Press. Baltimore, MD 279-299.

KASZKIEL, M. AND ZOBEL, J. 1997. Passage retrieval revisited. In *Proceedings of the 20<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Belkin, N.J., Narasimhalu, A.D., Willett, P. and Hersh, W. Eds. SIGIR'97. ACM Press, New York, NY 178-185.

KOZIMA, H. 1993. Text segmentation based on similarity between words. In *Proceedings of the 31<sup>th</sup> Annual Meeting on Association for Computational Linguistics* (Columbus, Ohio, United States, June 22 – 26, 1993). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ. 286-288.

LEGGETT, J.J. AND SHIPMAN, F.M. 2004. Directions for hypertext research: exploring the design space for interactive scholarly communication. In *Proceedings of the 15<sup>th</sup> ACM Conference on Hypertext and Hypermedia*. Santa Cruz, CA, USA. HYPERTEXT'04. ACM, New York, NY 2-11.

LESGOLD, A.M. AND PERFETTI, C. 1981. Interactive processes in reading: Where do we stand? *Interactive Processes in Reading*. Lawrence Erlbaum Associates, Hillsdale, NJ 387-407.

LIU, S., LIU, F., YU, C. AND MENG, W. 2004. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In *Proceedings of the 27<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'04. ACM Press, New York, NY 266-272.

LUNZER, E. 1979. From learning to read to reading to learn. *The Effective Use of Reading*. Lunzer E. and Gardner K. Eds. Heinemann Educational Books for the Schools Council, London 7-36.

- LYON, C., MALCOLM, J. AND DICKERSON, B. 2001. Detecting short passages of similar text in large document collections. In *Proceedings of the 2001 conference on Empirical Methods in Natural Language Processing*. 118-125.
- MAKHOUL, J., KUBALA, F., SCHWARTZ, R. AND WEISCHEDEL, R. 1999. Performance measures for information extraction. In *Proceedings of the DARPA Broadcast News Workshop*. 249-254.
- MIHALCEA, R. AND MOLDOVAN, D. 2000. Semantic indexing using WordNet senses. In *Proceedings of ACL Workshop on IR and NLP*. Hong Kong.
- MILLER, G.A. 1995. A lexical database for English. *Communications of the ACM*, 38, 11, November 39-41.
- MURRAY, J.H. 1997. *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. The Free Press, New York.
- MYSQL DOCUMENTATION. 2008. The MySQL online documentation for full-text search functions. <http://dev.mysql.com/doc/refman/5.0/en/fulltext-search.html> [Date accessed: January 2008].
- O'HARA, K. 1996. *Towards a Typology of Reading Goals*. Technical report EPC-1996-107. Rank Xerox Research Centre Cambridge Laboratory, Cambridge, U.K.
- PATEL D. AND MARSDEN, G. 2004. Customizing digital libraries for small screen devices. In *Proceedings of SAICSIT 2004*. 234-238.
- ROBERTS, I. AND GAIZAUSKAS, R. 2004. Evaluating passage retrieval approaches for question answering. In *Proceedings of 26<sup>th</sup> European Conference on Information Retrieval*, 72-84.
- RUMELHART, D.E. 1997. Toward an interactive model of reading. In *Theoretical Model and Processes of Reading (3<sup>rd</sup> ed.)*. Singer H. and Ruddell, R.B. Eds. Academic Press, New York, 573-603.
- SALTON, G., ALLAN, J. AND BUCKLEY, C. 1993. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pittsburgh, PA, June 27 – July 01, 1993). Korfhage, R., Rasmussen, E. and Willett, P. Eds. SIGIR'93. ACM, New York, NY 49-58.

- SALTON, G., ALLAN, J., BUCKLEY, C. AND SINGHAL, A. 1994. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264, 5164. June 1421-1426.
- SALTON, G., SINGHAL, A., BUCKLEY, C. AND MITRA, M. 1996. Automatic text decomposition using text segment and text themes. *Hypertext'96*. March 53-65.
- SCHILIT, B.N., GOLOVCHINSKY, G. AND PRICE, M.N. 1998. Beyond paper: supporting active reading with free form digital ink annotations. In *Proceedings of CHI9*. ACM Press. Los Angeles, CA 249-256.
- SRINIVASAN, S., PONCELEON, D., PETKOVIC, D. AND VISWANATHAN, M. 2000. Query expansion for imperfect speech: applications in distributed learning. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries (Cbaivl'00)*. IEEE Computer Society, Washington, DC 50.
- TELLEX, S., KATZ, B., LIN, J., FERNANDES, A. AND MARTON, G. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'03. ACM Press, New York, NY 41-47.
- THOMPSON, F.C. 2005. *Thompson Chain-Reference Study Bible-NIV*. Kirkbride Bible Company. Indianapolis, IN.
- URBINA, E., FURUTA, R., SMITH, S.E., AUDENAERT, N., DENG, J. AND MONROY, C. 2006. Visual knowledge: textual iconography of the *Quixote*, a hypertextual archive. *Literary and Linguistic Computing* 2006, 21, 2, 247-258.
- VAN RIJBERGEN, C.J. 1979. *Information Retrieval*. Butterworths. London, United Kingdom.
- VIPOND, D. AND HUNT, R.A. 1984. Point-driven understanding: pragmatic and cognitive dimensions of literary reading. *Poetics*, 13. June. 261-277.
- VOORHEES, E.M. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Croft, W.B. and Van Rijsbergen, C.J. Eds. Annual ACM Conference on Research and Development in Information Retrieval. Springer-Verlag New York, New York, NY 61-69.

- WACHOLDER, N., EVANS, D.K. AND KLAVANS, J.L. 2001. Automatic identification and organization of index terms for interactive browsing. In *Proceedings of the 1<sup>st</sup> ACM/IEEE-CS Joint Conference on Digital Libraries* (Roanoke, VA). JCDL'01. ACM, New York, NY 126-134.
- WORDNET MANUAL. 2008. The WordNet's morphological processing manual. <http://wordnet.princeton.edu/man/morphy.7WN.html> [Date accessed: January 2008].
- YANG, C.C. AND WANG, F.L. 2003. Fractal summarization for mobile devices to access large documents on the web. *WWW 2003*. May 215-224.
- YIN, X. AND LEE, W.S. 2004. Using link analysis to improve layout on mobile devices. *WWW 2004*. May 338-344.
- ZELLWEGER, P.T., MANGEN, A. AND NEWMAN, P. 2002. Reading and writing fluid hypertext narratives. *Hypertext'02*. June 45-54.

## VITA

Jie Deng received her Bachelor of Science degree in environmental engineering from Tsinghua University in Beijing in 1999. She received her Master of Science degree in computer science from the University of Nebraska – Omaha in 2002. In December 2008, she received her Doctor of Philosophy degree in computer science from Texas A&M University in College Station, Texas.

During her doctoral studies, Jie Deng has worked as a Graduate Teaching Assistant for various classes, including introductory programming and data structure, for the Department of Computer Science at Texas A&M University. She has also worked as a Graduate Research Assistant in the Cervantes Project with a focus on applying web technology for the management and visualization of rare book textual iconography. She has initiated her own research topic exploring the reading behavior that concerns a specific topic of interest.

She has published and coauthored papers presented at international conferences and in journals such as: ACM/IEEE Joint Conference on Digital Libraries, Digital Humanities, and Literary and Linguistic Computing.

Jie Deng can be contacted at:

Texas A&M University  
Department of Computer Science  
TAMU 3112  
College Station, TX 77843-3112

jdeng@cs.tamu.edu