Texplore - exploring expository texts via hierarchical representation

Yaakov Yaari
Bar Ilan University, Ramat Gan

Abstract

Exploring expository texts presents an interesting and important challenge. They are read routinely and extensively in the form of online newspapers, web-based articles, reports, technical and academic papers. We present a system, called *Texplore*, which assists readers in exploring the content of expository texts. The system provides two mechanisms for text exploration, an expandable outline that represents the hierarchical structure of the text, and a concept index, hot-linked to the concept references in the text. The hierarchical structure is discovered using lexical cohesion methods combined with hierarchical agglomerative clustering. The list of concepts are discovered by n-gram analysis filtered by part-of-speech patterns. Rather than the common presentation of documents by static abstracts, *Texplore* provides dynamic presentation of the text's content, where the user controls the level of details.

1 Introduction

Ever-faster computers, the Internet together with large information repositories linked by high-speed networks, are combined to provide immediate accessibility to large amounts of texts. The urgency of exploring these texts varies depending on the consumer - students, researchers, professionals, decision makers, or just anybody. In any case the amounts of texts are beyond our ability to digest them.

Research in information retrieval (IR) has been focused until now on the task of presenting relevant documents to the user. Commercial tools followed suit, as evident by the many powerful search engines now available on the Web. Typically, the relevant documents are presented by some automatically computed abstract.

Our work focuses on medium size and longer documents where the user needs some further assistance in exploring the content of the retrieved document. The idea then is to extend the applicability of IR methods, beyond the document retrieval, to the task of a reading assistant.

We might expect from a reading assistant, or a text exploration tool, to provide us with two basic capabilities:

1. A controlled view of the content of the document.
2. The list of concepts discussed in the text.

The first capability might be seen as an "electronic" table-of-contents, and is the key vehicle for efficient text exploration. The second can be seen as an "electronic" index, and provides an orthogonal access vehicle to the mechanism of a table-of-contents.

We have implemented such a text exploration system in *Texplore*. It is designed for expository texts such as informative articles in magazines, technical reports and scientific papers. In Section 2 we discuss some characteristics of expository texts that make possible the development of the above text exploration capabilities. In Section 3 we focus on the importance of hierarchical representation of texts as a visualization tool. Section 4 details the *Texplore* system itself, again focusing on the hierarchical content representation. We conclude by discussing some shortcoming of the system and plans for improvements.

2 Some characteristics of expository texts

The following subsections consider the linguistic evidence that is needed to develop a hierarchical representation of expository texts. The next

	-Projected	+Projected
	Narrative	Procedural
+ Succession	1. First/Third person 2. Agent oriented 3. Accomplished time 4. Chronological time	1. Non-specific person 2. Patient oriented 3. Projected time 4. Chron. linkage
	Expository	Horatory
- Succession	1. No necessary reference 2. Subject matter oriented 3. Time not focal 4. Logical linkage	1. Second person 2. Addressee oriented 3. Mode, not time 4. Logical linkage

Table 1: Discourse types

subsection discusses expository text in general and its relation to other discourse types. Hierarchical structure in discourse is discussed next, with the paragraph as its basic unit. The final subsection considers lexical cohesion as the basic technique for identifying structure.

2.1 Expository text and other discourse types

In order to understand the particular domain of expository text is important to see it in the larger context of other possible discourse types. Longacre (1976) presents a 2×2 model of four discourse types, *Narrative*, *Procedural* (Instructive), *Expository*, and *Horatory* (Sermon), shown in Table 1.

Expository text is seen to be less modal since its discourse is determined by its subject, and the logical structure built in its exposition, rather than by who the speaker is, the audience or the temporal order of the speech acts. This is not to say that two authors are expected to produce the same text on the same subject. Personal style is a factor here as in any human writing. However, we can take advantage of the modeless character of expository text when creating a representation of its content. If the discourse relations between two segments can be assumed to be modeless we can expect these relations to be manifested, to a large extent, in their lexical context. In other words, we can expect the robust techniques of information retrieval to be useful for identifying the information structure of expository texts.

2.2 The paragraph unit in hierarchical discourse structure

Hierarchical structure is present in all types of discourse. Authors organize their works as trilogies of books, as chapters in books, sections in chapters, then subsections, subsubsections, etc. This is true for an instruction manual, the Bible, for *The Hitchhiker Guide to the Galaxy*, *War and Peace*, and, in a completely different category, this humble paper.

Previous research shows that this hierarchical structure is not just an author's style but is inherent in many language phenomena. A number of rhetoric structure theories have been proposed (Meyer and Rice, 1982; Mann and Thompson, 1987) which recognize distinct rhetorical structures like problem-solution and cause-effect. Applying this model recursively forms a hierarchical structure over the text. From the cognitive aspect, Giora (1985) proposes a hierarchical categorial structure where the discourse topic functions as a *prototype* in the cognitive representation of the unit, i.e. a minimal generalization of the propositions in the unit. Finally, the hierarchical intention structure, proposed for a more general, multiple participants discourse, is a key part of the well-accepted discourse theory of Grosz and Sidner (1986).

Hierarchical structure implies some kind of basic unit. Many researches (Longacre, 1979; Hinds, 1979; Kieras, 1982) have shown that the paragraph is a basic unit of coherency, and that it functions very similarly in many languages of vastly different origin (Chafe, 1979).

Not only the paragraph is a basic unit of coherency, its initial position, the first one or two sentences of the paragraph, provides key information for identifying the discourse topics (Yaari et al.,). Again, as Chafe shows, this is true for many varied languages. The initial position of a paragraph is thus a key heuristic for general purpose document summarization (Paice, 1990).

2.3 Cohesion

Lexical cohesion is the most common linguistic mechanism used for discourse segmentation (Hearst, 1997; Yaari, 1997). The basic notion comes from the work of Halliday and Hasan (1976) and further developed in (Halliday, 1994).

Cohesion is defined as the non-structural mechanism by which discourse units of different sizes can be connected across gaps of any texts. One of the forms of cohesion is lexical cohesion. In this type, cohesion is achieved by choosing words that are related in some way -

lexically, semantically or collocationally. Lexical cohesion is important for a practical reason - it is relatively easy to identify it computationally. It is also important for linguistic reasons since, unlike other forms of cohesion, this form is active over large extents of text.

3 Hierarchical representation of text

In the previous section the hierarchical structure of a text was established as an inherent linguistic phenomena. We have also identified linguistic evidence that can be used to uncover this structure.

In this section we focus on the human-machine interaction aspects of this form of representation. From this point of view, hierarchical representation answers two kinds of problems: how to navigate in free text, and how to effectively communicate the content of the document to the user. These two issues are discussed in the following subsections.

3.1 Navigating in free text

The basic approach for free, unstructured, text navigation (and the basis for the whole internet explosion) is the hypertext method. Navigation follows what may be called a *stream of associations*. At any point in the text the user may hyper-jump to one out of a set of available destination sites, each determined by its association with narrow context around the link anchor. In spite of their popularity, the arbitrary hyper-jumps create a serious drawback by losing the global context. Having lost the global context, the navigator is destined to wander aimlessly in maze of pages, wasting time and forgetting what he/she was looking for in the first place. The use of a static ticker frame that allows an immediate deliverance from this maze (typically placed on the left part of the browser's window) is a recognition of this drawback.

Once NLP methods are applied on the text document, more sophisticated methods become possible for navigating in unstructured text. An important example is the use of lexical cohesion, implemented by measuring distance between term vectors, to decompose the text to themes (Salton et al., 1995). Themes are defined as a set of paragraphs, not necessarily adjacent, that have strong mutual cohesion between them. Navigation through such theme-linked paragraphs is a step forward in effective text exploration. However, the user navigates within the context of a single theme and still loses the overall context of the full text. Because there is only one hierarchy here, the user has to go through a selected theme to its end to find out whether it provides the sought information.

The answer proposed in *Texplore* is to discover and present the user with a hierarchical representation of the text. Hierarchical structure is oriented specifically to present complex information. Authors use it explicitly to organize large works. Scientists use it to describe complex flora and fauna. Manual writers use it to describe complex procedures. Our task here is somewhat different. We are presented with a given unstructured text and want to uncover in it some latent hierarchical structure. We claim that in so far as the text is coherent, that is, it makes sense, there is some structure in it. The more coherent the text, the more structure it has.

Combining the capabilities of hypertext and hierarchical representation is particularly attractive. Together they provide two advantages not found in other access methods:

1. Immediate access to the sought piece of information, or quick dismissal if none exists. In computer memory jargon we call this *random access*. This is the ability to access the required information in a small number of steps (bound by the maximum depth of the hierarchy).

2. User control over the level of details. Most navigation tools provide the text as is so the user has to scan at the maximum level of details at all times. However, for expository texts beyond a couple of pages in size, the user needs the ability to skim quickly over most of the text and go deeper only at few points. There is a need, then, to have good interactive control over the level of details presented.

3.2 Communicating document's content

Document summarization systems today are concerned with extracting significant, indicative sentences or clauses, and combining them as a more-or-less coherent abstract.

This static abstract should answer the question "what the text is about". However, because of the underlying technique of sentence extraction and its static nature, the answer is too elaborated in some of the details and insufficient in others.

Boguraev et al. (1998) discuss extensively this drawback of today's summarizers and conclude that good content representation requires two basic features: (a) presenting the summary extracts in their *context*, and (b) user control over the *granularity*. Their solution is based on identifying primitive clauses, called *capsules*, resolving their anaphoric references and providing them, through a user interface, at different granularities.

The expandable outline view of *Texplore*, built upon hierarchical representation of the text's contents, nicely meets the requirements of context and granularity, though the underlying NLP technology is completely different. In the next section we discuss the *Texplore* system in details, the supporting NLP tools as well as the front-end visualization system.

4 *Texplore* - system description

The overall data-flow in *Texplore* is shown in Figure 1. It starts with a preprocessing stage, a structure and heading analyses, leading to expandable outline display.

A typical screen of *Texplore* is shown in Figure 2[1]. It consists of three parts. The original text is shown on the right pane, the expandable outline on the upper left and the concept index on the lower left pane.

The following subsections describe the different parts of the system, focusing on the visualization aspects related to content presentation.

4.1 NLP preprocessing

The first two preprocessing steps, sentence analysis part-of-speech (POS) analysis, are pretty standard. The result is a list of POS-tagged sentences, grouped in paragraphs.

In the N-gram analysis, a repeated scan is made at each i'th stage, looking for pairs of consecutive candidates from the previous stage. We filter each stage using mutual information measure so the complexity is practically $O(N)$. Finally we remove those N-grams whose instances are a proper subset of some longer N-grams, and then apply part-of-speech filter on the remaining candidates leaving only noun compounds. That last step was found to be extremely useful, reducing false N-grams to practically nil.

[1]The text in this screen, as well as in the other screen captures, is from the article *staryazers* (Hearst, 1997). It deals with the possibility of life on other planets in view of the unique combination of earth and its moon

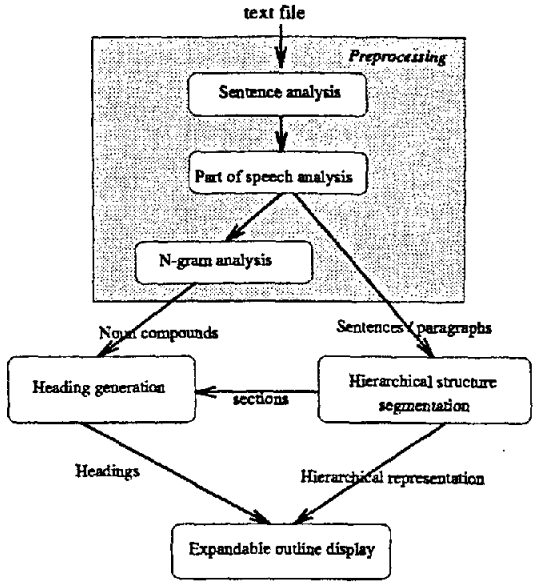

Figure 1: Data flow in *Texplore*

4.2 Hierarchical structure segmentation

The core of the system is the hierarchical structure segmentation. The method used for segmentation, called hierarchical agglomerative clustering (HAC), was described in detail by Yaari (1997). In HAC the structure is discovered by repeatedly merging the closest data elements into a new element. In our application we use paragraphs as the elementary segments of discourse and apply a lexical cohesion measure as the proximity test. The lexical cohesion measure, $Proximity(s_i, s_{i+1})$, is adapted from the standard Saltonian term vector distance. It computes the cosine between two successive segments, s_i and s_{i+1}.

$$Proximity(s_i, s_{i+1}) = \sum_{k=1}^{n} \frac{w_{k,i} \cdot w_{k,i+1}}{||s_i|| \cdot ||s_{i+1}||} \quad (1)$$

Here $w_{k,i}$ is the weight of the k'th term of s_i, and $||s_i||$ is the length of the vector. The as-

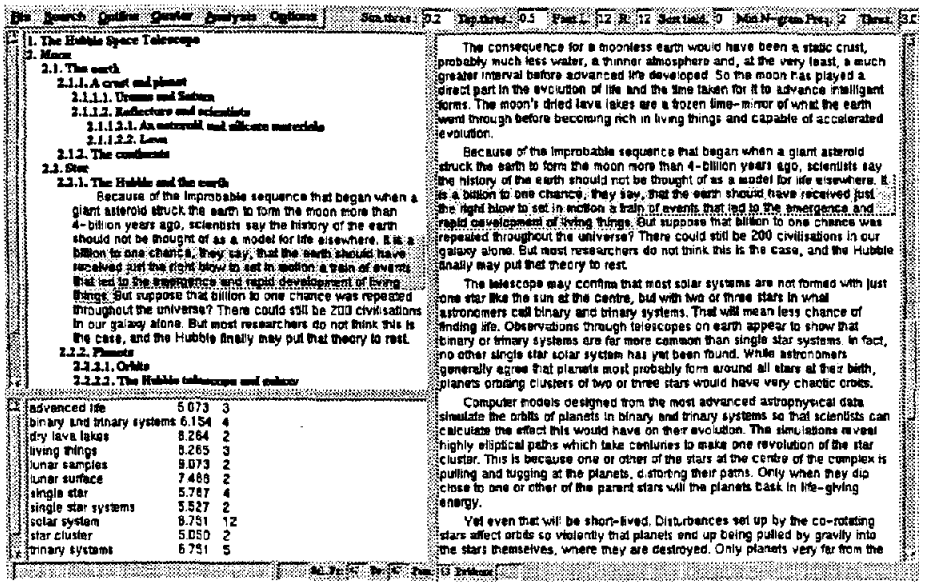

Figure 2: *Texplore* screen. The original text is shown on the right, the expandable outline in the upper left and the concept index on the lower left. The outline is shown collapsed except for one section

sumption that only adjacent segments are compared is not necessarily the case, see (Salton et al., 1995). However, it allows us to create the more conventional 2-D structure of a table-of-contents, instead of the arbitrary graph structure that would have been formed otherwise. Another modification is the way the term weights, $w_{k,i}$, are determined. We found that having the weight proportional to the term's IDF (Inverse Document Frequency, measuring its general significance) and the position of the sentence in the paragraph, improves the quality of the proximity measure, by giving higher weight to terms with higher relevance to inter-segment cohesion.

The result is shown in Figure 3. Inter-segment boundaries are set at points where lexical cohesion falls below some specific threshold. The resulting nesting could be quite deep (in this example there are 10 levels). Human authors, however, rarely use a hierarchy depth greater than 3 (except possibly in instructional discourse). The rather deep nesting is then smoothed, between the detected boundaries, to fit human convenience, as seen in Figure 4. This smoothed structure is superimposed over the original text, producing the expandable outline shown in the left pane of Figure 2.

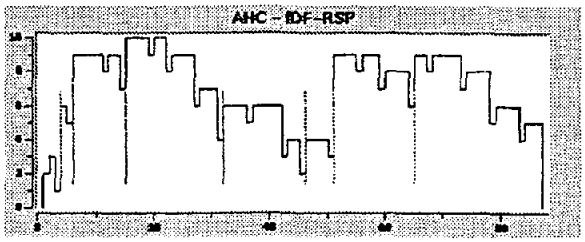

Figure 3: Result of hierarchical segmentation. Paragraphs are along the X-axis. Y-axis indicates proximity. Higher outline implies closer, and thus deeper-nested, adjoining segments. Vertical lines indicate inter-segment boundaries.

The hierarchical structure thus discovered is certainly not the only one possible. However, experiments with human judges (Hearst, 1997) showed that segmentation based on lexical cohesion is quite accurate compared to manual ones.

4.3 Heading generation

The next step, after the hierarchical structure of the text is determined, is to compose headings for the identified sections. Figure 5 shows the outline pane representing the hierarchical structure.

The generated headings are, at the moment,

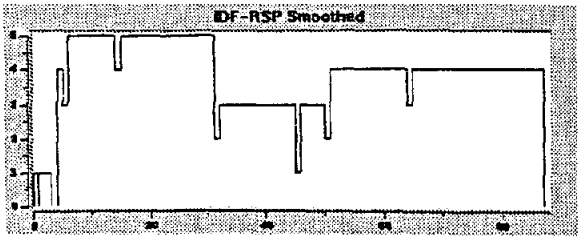

Figure 4: The hierarchical structure in Figure 3 after smoothing

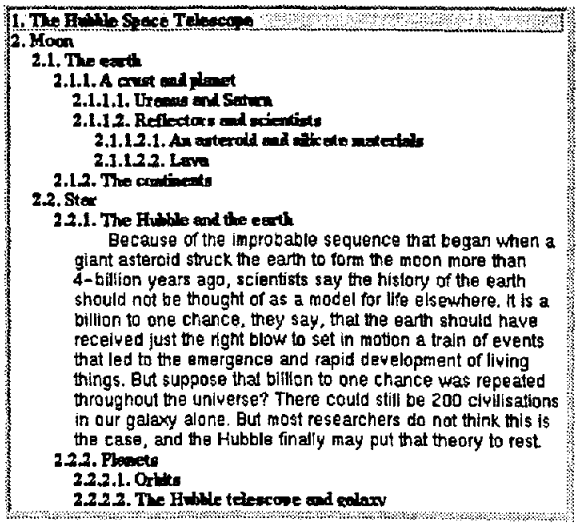

Figure 5: The text' outline representing the its hierarchical structure.

quite simplified. The saliency of noun compounds (NPs) is scored for each section by their frequency, mutual information (for N-grams with $N > 1$), and position in the section. A higher score is given to NPs in initial position, that is, the first one or two sentences in the paragraph.

The syntax of headings follows standard conventions for human-composed headings. The most common case is the simple NP. Another common case is a coordination of two NPs. With these guidelines we came up with the following heuristics to compose headings:

1. Remove from the list any NP that appears in an enclosing section.

2. If the saliency of the first NP in the list is much higher than the second, use this NP by itself. Otherwise, create a coordination of the first two NPs.

3. Prefix each NP in the heading with the determiner that first appeared with that NP, if any. *This rule is not very successful and will be modified in the future.*

The first rule exemplifies the power of this kind of content representation. Once an NP appears in a heading, it is implied in the headings of all enclosed sections and thus should not appear there explicitly. For example, in Figure 5 the NP *Moon* appears in the heading of section 2. Without the first rule it would appear in a few of the enclosed subsections because of its saliency. We, as readers, would see this as redundant information.

4.4 Expandable outline display

Figure 5 also illustrates the importance of context and granularity, mentioned earlier as key points in dynamic content presentation. The outline functions as a dynamic abstract with user control over the level of granularity. The user is able to read section 2.2.1, `The Hubble and the earth` in its full context. He/she is not *lost* anymore.

In fact, the two panes with the original text and its outline are synchronized so that the outline can act as a ticker-only pane viewing the text on the larger right pane, or be used as a combined text and heading pane.

The outline pane also supports the standard controlled-aperture metaphor. Double-clicking on a heading alternately expands and collapses the underlying text. The user can thus easily increase the granularity of the outline to see further detail, or close the aperture to see only the general topics.

The heading acts cognitively as the surrogate of its text. Thus if a heading of a collapsed section is selected, the full text corresponding to this heading is selected on the right. This strengthens the concept of the outline as a true, thought compact, representation of the original text. Figure 6 shows the same outline, this time with reduced granularity, highlighting the correspondence between a selected heading and its text. The concept index, shown in the lower left pane of the window, is discussed in the next section.

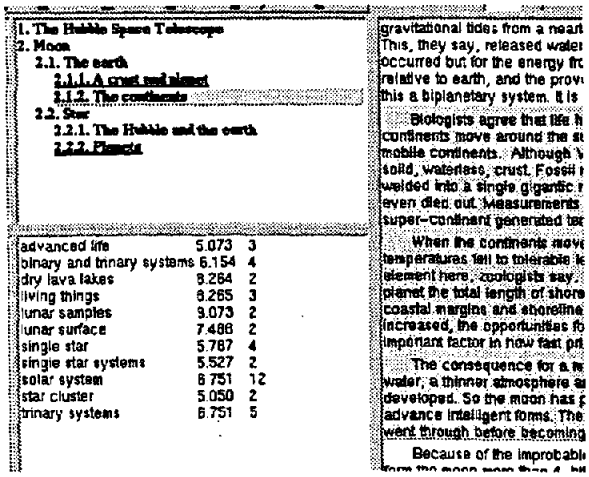

Figure 6: Collapsed outline showing correspondence between a heading and its underlying text.

4.5 Concept index

The N-gram analysis, combined with part-of-speech filtering, identifies a set of noun compounds that are used as a concept index for the text. They are termed *concepts* because the information they carry reveals a lot about the text, much more than simple one word nouns. Consider, for example, the first three N-grams: `lunar samples`, `living things`, and `dry lava lakes`. In contrast, the composing words of each N-gram, e.g. `lava`, `lake`, `living`, or `things`, reveal very little.

The high information content of the concept index makes it a very concise representation of what the text is about, though certainly secondary to the outline. Also, having these "concepts" hot-linked to their references in the text forms a hot-link index of key topics of the text.

5 Conclusions and future plans

We propose a new approach for dynamic presentation of the content of expository text based on uncovering and visualizing its hierarchical structure. Using this "electronic" table-of-contents the user has the advantage of exploring the text while staying within the full context of the exploration path. The user has also full control over the granularity of the displayed information. These characteristics are beneficial both for navigating in the text as well as communicating its content, while overcoming drawbacks of existing summarization methods.

The weakest point in *Texplore* is the generation of headings. The current approach is too simplistic, both in the criteria for selecting NPs and in the way they are composed to headings. We have analyzed the way headings are formed by human authors (Yaari et al.,) and the results were used to form a machine-learning system which identifies the NPs of a given section using multiple sources of information. The system constructs headings for the text hierarchy using a fixed set of syntactic formats (found to be common in heading syntax). We are in the process of integrating this system into *Texplore*.

The hierarchical structure segmentation is also too simplistic, based solely on the proximity of term vectors. Again, we are working on a machine learning system that uses a set of structured articles to learn segmentation rules. The basic approach is to divide the task into two steps, determining the boundaries and forming the hierarchy. We are using various cohesion cues, associated with each paragraph, as the learning attributes: lexical similarity, cue tags, cue words, number of starting and continuing lexical chains, etc.

Using machine learning has the advantage of a built-in evaluation against the segmentation done by human subjects. We also plan to evaluate the usefulness of the hierarchical presentation in terms of reading effectiveness.

References

B. Boguraev, Y.Y. Wong, C. Kennedy, R. Bellamy, S. Brawer, and J. Swartz. 1998. Dynamic presentation of document content for rapid on-line skimming. To be published in the Spring 1998 AAAI Symposium on Intelligent Text Summarization.

W.L. Chafe. 1979. The flow of thought and the flow of language. In T. Giv'on, editor, *Syntax and Semantics: Discourse and Syntax*, volume 12, pages 159–182. Academic Press.

R. Giora. 1985. A text based analysis of non-narrative texts. *Theoretical Linguistics,*, 12:115–135.

B.J. Grosz and C.L. Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics,*, 12(3):175–204.

M.A.K. Halliday and R. Hasan. 1976. *Cohesion in English.* New York: Longman Group.

M.A.K. Halliday. 1994. *Introduction to Functional Grammar, second edition.* London: Edward Arnold.

M. A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, March.

J. Hinds. 1979. Organizational patterns in discourse. In T. Giv'on, editor, *Syntax and Semantics:: Discourse and Syntax*, volume 12, pages 135–158. Academic Press.

D.E. Kieras. 1982. A model of reader strategy for abstracting main ideas from simple technical prose. *Text*, 2(1–3):47–81.

R. Longacre. 1976. *An anatomy of Speech Notions.* Peter de Ridder Press, Lisse.

R.E. Longacre. 1979. The paragraph as a grammatical unit. In T. Giv'on, editor, *Syntax and Semantics: Discourse and Syntax,*, volume 12, pages 115–134.

W.C. Mann and S.A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, ISI.

B.J.F. Meyer and G.E. Rice. 1982. The interaction of reader strategies and the organization of text. *Text*, 2(1-3):155–192.

C.D. Paice. 1990. Constructing literature abstracts by computer : techniques and prospects. *Information Processing and Management,*, 26(1):171–186.

G. Salton, A. Singhal, C. Buckley, and M. Mitra. 1995. Automatic text decomposition using text segments and text themes. Technical Report TR95-1555, Cornell Unievrsity, November.

Y. Yaari, Y. Choueka, and M. Elhadad. Analysis of headings and sections' structure in expository texts. Not yet published.

Y. Yaari. 1997. Segmentation of expository texts by hierarchical agglomerative clustering. In *Proceedings of RANLP*, pages 59–65, Tzigov Chark, Bulgaria.