Hypertext:
The Importance of being Different

MSc Dissertation

Einat Amitay

Centre for Cognitive Science
The University of Edinburgh

September 1997

Acknowledgements

There are many people to whom I would like to thank: to Jon Oberlander for being attentive and patient, even in bad moments; to Chris Brew whose door has always been open; to Richard Thacker who always agreed with me enthusiastically when asked to; to Richard Tobin, David McKelvie, Andrei Mikheev and CST for their UNIX, SGML and Java skills; to *all* the MSc students who suffered me silently even when I was navigating the web noisily while they were trying to do real work; and last but not at all least - to my funding buddy - Uri Parker.

abstract

Dillon et al. (1993) observed, when the hypertext authoring on the web was just beginning to become popular in the non-academic world, that there is a problem of schemata, or genre conception, in hypertext, because of the flexible nature of language and the varied layout used in its creation. Today, almost five years later, the web is used by many people and there are conventions which evolved from usage and experience. In the years that passed since then, users became aware of the existence of other users by interacting with their hypertext documents and by creating their own homepages. Through analysing two corpora consisting 1000 HTML files retrieved from the World Wide Web, this study describes the linguistic conventions with which hypertext documents are being written. It is claimed here that hypertext is a new linguistic genre and that it should be treated as such in future studies. It is also suggested in this dissertation that studying these conventions and applying the gained knowledge to existing academic work, would be beneficial to both hypertext users and the research community.

He had bought a large map representing the sea,
Without the least vestige of land:
And the crew were much pleased when they found it to be
A map they could all understand.

'What's the good of Mercator's North Poles and Equators
Tropics, Zones, and Meridian Lines?'
So the Bellman would cry: and the crew would reply,
'They are merely conventional signs!

'Other maps are such shapes, with their islands and capes!
But we've got our brave Captain to thank'
(So the crew would protest) 'that he's bought us the best-
A perfect and absolute blank!

This was charming, no doubt: but they shortly found out
That the Captain they trusted so well
Had only one notion for crossing the ocean,
And that was to tingle his bell.

The Bellman's speech
The Hunting of the Snark
Lewis Carroll

Chapter One

Introduction

The study of text and its linguistic characteristics is one of the leading interests in computational linguistics. These characteristics include phenomena such as preferred syntactic structure, usage of discourse context, choice of words, text layout, etc. The study of the way people produce text can help in designing automatic machines to generate human-like texts where these can save time and effort. The understanding of these conventions in writing texts can also aid in improving the comprehensibility of language interfaces (in contrast with iconic interfaces). Since the language conventions most scientists explore lie in the written text and speech genres, they choose to apply this knowledge to the creation of hypertext.

Hypertext is a term coined by Ted Nelson in 1965 (Nelson, 1974; 1988). But the idea behind this concept was put into writing in July 1945 in an article called "As We May Think" by Vannevar Bush (Bush, 1945). Bush was the director of the American Office of Scientific Research and Development during World War II. Bush named his associative word links - "memex", *memory extension*. The idea behind the *memex* was to first create a central electronic archive which would gather information in the form of on-line associations from all the scientists that could contribute to a given subject, directly or indirectly. Secondly, the *memex* could be used to solve complex problems where the information held by one expert in their field was insufficient. Today most people know hypertext as the genre of text used on the World Wide Web but hypertext is also used in much software for the design of online help manuals or as a tool for managing large textual data bases.

The main difference between flat-text and hypertext is that the former is self contained while the latter can never be. Landow (1992) names this feature *textual integrity*. For example, a flat-document describing hypertext would probably include an introduction to the applications and usage of hypertext, but in the electronic medium there is no need to redefine terms every time we use them. It is enough to link the reader to a site (another online document) which specialises in these aspects, and to concentrate on what the paper is really about. The reader, on the other hand, can then choose whether they need this information in order to understand the article, or not. Hypertext can not be self-contained because then there will be no point in having a link to another document explaining another aspect of the subject discussed. It is true that there are different conventions within different knowledge domains and that people modify their writing for different audiences such as novice or experts, but the fact is that hypertext allows the document to be more flexible since each reader can choose their own path in the knowledge presented.

In this dissertation we would like to test the assumption that flat text and hypertext are different in nature through a comparative linguistic analysis. Significant differences, if to be found, between flat text and hypertext should be taken into account in various applications: from text to hypertext generators to hypertext generation systems (both are currently using the approach that hypertext is only flat text with linking devices instead of pages). The study proceeds by gathering and analysing a specially designed HTML corpora retrieved from the World Wide Web. The results obtained will be compared with other findings obtained by researchers in the fields of text and conversation analysis.

We will adopt here an approach analogous to that suggested by Suchman (1987), in the context of planning:

> *"common-sense notions of planning are not inadequate versions of scientific models of actions, but rather are resources for people's practical deliberations about action... Rather than attempting to abstract action away from its circumstances and represent it as a rational plan, the approach is to study how people use their circumstances to achieve intelligent action. Rather than build a theory of action out of a theory of plans, the aim is to investigate how people produce and find evidence for plans in the course of situated action."* (pp 49-50).

By analogy, we can usefully look at the way in which hypertext designs are accomplished on a day-to-day basis, and extract generalisations from this study. Such a naturalistic, descriptive approach presents a viable alternative to more rationalistic, prescriptive views on hypertext design. Arguably, there are already robust conventions regarding the use and production of hypertext, followed by millions of people every day, embodied in the World Wide Web. The WWW allows people to create their own sites and to join the virtual community. Over the last few years people have evolved certain conventions regarding this means of communication. Consumption of existing hypertext designs has been translated into production of further designs.

There might be an argument against such a descriptive approach claiming that there is a “chicken and egg problem” here - namely whether conventions are caused by, or give rise to, the existing software. There might be truth in saying that users of software are influenced by its design and therefore would use their interpretation of the design to build their own web sites. This claim is not treated here because it is part of the answer to how and why people come up with new conventions, but it does not give a description of what these conventions are. We simply want to assert that there are already linguistic regularities and linguistic behaviour patterns within the WWW which are the fruit of a simple language evolution rule: “*maximum communication with minimal effort*”.

The descriptive approach constitutes a new, general, method for informing hypertext interface and generation design. In this particular study we would like to begin with the smallest unit: the nodes themselves. In many designs of hypertext systems attention is directed to the spatial linking. Even studies such as McDonald and Stevenson (1996; in press), and Wright (1993), which focus on the text within a node, do not relate to the content of the actual links. Their conclusions relate to the structure of the document and to its form, but never to its content, or to the content of the referring anchors (but see Baron, Tague-Sutcliffe, Kinnucan, and Carey (1996) for a suggested classification of anchor content). Many of the studies actually put the blame for the "lost in hyperspace" phenomenon (Conklin, 1987) on the referential space between documents, and not on the labelling of the links within the document. Since one of this dissertation's goals is to study whether this assumption is true, it has been decided to examine, amongst other patterns, the regularity of anchors within hypertext documents: exploiting the endless resources offered by the Web by simply retrieving and extracting information from it. Two corpora of HTML files were gathered and analysed. The data and findings presented here are only the tip of the iceberg of the many conventions which rule the web and which are practised every day by users.

Since the range of issues which a descriptive study could pursue is too large to address in one MSc dissertation, and since this study aims to examine the question whether or not there hypertext is a new distinct genre, this dissertation focuses on the linguistic conventions practised by web users. Some of the questions which will be addressed concern the linguistic structure of hypertext documents. The following paragraphs describe some of these questions.

Domains - On the WWW there are many types of hypertext, authored by many different users and for divers purposes. These hypertext document can be personal homepages; scientific articles authored for the web (vs. the PostScript or plain text versions); *.com* sites, which are central pages representing commercial companies and firms; special interest sites, for example sites for electronic games collectors, special food sites, on-line shops and on-line museums; web fiction sites which are already defined as a new trend in literature; geographical descriptions of cities and tourists sites which are promoted by tourist boards all around the world, etc. Identifying the different domains and the motivations behind publishing them online might be of interest in the user modelling stage of hypertext design. Distinguishing the domains from each other can also be beneficial for designing templates for WYSIWYG (What You See Is What You Get) editors for hypertext authoring. Language conventions can also change from one domain to the other and a domain has to be defined before differentiating it linguistically from other domains.

Document level - In this study the term hypertext document refers to a single HTML file retrieved from the web. A hypertext document may contain text, links, images, and buttons. Each document is treated in isolation, i.e. as a distinct entity, and the only contexts of each entity is its content and the fact that it appears on the web. Isolating the hypertext documents from other documents it is linked to can inform us of the real context in which users see the document. Studying hypertext documents in isolation enables us to compare it with flat text. The question which arise from such an isolation is whether this definition of hypertext document can teach us new facts about the use of language in the context of the single and disconnected hypertext document.

Paragraph level - An interesting question regarding hypertext document is their layout. Some of the paragraphs in hypertext documents can actually be constrained by the layout effect the author had in mind. One of the most important factors in document layout on the web is the fact that it is controlled by screen size. This constraint might affect the size of paragraphs people choose to write. Then the first question regarding paragraphs in hypertext documents is if traditional, linguistic division controls the decision of where a paragraph should end and another begin, or have paragraphs become motivated by mere layout design? In other words, is there a paragraph division similar to the one existing in flat text. Assuming that paragraph breaks might be affected by font size, screen size, etc. what is their average length in words? Is it consistent with professional recommendations for screen layout? If it is consistent with these design recommendation then how was this convention formed. If they are not consistent with the recommendation then what is the reason for such a deviation? Can the subjects according to which paragraphs are divided be compared with flat text subject division? Are paragraphs formed around links, i.e. people write their text having the links they want to insert in mind? If so, what is the link density within paragraphs?

Sentence level - If the existence of paragraphs is in doubt then the existence of sentences as we know them can also be questioned. Tailoring sentences to surround links with context might result in new linguistic constructions. For that reason the position of links within sentences should be studied. Whether the link is introduced in the beginning, middle or end of the sentence might interfere with the comprehension of both the sentence and the anchor. If new information tends to appear toward the end of the sentence then we might find that many anchors are positioned at the end of sentences. The length of sentences in which links appear might also create different levels of sentence-anchor cohesiveness. Is the anchor an integral part of the sentence or is it a separate unit? What percentage of the sentences is actually grammatical? Is there a pattern in syntactic structure? If there is, can it be modelled? Such modelling can help in generating new sentences automatically. Another issue which arises from such a study concerns the ability to automatically analyse the syntactic structure of the sentences. Such an analysis would be able to identify the linguistic components of the sentence and their local relations.

Lists - A feature unique to hypertext documents is lists of links. These lists consist separate items, often ordered with a consistent method (subjects, alphabetically, numbered, etc.). Sometimes all the items are highlighted and other times the items are partially highlighted. In the latter case there is an interesting question regarding the structure of the partially highlighted items: whether there is a constant linguistic structure to those items and what this construction is. Another interesting question might be how distinct the listed items are from each other in word choice, POS, number of words, consistent structure, rhythmic structure, upper case and lower case character choice, etc. Studying the optimal versus the average length of lists can be used in the design of software menus and, of course, in designing effective hypertext lists.

Anchor wording - The wording of both highlighted items and anchors in sentences is one of the main concerns of this study. These are interesting because they allow a glimpse into the mind of the author and the way they make an association or a leap from one document to another. Studying the wording of non-highlighted parts provides supporting evidence to the way people write and linguistically design their web pages. Such a study can also aid in software menu design since it provides insight into the way people prefer to name their links and list items, which are similar in function to menu items. And last but not least, the effect of highlighted or partially highlighted lists on reading from the aspects of reading speed, comprehension, and scanning ability. Reading highlighted words might result in a priming effect, especially in the case of partially highlighted lists.

This dissertation is not designed to answer all the questions raised here but to persuade the reader that there is case for defining a new linguistic genre, the hypertext genre. In the following chapters some of the aspects presented here will be discussed and some more questions will be introduced. The discussion will be supported by evidence from the corpora collected and the statistical study of the data. Each question will be put in the context of the research community which would benefit the most from studying the aspect of hypertext discussed. Corpus analysis was chosen as a tool because of the immediate results such an approach gives and because we believe that a descriptive attitude would strengthen claims and observations made about hypertext. The aim of this dissertation is not to answer all the question presented but to promote new thinking regarding hypertext and to encourage the academic community to use the immense raw data resource that the web provides freely, friendly and always up-to-date.

Chapter Two

The Corpora

Collecting the corpora

The web is an almost unlimited resource for data collection which is ignored by many scholars. This unawareness is exhibited in the fact that researchers use the web as a working tool, to communicate with other researchers, but not as a rich structure which has to be scientifically explored. It is a highly varied collection of human subjects from all around the globe and it is the cheapest way to get as many specimens needed without actually investing in fancy equipment and worrying about "neutral" experiment conditions. The web is used by many people every day and the ability to reuse the data to perform analysis of usage and authoring conventions is an appealing one. Two corpora are presented here, together comprising 1000 HTML files. The actual collection of this data took less then one night of unsupervised computer work. The conditions under which these files were generated are varied and as diverse as possible. This diversity in environment might enable the results to be unbiased and neutral, since they resemble most the natural environment in which people actually author their web pages.

Collecting a sufficiently representative sample of web-pages was one of the most important aspects of this study. Since the Web is a multilingual-multinational virtual place there was a need to constraint the corpora to a specific language but not to a specific geographic place or domain. By not limiting the domain we were trying to get a random selection of files without losing the representativeness needed in order to model the web. The Parker Corpus was thus collected by performing a web search on a very common English family name. The files which were not written in English were then filtered out by hand. Having to work with such a large collection of files presented new problems. The first was the style of HTML markup which proved highly variable from author to author. The second problem was that on such a large corpus it would have been very difficult to look at each and every file separately. The compromise was to use one corpus for statistical manipulation and a smaller corpus for a more detailed study of manually authored web pages. For collecting the second corpus, the Home Corpus, Another constraint was introduced: limiting the domain to personal homepages. The two corpora were compared in order to be able to use the smaller one for the study of language conventions and the larger one for statistical analysis and for confirming representativeness of the smaller corpus.

The Parker Corpus

The Parker Corpus comprises 845 HTML files. In order to gather the files randomly, URLs were retrieved from the HotBot search engine by looking for the word *Parker*. The URLs were then fed into a program that "fetched" each file and placed them under one directory. The corpus was "cleaned" of error messages and, as much as it was possible, of HTML frames, which use different files for code.

Two methods were used to retrieve the anchors from the files. One method was a simple UNIX command (grep -i '</a' *.html ¦ sed 's;</[aA]>.*;;' ¦ sed 's;:.*>: ;') which created a problem because of the nature of the commands and the way people write HTML files: UNIX commands related to one line only, while markups for links were spread over two or even three lines. The other method was to use a Perl program in order to avoid the line problem. Neither method was perfect, but while the UNIX command gave less information than there actually was, the Perl program gave more information than was needed (such as font types etc.). The anchors retrieved with the Perl program were then cleaned by hand taking out HTML markers and extracting text from textual images.

The HTML files consist of 2,198,306 words (including the HTML tags), which are stored in slightly more than 18 Mbytes. From this corpus of HTML files we extracted a smaller corpus of anchors. This anchors corpus consists of 1.7 Mbytes, representing 133,341 words (stripped from HTML tags). It was assumed that using a name, such as *Parker*, to retrieve the URLs would bias the results, since most files were expected to be personal Homepages of some sort. But as the information accumulated we realised that this was not the case. Since the search was not aimed to any specific domain, the results presented here are as general as possible. The choice of the word *Parker* also allowed restricting the language used in the documents to English, since it is a very common English name. However, there is no guarantee that all the files are in English because we did not check each and every one of them, although we did our best to filter most of the non-English ones. Since the Parker Corpus is used here only for statistical measuring and comparison with the Home Corpus, the results from the URL search are clean enough to be reliable. We intend to gather more specific information in the future, and to try and formulate domain specific patterns. Such patterns could be the use of indexes and tables of content, the insertion of content anchors (as opposed to structure anchors) within documents, and the use of domain-dependent vocabulary. The results described here, however, outline general patterns and tendencies in HTML document design, maintained by Web users all around the world.

The Home Corpus

The Home-Corpus comprises 155 HTML files. This corpus is a collection of personal homepages sent to us by people from various mailing lists and news-groups who actually mailed us their URLs. The collection is as random as possible, since we did not know who would reply to our messages. The URLs were then fed into a program that "fetched" each file and placed them under one directory.

We chose to restrict the domain to personal homepages because this is a domain unique only to the web in the sense that it is an unwritten convention that each user who wants to author their own web page will begin with a homepage and will introduce themselves to the virtual community. Thus, the most common individual pages are homepages. Another reason for this domain restriction is the fact that the easiest way to get permission to use the HTML files and still maintain randomality is to post messages on the web asking for people to send their URLs with no control over the responses. However, we did filter out homepages that were pornographic in nature or belonged to neonazis or contained other offensive material in our judgement.

Since we wanted to study the Home Corpus more thoroughly we were looking for tools to manipulate markup. The best solution seemed to be the SGML tools created at the Human Communication Research Centre at the University of Edinburgh. SGML (Standard Generalised Markup Language) is an ISO standard (ISO 8879:1986) which supplies a formal notation for the definition of generalised markup languages. SGML is an international standard for the definition of device-independent, system-independent methods of representing texts in electronic form (Sperberg-McQueen & Burnard, 1993). HTML (HyperText Markup Language) is a markup language for hypertext which is used for compiling and displaying files on Web Browsers.

Since SGML is mostly used by experts it requires accuracy and errorless coding. HTML is used by many non-expert people and as such it does not require accuracy and it will be viewed through the browser even though it has many coding errors. This feature of HTML is user-friendly and many people are not aware of the errors their HTML code contains. In the context of gathering data for the corpora this feature of HTML ensures the representativeness of the files because users can be both experts and non-experts, regardless of their coding skills. In the context of converting HTML to SGML, in order to be able to manipulate the files with the SGML tools, this feature presents difficulties. The SGML tools require accurate and errorless code which means that the HTML conversion has to account for and correct HTML errors. While trying to correct (whether manually or writing a program to do this automatically) a 1000 files of HTML requires a lot of effort, correcting only a small portion would appear to be a wiser thing to do. For that reason we decided to convert only the Home Corpus from HTML to

SGML and to leave the Parker Corpus as is. The SGML tools provide easy access to the arrangement of the markup without losing information over line breaks and text layout on the screen.

The SGML conversion does affect the files themselves. In order to be able to use the SGML tools some of the HTML file had to be changed a great deal. We tried not to affect the layout of the documents i.e. not to alter the way people would view the pages through a web browser (which in our case was netscape 3.0). The manipulation of the files did change some of the internal structure of the documents. For example: the DTD (Document Type Declaration) for HTML version 3.2 that we used (Graham, 1996) does not support common coding errors such as <h1>something</h1> instead of <h1>something</h1>. This error is so common that it can be found in most of the HTML files we have! This kind of error does not change a thing for the web browsers and most of them would interpret both orders as being the same. The only problem is the convention which the people who wrote the DTDs try to introduce. The DTD we use is the one recommended by the W3C (http://www.w3.org/), which is the highest authority for HTML markup on the web.

There are probably not many users out there who are really interested in DTDs or in markup conventions and these users will write their code in the way they find most convenient. The problem of manipulating large corpora of HTML from diverse sources is only a minor one at present but we predict it will be a great one soon. Accounting the most common denominator (common and harmless coding errors) in writing DTDs and other parsers and compilers for HTML would prove to be the wisest thing to do in the near future. The following example is the list of errors which the tool for SGML validation outputs for one HTML file. We must mention that these errors do not create any ambiguity in parsing the file and therefore are viewed by the netscape browser as a normal web page. Assuming that the author who sent us the URL for file number 085.html actually viewed it before sending it to me, we would think that the list is too long for a file that both the author and the reader can read perfectly:

:13:16:E: there is no attribute "TOPMARGIN"
:13:134:E: there is no attribute "LEFTMARGIN"
:17:7:E: end tag for element "STYLE" which is not open
:21:14:E: there is no attribute "VALIGN"
:21:24:E: value of attribute "ALIGN" cannot be "TOP"; must be one of "LEFT", "CENTER", "RIGHT"
:21:36:E: an attribute value must be a literal unless it contains only name characters
:26:25:E: value of attribute "ALIGN" cannot be "TOP"; must be one of "LEFT", "CENTER", "RIGHT"
:37:13:E: there is no attribute "FACE"
:37:25:E: "FONT" is not a member of a group specified for any attribute
:38:3:E: document type does not allow element "P" here
:39:4:E: document type does not allow element "P" here
:40:4:E: document type does not allow element "P" here
:41:4:E: document type does not allow element "P" here
:42:4:E: document type does not allow element "P" here
:43:4:E: document type does not allow element "P" here
:44:4:E: document type does not allow element "P" here
:47:15:E: end tag for element "CENTER" which is not open
:57:29:E: there is no attribute "VALIGN"
:57:65:E: document type does not allow element "TABLE" here
:60:30:E: there is no attribute "VALIGN"
:75:23:E: start tag for "LI" omitted, but its declaration does not permit this
:75:27:E: document type does not allow element "LI" here
:77:51:E: end tag for element "A" which is not open
:79:26:E: start tag for "LI" omitted, but its declaration does not permit this
:79:30:E: document type does not allow element "LI" here
:82:26:E: start tag for "LI" omitted, but its declaration does not permit this
:82:30:E: document type does not allow element "LI" here
:84:50:E: end tag for element "A" which is not open
:86:27:E: start tag for "LI" omitted, but its declaration does not permit this

:86:31:E: document type does not allow element "LI" here
:89:67:E: start tag for "LI" omitted, but its declaration does not permit this
:89:71:E: document type does not allow element "LI" here
:92:48:E: end tag for element "A" which is not open
:94:26:E: start tag for "LI" omitted, but its declaration does not permit this
:94:30:E: document type does not allow element "LI" here
:97:26:E: start tag for "LI" omitted, but its declaration does not permit this
:97:30:E: document type does not allow element "LI" here
:100:26:E: start tag for "LI" omitted, but its declaration does not permit this
:100:30:E: document type does not allow element "LI" here
:105:26:E: start tag for "LI" omitted, but its declaration does not permit this
:105:30:E: document type does not allow element "LI" here
:113:7:E: document type does not allow element "CENTER" here
:115:3:E: document type does not allow element "TR" here; assuming missing "TABLE" start-tag
:119:3:E: document type does not allow element "H6" here
:120:2:E: document type does not allow element "P" here
:121:20:E: document type does not allow element "HR" here
:125:19:E: document type does not allow element "H2" here
:125:48:E: document type does not allow element "CENTER" here
:126:21:E: there is no attribute "VALIGN"
:126:31:E: there is no attribute "WIDTH"
:126:34:E: document type does not allow element "P" here
:129:41:E: document type does not allow element "A" here
:131:136:E: document type does not allow element "A" here
:131:144:E: general entity "T" not defined and no default entity
:135:68:E: document type does not allow element "A" here
:138:47:E: document type does not allow element "A" here
:140:79:E: document type does not allow element "A" here
:146:45:E: document type does not allow element "A" here
:149:20:E: document type does not allow element "HR" here
:151:20:E: document type does not allow element "A" here
:151:25:E: document type does not allow element "H2" here
:151:54:E: document type does not allow element "CENTER" here
:151:88:E: end tag for "H2" omitted, but its declaration does not permit this
:151:22: start tag was here
:151:93:E: end tag for element "H2" which is not open
:153:20:E: document type does not allow element "HR" here
:155:21:E: document type does not allow element "A" here
:155:26:E: document type does not allow element "H2" here
:155:55:E: document type does not allow element "CENTER" here
:155:88:E: end tag for "H2" omitted, but its declaration does not permit this
:155:23: start tag was here
:155:93:E: end tag for element "H2" which is not open
:157:20:E: document type does not allow element "HR" here
:159:20:E: document type does not allow element "A" here
:159:24:E: document type does not allow element "H2" here
:159:53:E: document type does not allow element "CENTER" here
:159:91:E: end tag for "H2" omitted, but its declaration does not permit this
:159:21: start tag was here
:159:96:E: end tag for element "H2" which is not open
:161:20:E: document type does not allow element "HR" here
:163:21:E: document type does not allow element "A" here
:163:25:E: document type does not allow element "H2" here
:163:54:E: document type does not allow element "CENTER" here
:163:90:E: end tag for "H2" omitted, but its declaration does not permit this
:163:22: start tag was here
:163:95:E: end tag for element "H2" which is not open
:166:7:E: end tag for "A" omitted, but its declaration does not permit this
:125:0: start tag was here
:166:7:E: end tag for "FONT" omitted, but its declaration does not permit this
:118:1: start tag was here
:167:7:E: end tag for "CENTER" omitted, but its declaration does not permit this
:113:0: start tag was here
:167:7:E: end tag for "FONT" omitted, but its declaration does not permit this
:55:4: start tag was here
:168:6:E: end tag for element "FONT" which is not open
:168:15:E: end tag for element "CENTER" which is not open
:173:3:E: end tag for element "P" which is not open
:175:14:E: there is no attribute "VALIGN"
:175:24:E: value of attribute "ALIGN" cannot be "TOP"; must be one of "LEFT", "CENTER", "RIGHT"
:175:36:E: an attribute value must be a literal unless it contains only name characters
:179:7:E: document type does not allow element "P" here
:180:10:E: end tag for "TABLE" which is not finished
:189:11:E: unknown declaration type "HR"
:190:11:E: document type does not allow element "P" here
:192:11:E: document type does not allow element "P" here
:193:11:E: unknown declaration type "HR"
:195:16:E: document type does not allow element "CENTER" here
:201:29:E: document type does not allow element "HR" here
:203:9:E: end tag for "FONT" omitted, but its declaration does not permit this
:187:8: start tag was here
:203:9:E: end tag for "CENTER" omitted, but its declaration does not permit this
:186:8: start tag was here
:207:22:E: there is no attribute "SRC"
:207:43:E: there is no attribute "HEIGHT"
:207:52:E: there is no attribute "WIDTH"
:207:62:E: there is no attribute "ALIGN"
:207:78:E: there is no attribute "CONTROLS"
:207:98:E: "MASTERSOUND" is not a member of a group specified for any attribute
:207:108:E: there is no attribute "AUTOSTART"
:207:118:E: there is no attribute "LOOP"
:207:122:E: element "EMBED" undefined
:217:8:E: end tag for element "CENTER" which is not open

:217:20:E: there is no attribute "FACE"
:218:7:E: document type does not allow element "CENTER" here
:221:11:E: there is no attribute "SRC"
:221:31:E: there is no attribute "HIDDEN"
:221:41:E: there is no attribute "LOOP"
:221:42:E: element "EMBED" undefined
:225:6:E: end tag for "FONT" omitted, but its declaration does not permit this
:217:9: start tag was here
:225:6:E: end tag for "TABLE" omitted, but its declaration does not permit this
:21:0: start tag was here

File 085.html (Home Corpus) as it is viewed by the Netscape browser.

Another problem we had to face was a very interesting one. Some authors inserted or invented their own tag names to the markup they use. This, of course, does not show on the screen of the web browser but it is a feature that we did not like losing. However, because of the reasons we gave earlier, there was not much we could do but commenting these new tags out. Such interesting tags were:

file 009.html:

general entity "increase" not defined and no default entity

general entity "show" not defined and no default entity

general entity "link" not defined and no default entity

file 014.html:

there is no attribute "TARGET"

general entity "pad" not defined and no default entity

general entity "ft" not defined and no default entity

general entity "frgb" not defined and no default entity

general entity "trgb" not defined and no default entity

general entity "df" not defined and no default entity

general entity "dd" not defined and no default entity

In order to detect how common these errors are in the 155 files of the Home Corpus, we manually checked and corrected all the files and marked files which contained more than 25 lines of error messages or had many tags which were not defined in the DTD we had. Then we calculated the percentage of the marked files out of the 155 files. We found 61 files which had more than the 25 lines of error messages or many "invented" tags. This means that 39% of the files had too many errors to correct manually. Although we did not take into account all the files which had fewer errors, we would like to mention the fact that most of the rest of the files had at least five or more errors. The reasons for emphasising this point are the following:

1. Assuming people view their web pages on the browser before they decide to release it on the web, compilation errors are ignored as long as the document layout they view on the screen is satisfactory.

2. The DTDs (or any other markup syntax parsing methods) used by commercial browsers to compile the HTML files on the web are much more robust and advanced than the ones used for research purposes. This leads to a divergence in the HTML markup usage because the more flexible the browsers become the less attention is paid to using the correct HTML code by users.

3. In future work with manually written HTML files a robust DTD should be used which would account for more errors than the one used here (the W3C published and recommended version). Such a DTD should:

    a) parse the HTML code regardless of tag ordering, as long as this does not create ambiguity

b) ignore unknown tags which are marked by < > bracketing, or preceded by & and end with space

c) interpret tags as being layout commands when they appear to stand on their own without any end-tag e.g. which is commonly used to insert a bulleted item (this requires some sort of preprocessing stage, to identify stand-alone tags)

d) assume a measure of error that can be ignored as long as it does not interfere with the display of the document on the screen

e) be appropriate for use with existing SGML or XML tools

The basic assumption in manipulating files taken from the web should be that people know what they want to see and have done what is sufficient to achieve their goals. Thus we need only to observe and interpret what they write. Trying to educate the users might result in a higher quality of web documents but will also decrease the quantity of unconstrained authors and users. One of the reasons for the flourishing of the web is that putting a personal homepage on the web does not require high level programming skills. Authoring a web document involves a lot of verbal creativity and some insertion of layout commands in the form of HTML tags in the limits of the hypertext convention. Since part of the motivation for this study is the wide usage of the web, the "neutral" environment, which is a crucial element in this study, will be damaged if the real world will be fitted into the constraints of the research environment i.e. the limitations of our compilation techniques. In the future we intend to improve the DTD used in order to be able to gather and manipulate a larger corpus of web pages.

comparing the two corpora

The issue of similarity between corpora is a very interesting one. Kilgarriff (1996; in press; et al. 1996) suggests many different measures and surveys the current techniques for comparing two corpora of text. The techniques suggested are based on the study of flat text and include comparing χ^2 (chi square) statistical measurements for words which occur more than 5 times in both corpora: Mann-Whitney ranks set, which compare frequency of occurrences of word samples from both corpora; Mutual Information statistical measure (Church and Hanks, 1989); Log-likelihood (G^2) (Dunning, 1993); and the Information Retrieval method which measures the co-occurence of related terms in both corpora. Kilgarriff (1996) suggests applying the clumpiness of words, identified by Katz (1996), and Poisson Mixtures (Church and Gale, 1995), to measure corpora similarity. These methods account for both the global frequency and the local density of words in a given document. The measure suggested is to try to compare both of these behaviours for different words in both corpora. Another measure suggested is of content and domain comparison. This comparison is not based on any lexical similarity measure but takes into account the authors of the documents, the target audience and the purpose of the document.

All of the methods for comparing lexical distribution between two corpora assume that there is a "normal" behaviour for words within a text. The problem, of course, is that the behaviour expected in these cases is within flat text documents, which is completely different to the behaviour of words within hypertext documents. To see that this is so, consider the following: Both Church and Gale (1995) and Katz (1996) describe a pattern of behaviour for words in flat text: if a word is a topical word then it will appear more than once in the document and in a clump. This assumption was empirically validated, in the context of text segmentation, by Richmond, Smith and Amitay (1997). Hypertext, however, has a property which allows people to mention topical words only once without losing the cohesion of the document. This means that when trying to identify clumps of identical topical words within a hypertext document, the behaviour of the word frequency might be different. Since lists and indexes are topical words organised in a methodical way, and very often are not repeated in the content of the document, there is a need to test the assumptions made for text, on hypertext. Topical words might reoccur in related or linked documents, but if the documents are collected randomly and are treated as once-occurring HTML files, then the word frequency behaviour should be expected to be different from that of words in flat texts.

A good example of this different behaviour of words in hypertext is files number 2 and 10 of the Home Corpus. Both files contain only one (!) occurrence of the word *the*. In the case of file 010.html there are a 100 words (not including tags), and in file 002.html there are 64 words (again, not including tags). In the British National Corpus (BNC - more than 100 million words) the word *the* is more than 6% of the whole corpus. This is a very impressive percentage which, if dealt with flat text, is part of the linguistic and grammatical conventions of producing the language which is documented in the BNC. However, looking at files 2 and 10 of the Home Corpus the usage of *the* is in naming places (“*The Netherlands”*, file 2) and projects (“*The Migraine Project”*, file 10), where the determiner is part of the proper name. There was no other mentioning of the word *the* in these files.

There is a case here, then, to use a method which is less lexically dependent, until a better understanding is achieved regarding the behaviour of words and their frequency in hypertext documents. The methods used here compare authors, audience, content and the statistical distribution of anchors and lexical bigrams (which would indicate content).

	Parker Corpus	Home Corpus
authors	Unknown	people who read mailing lists and newsgroups
audience	other Web users	other Web users
content	mostly homepages, but the randomality of the retrieval yielded some other sorts of pages	homepages
No of files	845	155 (152 after cleaning three files out because of SGML errors)
No of words (including tags)	2,198,306	65,611 (equivalent to 2.98% of the Parker corpus)
words/anchor	2.89	2.74
anchors/file	35.9 (median = 14-15)	17.73 (median = 13-14)
20 most frequent bigrams in the anchors	*home page return to *back to *of the more info to top issue # table of of contents *on the *to the info request to table *in the *click here go to top of and the the internet science fiction	*home page university of department of cognitive science centre for of edinburgh for cognitive *of the natural language *click here easter island computer science research interests language processing computational linguistics *back to artificial intelligence *to the *on the *in the

Let us look at the above table and examine the results. Since the Parker corpus was collected randomly from the web it was not possible to retrieve only homepages. However, because of the name used to retrieve the URLs - Parker - the likelihood of getting back URLs of personal pages of people named Parker, was very high. As can be seen in the list of bigrams, in both corpora the phrase "home page" is the first to appear. The comparison between two statistics of words/anchor and anchors/file is reinforcing the findings from the Parker corpus. The words/anchor statistics indicate an average of almost 3 words per anchor, and the anchors/file indicate an average of between 13-15 anchors per web page. Such an average is of course more flexible but the figures in the table show that even with such a flexible parameter there is consistency. This consistency is very remarkable. It means that there is a well established convention regarding the length of anchors and their density within a document.

The bigrams indicate content. The bigrams which are given here are the top 20 bigrams from the extracted anchors of each corpus. The reason for using only the anchors and not the whole of the HTML corpora is simple: since the latter contain markup which would probably interfere with the result we chose to examine only the anchors. Another interesting consideration was the fact that most of the anchors contain more accurate information regarding the content of the page and the links people choose to put in it. As it can be seen in the table, the content of the Home Corpus is more homogeneous. Many of the links are related to research and academic institutions, especially in the area of natural language processing and cognitive science.

However, if only the words which appear in both lists are examined an interesting pattern is revealed: There are 6 bigrams which appear in both lists: the first in both is "home page" which indicate that most files are linked in one way or another to a homepage of some sort. If the actual use of this bigram is studied then a linguistic conclusion can be drawn: In the singleton frequency list the word *homepage* is at the 60th place in the Parker Corpus anchors frequency list and at the 68th place in the Home Corpus anchors frequency list. This means that the compound word *homepage* is still not as commonly used as the word combination *home page*. This interesting formation of a new compound word can be studied from the web resources suggested in this dissertation. There are probably many new compounds that are being formed daily on the web and a diachronic study of language could benefit from investigating this vast resource of linguistic examples and information.

All of the bigrams shared by the two lists concern navigation aspects and directions on the web. The seven bigrams are:

home page
back to
click here
of the
to the
on the
in the

Three of the bigrams are unique to web pages and an assumption can be made that they will not appear in flat text which has nothing to do with the internet. The reason for not finding more instructional combinations in the Home corpus, as can be found in the bigram list of the Parker Corpus, is probably the homogeneous collection of files in the Home corpus. Since the Parker Corpus comprises more than 5.5 times the data in the Home corpus, the navigational instruction, which are shared by most of the files, become more significant. Nevertheless, both *back to* and *click here* remain high in the list of bigrams for both corpora. This indicates that the two straightforward instructions are inseparable from the navigation activity between documents and that authors prefer literally directing their readers towards their interests instead of using the "normal" linguistic tools. This kind of direction indication and iconic wording is unique to web navigation. A more detailed discussion about the content and context of anchors within the Home Corpus is given in chapter five of this dissertation.

From this point onward most of the analysis of data will be performed on the Home Corpus. In some cases data from the Parker Corpus will be used to emphasise an idea or to strengthen a claim. It is thus assumed that the Home Corpus is a representative sample of the overall 1000 files and that therefore it can be used for characterising the convention in writing hypertext documents.

Chapter Three

Hypertext Usability and Design

For more than a decade there has been a growing interest in the study of hypertext. Since the mid-eighties there have been hundreds of articles and studies published involving the theory and application of this multi-dimensional electronic text. Some of the articles published study the cognitive aspects of hypertext, others examine its practical aspects. However since there are so many hypertext systems, designs and theories, it is not easy to see the wood for the trees.

Chen and Rada (1996) compare 23 published experimental studies and doctoral dissertations on hypertext, which appeared between 1988 and 1993. They compare users, tasks and tools, finding many discrepancies between results. In their discussion they point out that

> *"...experimental designs vary as the hypertext systems vary. Some systems have only the most primitive hypertext facilities implemented. For instance a page of a document is displayed on the screen at a time and hypertext links are provided at a page-to-page level. Some systems are advanced and include graphical browsers, a persistent window to display the current focal point, footprints, and search facilities."*

They note also that users' characteristics affect the results and that these characteristics should be addressed in future research. The authors conclude by inviting researchers to formulate precedents for hypertext experiment designs. Since the studies considered were published, there has been no substantial change in experimental design. Considering, for instance, McDonald and Stevenson (1996; in press), Stanton (1994), Baron, Tague-Sutcliffe, Kinnucan and Carey (1996), and Wenger and Payne (1996), it is clear that little effort has yet been made to compare such experimental conditions. It is true that benchmark measures had already been formulated by Anderson, Berre, Mallison, Poerter and Schneider (1990) but these are not widely used, and are inadequate for recent hypertext systems.

Theory and implementation of hypertext design also suffer from lack of consistency. Theoretical papers such as Recker and Ram (1994), and system design illustrations such as Nanard and Nanard (1995) and Maioli and Vitali (1994) refer to an undefined user, ignoring individual cognitive differences, and variations in knowledge domains. Nygren, Allard and Lind (1995b), and Nygren (1996) point out that hypertext design should aim to improve the ability of skilled users to navigate within the systems. Nygren (1996) defines the difference between the design for novice and the design for skilled users:

"A design fit for the untrained user will focus on ease of learning, while a design fit for the expert will focus on efficiency in daily use."

She also claims that the studies done today do not satisfy the demand for better daily user interface design, but aim instead toward the novice user. Since the use of computers is very widespread, and since menus and hypertext are now an integral part of much simple software, we would like to focus on this audience of skilled users.

Being an overflowing source for HTML files, the web enables researchers to collect material for their studies while constraining their experiments without biasing the experiment's conditions and results. The web is probably the most direct channel to people's minds and houses. In the corpora collected there are a 1000 files, which are assumed to be equivalent to an almost as large collection of human subjects in a given experiment of writing hypertext documents. The data gathered is already typed electronically and arranged in the most natural and neutral way possible, in the subjects' normal environment. All of this is achieved overnight and without causing any interference in the final outcome.

A data intensive linguistic approach to finding design conventions on the WWW can be very appealing if we consider the number of participants, in one of the largest and most comprehensive experiment ever to be made in HCI. There are many questions relevant to the interface design community which can be answered very easily with the most simple statistical tools.

One of the most sought-after answers, to a question presented in various mailing lists, is how many outgoing links (as opposed to internal links) there are in an average Web document. One of the problems with such a question is that the definition of outgoing links and internal links depends on screen size, font size, and page design. Any given link might imply a physical leap from one document to another, or from one chunk of text at the beginning of the document to the acknowledgements at the end of the same document. For this reason we would like to include all links, whether outgoing or internal, in our analysis, making no distinction between their potentially different destinations. This approach to the non-existing distinction between links' destination is also taken by Myka and Güntzer (1994), who state that

"in addition to implicit links, many documents contain explicit references to other parts of the same document or to completely different documents. Linking these references to the referenced document part supports the controlled navigation of a user who, otherwise, would have to brows sequentially through the pages".

The above question - how many links there are in an average document - was the first to be addressed. At first we calculated the overall number of links in the Parker Corpus, including textless icons. The number we have computed seems to be quite high: 36.87 links. Looking for deviations from this number we have found within the Parker Corpus one file which consists of more than 2100 links! It is interesting to note that the documents containing these links presented, then, not as a list of items, but as embedded anchors within the text. There are many documents with more than 200 links, and of course, there are some documents with no links at all. Calculating the average number of verbal-links per page yielded the result of 35.9 links per page. A verbal-link is usually marked in the HTML language between two angled brackets. When the text appears inside an icon it is usually specified in the markup as ALT="text". We extracted this information manually from the corpus.

Since the average number of links per electronic page was much higher than expected it was suggested that the median Number will be taken. The median is 15 links per page and this seems intuitively correct. It should be noted, however, that text-to-hypertext systems such as LaTeX2HTML insert 12 links automatically without any non-linear, real hypertext links. These links are usually 3 buttons (*previous*, *next* and *index*), and 3 titles corresponding to the pages which the buttons refer to. The 6 links appear twice on the page - top and bottom, regardless of the length of the actual segment of text.

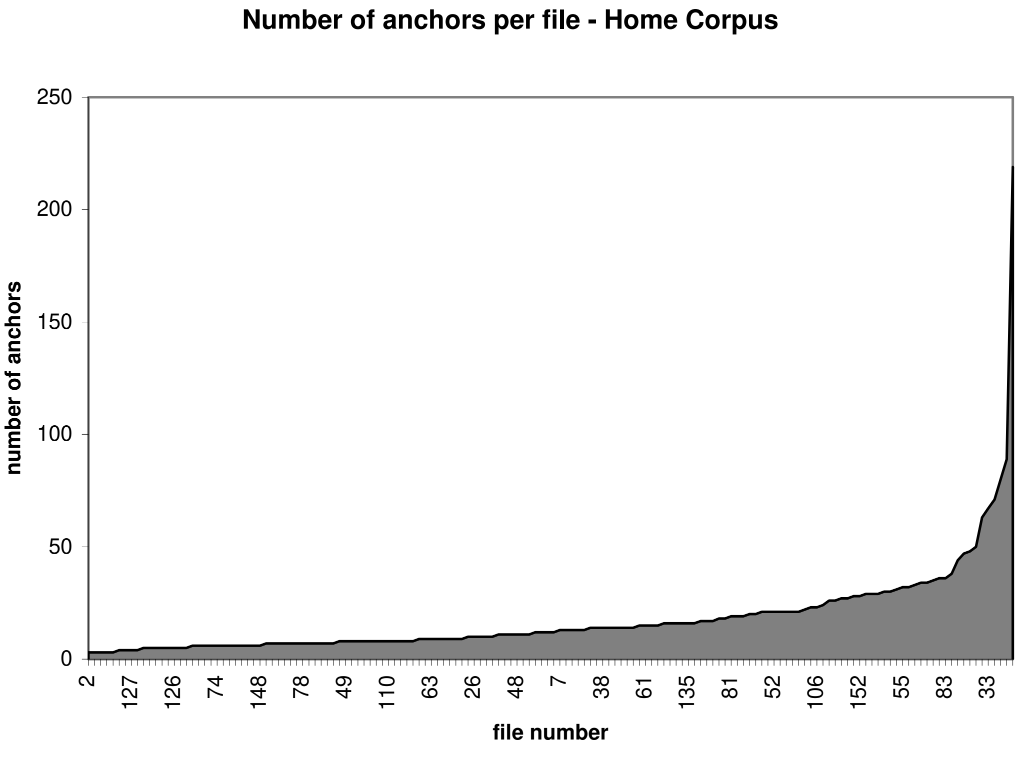
The next question to be addressed was: How many non-verbal icons are used in the average Web document? The answer to this question is quite surprising, just 2.6% of the retrieved anchors were actually textless. The definition of textless is that there are no printed characters on the icon. This finding is very interesting because it demonstrates that the convention precedes the science: In a recent study by King, Boling, Annelli, Bray, Cardenas and Frick (1996) it is concluded that

> *"...buttons with both pictorial symbols and text labels resulted in significantly less user confusion than did buttons with pictorial symbols only. Buttons with text labels only also produced significantly less confusion, compared to those with pictorial symbols only".*

It is unlikely that any of the authors of the 845 HTML Parker Corpus files has ever read an article related to hypertext design. Nonetheless, it is suggested that the use of "ISMAP", which is a clickable map of textual and pictorial icons, is becoming very common. It is possible that this combination of text and picture will increase in use at the expense of textless icons. One of the reasons for the low percentage of the use of images as links might be the loading speed. Since loading documents with many images slows their retrieval time, the authors of Web documents usually try to control this phenomenon by inserting fewer clickable images to their creations. Another aspect of the loading speed is that if links are to be presented as images they might appear after the texts appear; or in very busy hours of the day it may be the case that the images will not appear at all. Users are apparently well aware of the problem and the consequence is that fewer images are being used as anchors.

The number of words per anchor might be one of the most important findings of this study regarding design issues. One of the major distractions in hypertext designs can be the length of the phrase chosen as anchor. We found that the average in the Parker Corpus stands at 2.89 words per anchor, and this time there were no large deviations from the average.

Repeating the experiment on the Home Corpus gave very similar results: an average of 17.73 links per page (excluding textless images), and the median of 13-14 links per page. The average is not as high in the Home Corpus as in the Parker Corpus, probably because of the fact that homepages do not usually contain large indexes but references to documents which are a collective index, while the Parker Corpus probably include some indexing pages. Within the Home Corpus there is one file containing 219 links and the second highest contain 89 links. There are no files which contain less than 3 links. The chart below shows the distribution of links per files in the Home corpus.



The other finding supported by the repetition of the experiment on the Home Corpus, is the number of words per anchor. This time the number of words per anchor is 2.74 and it is consistent with the previous figures calculated for the Parker Corpus. These results validate the claim that there is a well defined convention regarding web pages design and that those conventions should be reflected in professional hypertext designs

People tend to make the message short, and more important, they tend to distinguish it from other texts or anchors by using nouns, proper names, and adjectives, omitting determiners, prepositions and apostrophe 's'. The use of determiners is quite consistent: the usage definite article is used particularly if the anchor refers to a proper name such as a title of a book, say "The Wind in the Willows". Determiners are omitted from the anchor when they are not part of the proper name, or when they might distract the reader from the core idea. Verb usage is also intriguing and there is a tendency to use imperatives and present participle morphology such as '*climbing*', '*jogging*', '*reading*', '*downloading*', etc. These forms appear very often in vertical lists of links, apparently to shorten the message and to make it as compact as possible.

This observation is supported in another field of HCI - menu design. The similarity between the wording of anchors and the wording of menu leaf information is remarkable. Norman (1991) describes in his book, "The Psychology of Menu Selection", the linguistic attitude that should be taken when designing a menu:

> *"A number of guidelines have been suggested for the wording of leaf information. Alternatives should be unambiguous, mutually exclusive, exhaustive, and nonoverlapping. The wording of alternatives should emphasize the critical features of their differences rather than the similarity. Furthermore, it has been suggested that alternatives should be brief noun phrases (Shneiderman, 1987). Most of these guidelines have yet to be validated by empirical research."*

Since the problem of anchor wording is similar to the problem of leaf wording in menu design, we would like to suggest that this study is as relevant to menu design as it is to hypertext design. The community of interface design might benefit from studying the way people choose to list items in their hypertext documents, and from the naming of these objects. This process of categorising items is very similar to creating a local hypertextual menu, which is a basic individual interface, introducing a new local environment within the web. In chapter five this linguistic environment is studied from the syntactic and pragmatic aspects. Baron, Tague-Sutcliffe, Kinnucan, and Carey (1996), and Trigg (1983) also present models, which try to define a taxonomy of link types. In the light of their research such taxonomy would aid designers to create a more human-like environment. Another community which could benefit from this kind of study is the one which studies the online technical writing. By learning the linking patterns people are capable of making, a better understanding of indexing methods and users' needs can be achieved.

Chapter Four

Hypertext and NLP Techniques

This chapter illustrates the current Natural Language Processing (NLP) techniques which apply to hypertext. These techniques are examined in the light of findings from the corpora collected and some comments and suggestions for improvement, regarding these techniques, are made. The subject of Part Of Speech (POS) tagging which is discussed in chapter two is not dealt with again in this chapter. However it is assumed that this discussion is relevant to this chapter as well and therefore some reference is made to it.

Text-To-Hypertext

The task of converting flat linear text into hypertext is one of the most challenging problems in the Information Retrieval and Information Extraction communities. A good solution to this problem can facilitate the conversion of hardcopy libraries to virtual ones in a fast and efficient manner. This conversion will allow faster and more convenient access to information. In this section we would like to discuss the existing methods for converting text to hypertext and then to analyse their underlying assumptions in the light of the findings from the corpora collected.

One of the main differences between a flat linear document and a hypertext document is that the former is usually self-contained (although see Myka, Argenton and Güntzer (1996) who claim that the structure of some flat linear documents is broken by internal references to page numbers and footnotes, but they also claim that this structure is part of the linear structure conventions). Hypertext, on the other hand, needs "supporting" linking to other hypertext documents in order to give the same full information (and sometimes even more than that given by flat documents). This is mostly true where hypertext documents are generated manually. It is not true when they are created automatically with the current state-of-the-art techniques. Since there is a lack of research in studying this particular difference between hypertext and flat text, researchers assume that there is no real difference between the two, and that the structure of flat text can be used, exclusively, to determine the structure of the converted hypertext. The following discussion and survey are an attempt to partially bridge this gap.

Assume that there are a number of stages that a text has to go through between "being" a text to "being" hypertext: The first stage is segmentation. Segmenting linear text is a very common problem in IR and it is solved with different techniques. Some of these techniques are explored by Hearst (1994), and by Richmond, Smith and Amitay (1997). Segmenting linear text into smaller subject-based units enables better linking between documents, based on subjects and concepts found. Another motivation for segmenting the flat text into smaller units is the physical limitations presented by the online electronic environment. There is a technical constraint which allows the reader to see only a portion of the document at a time (usually a 14" or 17" screen). And thus it is not efficient to read a 300 pages book in one piece i.e. to scroll the screen along 60 or so meters of text! And therefore, in order to put some order into the reading process, each flat text should be divided into many small sections which would allow the reader to navigate between them. Shneiderman (1989) explains in his article that:

> *"A major concern to authors of hypertext databases is determining the optimal length for documents. Research suggests that many short documents are preferable to a smaller number of long documents. An experiment was performed at the University of Maryland using the* ***Hyperties*** *system in which the same database was created as 46 short articles (from 4 to 83 lines) and as 5 long articles (104 - 150 lines). Participants in the study were given 30 minutes to locate the answers to a series of questions by using the database. The 16 participants working with the short articles answered more questions correctly and took less time to answer the questions."*

Carlson (1989) tries to limit this segmentation by claiming that [*not all texts are suitable for hypertext representation. Hypertext makes the deep structure of a knowledge domain explicit.... Although there is no universal model for hypertext implementation, experience suggests that if the document is closely interwoven through rhetorical devices, then decomposition into chunks and links will be difficult, with loss of information and confusion of meaning of a potential result. For some documents, this conversation is either impossible or not desirable because it destroys the subtle interconnections of theme, argument, metaphor, and word choice*.] This claim is not based on evidence but on a suggested *experience*. The problem of segmenting what is called *expository text* is the heart of the segmentation algorithms introduced before. These techniques make use of discourse markers as well as detection of topical words.

The second stage in this process of text to hypertext conversion is finding the place from which one document would be linked to the other. This can be done in many ways. The following paragraphs present a survey of linking-between-document techniques and approaches:

In his MSc thesis Blustein (1994) identifies three common approaches to text to hypertext conversion. The first approach is the Highly Structured Document approach which makes use of markup languages such as SGML and link type markup inserted by manual preprocessing. The linking between documents is done on the basis of similar markup as well as similar semantics of markup. The weakness of this approach is in its dependence on markup. The first problem is that it requires a lot of supervised work by skilled people. There is also the problem of standard markup which demands an agreement between coders with regard to the convention they will take in marking the text. Salminen et al. (1995) use indexing methods for finding markup, then listing them in an index from which links are created. The index words, which are named index-elements, are therefore part of the markup and have to be encoded into the text.

The authors assert that the conversion is not always a straightforward process and that it is time consuming and costly.

The second approach described (Blustein, 1994) is the AI approach. This approach deals with documents from a restricted domain which are analysed on the basis of a knowledge base. The domain restriction eliminates real world linking which people choose to do in almost every document they author. This kind of restriction is damaging to the richness of hypertext documents created with the AI approach. Since with this approach the text has to be parsed, it requires grammatically correct text. This requirement is very unrealistic if we take the real world hypertext which is produced by web users. As mentioned in chapter two, the web has its own grammar conventions and these are generally not compatible with any training corpus publicly available for research today. The links created on the basis of the knowledge base depend on the information encoded in the knowledge base and each concept has to be inserted into the knowledge base in order to be considered a candidate for linking between documents.

The third approach is the IR approach (Bernstein 1990; Salton et al. 1994, Blustein, 1994; Golovchinsky 1997). This approach is not domain restricted and does not need structured documents in order to understand their information structure. The IR community is tackling the problem of hypertext by performing pattern matching between strings of words in different documents i.e. the more similar patterns are found the more likely it is that there is similarity in content and therefore there is a reason to assume that the documents should be linked. The basic assumption underlying this approach is that a link between documents should be created only if they share the same topic. However, there are many other types of links between hypertext documents and these are overlooked by the IR community. For example, a link which would explain a whole concept discussed briefly in one sentence, or a link which would lead to an action such as downloading a software or sound effects. A suggested restricted typology of hypertext links can be found in Trigg (1983). Another problem this approach faces is the need to limit the number of links created because of the endless possibilities of pattern matching. This means that as the number of documents processed increases the number of potential links increases too.

Golovchinsky (1997) took only capitalised nouns, generally proper names, for creating new anchors in online documents. This treatment might be a good solution for choosing only topical words and proper names. The percentage of words which appear in initial upper-case character in the Home Corpus is 67% out of all the words in the anchors. This is a very high proportion and constraining the words chosen to be anchors to almost 70% upper-case-initial proper nouns might prove to be a very good intermediate answer to this problem. Moreover, 78.3% of the full-phrase anchors actually begin with an upper-case character, leaving prepositions and determiners with initial lower-case characters. This seems, then, to be a good indicator to what is a best candidate word for being an anchor. Notice that in the following examples from the Home Corpus all the words but *of, for, in* and *the,* i.e. function words, are capitalised. Thus an interesting solution might be to detect strings of capitalised words in the text as highly probable candidates for anchors:

Department of Computer Science (040.html, Home Corpus)
Ancient Scripts of the World (066.html, Home Corpus)
Need for Work Management in the Utilities Industry (085.html, Home Corpus)
Information for All (096.html, Home Corpus)

This is only a partial solution and there are many other problems arising from it of course. Such problems are vocabulary choice differences between different authors (which would make pattern matching difficult), lexical ambiguity, and placement of the link in the right context (see Carter 1996). The last problem is probably the most difficult one because if the document contains the word more than once there is a need to understand the context in which an anchor would be most appropriate and coherent.

All three approaches discussed here may also use synonymous information to find more possible links. Synonymy can be found at the word level as well as at the markup tag name level. This expansion of concepts often leads to a greater number of linking options and then there are still more decisions to be made as to the nature of the links and to their context. Using additional thesaural information does not necessarily aid in determining document connectivity because it creates more potential places to link between documents and it can also add ambiguous information to the existing one.

Another approach which is not discussed in the above mentioned thesis is the page layout approach. This approach, described in Myka and Güntzer (1994) and Leibowitz-Taylor et al. (1994), concerns the inner-linking within lengthy flat documents. The authors make use of document layout, such as titles, paragraphs, pictures, and numbered lists in structured documents, to locate a potential link. They also try to tackle the problem of segmenting such lengthy documents by means of document layout. This approach is a very simple substitute to the first approach described by Blustein (1994), that is the Highly Structured Document approach, only this time the markup is amalgamated in the graphical layout of the document. This document layout approach answers the need for dealing with documents which were only processed by an OCR program. This technique also makes use of font types and bold print, identifying these as better candidates for anchors. One problem with this approach is that it depends on the linear structure of the documents and that it maintains this very same structure after the text to hypertext conversion (Myka, Argenton and Güntzer, 1996).

Following Richmond, Smith and Amitay (1997), a combined approach might be taken. The authors suggest a simple algorithm for detecting subject boundaries within flat texts. The algorithm also finds topical words in each segment, enabling the detection of content and context. It also disambiguates the necessary words by allowing access to their immediate context. The topical words found can be used both for indexing the page and marking potential words for anchors. This method can be used to break the structure of the flat text and free the segments from the hierarchical ordering of the flat document. A summarisation technique might be a good solution for changing the syntax from that of flat text to the one of hypertext, shortening and simplifying the sentences when needed and making the anchors coherent with the text. The problem of action-links, or expansion-of-idea links creation can be partially solved by combining the AI knowledge database approach.

Hypertext Generation

Generating online hypertext systems such as online tourist guides, telephone directories, university courses etc. can be useful to both the users and the operating companies or institutions behind those systems. Such online systems can aid in creating the online classroom and learning environment of tomorrow. These systems are supported by a knowledge database and a mechanism for generating grammatically correct and coherent texts. Most of the attention in today's generation systems is paid to discourse planning, knowledge representation, user modelling and phrasal lexicon planning. There are no knowledgeable adjustments in the techniques used for flat text to hypertext generation but researchers are intuitively adapting their systems to conform with what they know about hypertext, being users and authors of hypertext themselves. Milosavljevic et al. (1996), for example, write that hypertext generation means that changes have to be inserted into previous models of text generation.

Researchers in the field of Natural Language Generation (NLG) emphasise the benefits from hypertext generation (Dale et al., 1997; Milosavljevic et al., 1996; Moore, 1995; Reiter at al., 1992). Among the advancements of applying generation techniques to generate hypertext on the web they enumerate the following facts: the ability to relate and link directly to other available resources on the web; using the already available interface, i.e. browser, to view the generated hypertext without designing a new environment and a new system; being exposed to a varied audience without any additional cost or trouble; using the multi-modal display of the interface which allows an easier segmentation of the text into smaller units, facilitating the navigation process; and being able to use discourse history as an online indicator for users' needs and as a feedback for the system.

The NLG community is still assessing the possibilities and facilities, provided by the WWW, for testing different environments for users. The question of tailoring different texts for different users is a very crucial one in designing hypertext and most of the research today is examining this question from all its aspects. Dale et al. (1997) distinguish between textual coherence between pages and within pages. They bring up the question of how to insert and create unambiguous links and anchors in the right context, since the same wording might appear in different contexts.

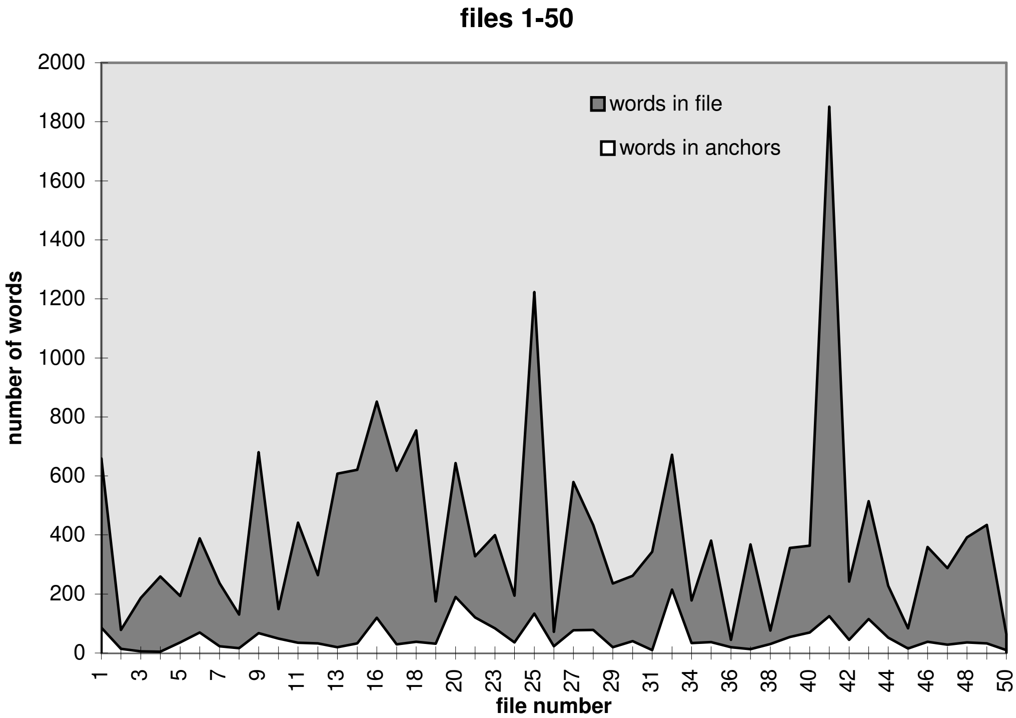
The next chapter tries to look at what context people provide in their homepages for links and anchors, and some questions that are raised by both research communities, Text-To-Hypertext and NLG, will be examined. It is suggested here that research of hypertext as a genre would be beneficial to both communities and that there are many questions which could be answered by a statistical study of the web and its users.

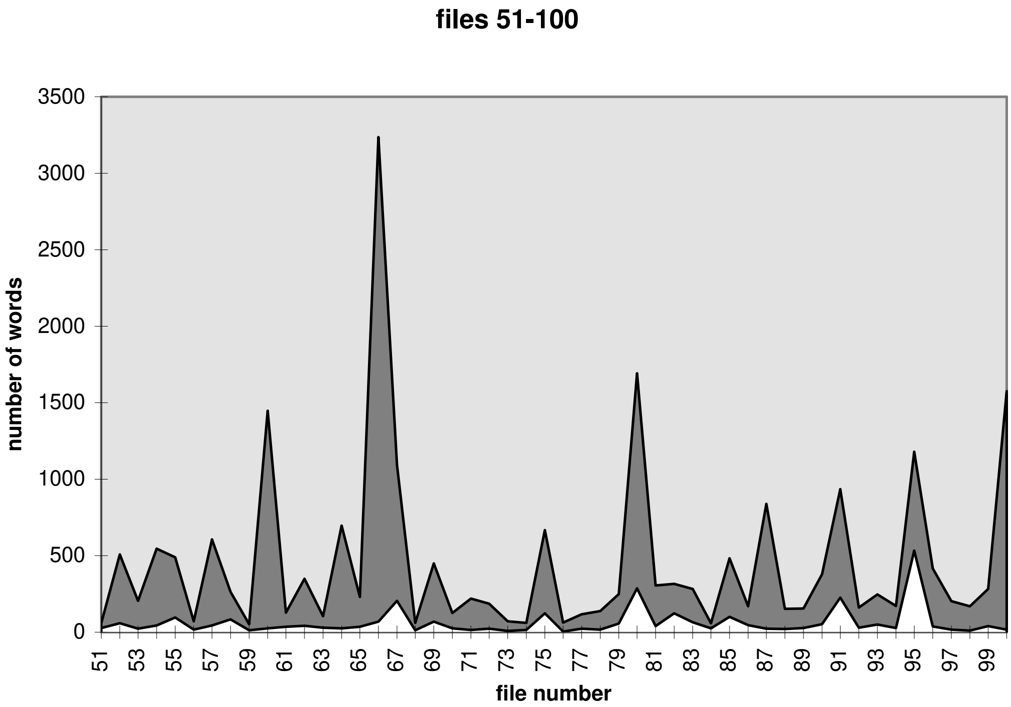
Chapter Five

Anchors and Context

In order to better understand the linguistic environment in which people choose to put their hypertext, a detailed study would examine the content of each hypertext document, trying to infer the context in which they were created. In this chapter the content of hypertext documents will be studied through a statistical analysis of the words which consist the HTML files of the Home Corpus. These words were divided into two groups: the words which appear on the screen (excluding, for technical reasons, the words which appear in image form), and the other group, which is a subset of the first, includes all the words which form the anchors. The number of words consisting the first group is 48,963. The second group has 7391 words.

The ratio between number of words and words in anchors is first to be calculated here because this might be the place where this information would help us understand the real context of anchors. The ratio between all the words in the file (grey) and the words in the anchors (white) is shown for each file from 001.html -100.html (Home Corpus) in the graph below:





As can be seen from the graphs, there is, in some of the files, a visible increase in the number of words used as anchors as the number of words per file increases. However this is only a local tendency and it is *not* linearly proportionate with the number of words. This can be explained by the fact that the number of links per page ranges between 3 at the lower end to 219 at the highest end. The range of actual number of words in hypertext files is much larger and stretches from 24 to 3167 words in one file. This range is very large and therefore the calculated average of links per file is probably more useful than that of links per a certain amount of words. The average, as mentioned in chapter two, is 17.73 links per page, or if the median is preferred then 13-14 links per page.

This is an interesting finding because it means that the length of the document does not dictate the number of links and it also means that links' density is not a factor in inserting links into the hypertext document. This raises a very intriguing question: If link density is not a constant factor in writing hypertext files then what causes people to insert links into the text? Finding the answer to this question might also solve the problem of context presented in previous chapters. If the research community could have an answer as to how many and in what context links can be created then the way to creating text-to-hypertext systems, as well as hypertext generation systems would be practically solved. The question of whether or not density is a factor in creating new links can be addressed by the above charts, and the answer is that there is no linear correlation between how many words the document contains and how many links can be created. The question of the context in which the links occur is the main issue of this chapter.

General linguistic environment of hypertext documents

Defining the general content of hypertext documents might seem ambitious to the reader but the purpose of this definition is to highlight the common linguistic factors which contribute to the creation of hypertext documents. Two strategies were taken in finding the linguistic environment in which hypertext documents are written. The first is studying frequency lists of words from both the anchor and other words in the files. The second is a detailed study of examples from the Home Corpus.

The following list is a frequency list of words which appear throughout the Home Corpus files. The list analysed here is only the top 35 words:

2215 the	303 at	185 as
1633 of	281 you	177 that
1335 and	277 this	177 have
1003 in	266 with	176 research
957 a	231 university	165 m
922 to	231 page	165 be
627 i	220 are	161 here
518 for	205 from	158 or
459 is	204 by	149 home
409 on	194 an	139 information
362 my	192 it	130 me
331 s	190 language	

There are many interesting fact which could be extracted from this short list. The personal voice people use in writing hypertext is one of them. Since the Home Corpus is a collection of personal homepages it is not surprising that the words *I, my* and *me* would be very high in the list. However the word *you* is also placed very high thus indicating a tendency to use direct and informal language - *from me* [the author] *to you* [the reader]. This informal environment also extends to a physical location: *here* , *home* and *page,* and to some extent the word *this* also indicate placement. Later in this chapter it will be shown that the words *page, home* and *here* are part of a sequence of words which name their context more specifically.

As mentioned before, since the Home Corpus was gathered from people who subscribed to mailing lists and newsgroups, the authors of the Home Corpus seem to have a homogeneous background. They seem to be involved in *research* in the context of a *university*. The words *language* and *information* can also be part of the academic world but since these particular words could also indicate other things they can not be put into one group or another. The word *language* can appear in the context of computers as well as in the context of linguistics and literature. The word *information* can also be associated with computers, but can also be part of the hypertext environment where the authors supply a link to more information on the subject discussed.

The other words in the list are mainly function words and when compared to the top 35 words of a frequency list taken from the British National Corpus (BNC - more than 100,000,000 words) the following similarities and dissimilarities exist:
√ exact place
* exist
— not in the list

	BNC		Home Corpus
√	6187925 the	√	2215 the
√	2941786 of	√	1633 of
√	2682874 and	√	1335 and
*	2560344 to	*	1003 in
√	2150872 a	√	957 a
*	1883290 in	*	922 to
*	1115377 that	*	627 i
*	1089558 it	*	518 for
√	998857 is	√	459 is
—	923972 was	*	409 on
*	905318 i	—	362 my
*	851722 for	*	331 s
*	807305 's	*	303 at
*	724195 on	*	281 you
*	695595 you	*	277 this
—	681374 he	*	266 with
*	664778 be	—	231 university
*	652050 with	—	231 page
*	517783 as	*	220 are
*	513075 by	*	205 from
*	478177 at	*	204 by
*	473691 have	*	194 an
*	470949 are	*	192 it
*	463235 this	—	190 language
—	462776 not	*	185 as
—	456071 but	*	177 that
—	445396 had	*	177 have
—	433594 his	—	176 research
—	433475 they	—	165 m
*	413532 from	*	165 be
—	380284 she	—	161 here
—	372031 which	*	158 or
*	370855 or	—	149 home
—	358792 we	—	139 information
*	344045 an	—	130 me

There are 10 items in each list which do not appear in the other. For the BNC the items can be divided into three groups: *was* and *had*, which relate to tense and aspect; *he*, *his*, *they*, *she* and *we*, which relate to person; and *not*, *but* and *which* that are part of a stylistic convention. Examining the first group, an interesting phenomenon is revealed. While the BNC frequency list contains all the tense variations for *be* (except for am) and *have*, the Home corpus frequency list include only the present tense of these verbs. This might suggest that there is a strong preference to use the present tense in writing hypertext. This preference is probably due to the fact that hypertext is changeable and that people modify their files whenever the facts change, thus at the time of reading the hypertext document the present is assumed to be representing a fact. Since homepages are collections of facts about people, it is not surprising to find that the tense used for writing them is the present one.

The fact that there are no third person, animate, pronouns in the Home Corpus top 35 frequency list suggests that the "conversation" in hypertext is only between two participants: the author - first person singular, and the reader - second person sg/pl. Bringing together the findings from both groups adds up to a very simple environment: people write with a specific audience in mind and they use tense to emphasise and assert the accuracy of their descriptions and claims.

The group of *not*, *but* and *which* is also intriguing and much can be learnt from its exclusion from the list. The words *but* and *which* appear mostly as connecting two sentences and making them one, or expanding the context of the noun phrase. The word *not* is probably excluded from the list because of the abbreviations people choose to make (*don't*, *can't* etc.), but also because it contradicts the factual perspective that people want to create. From the exclusion of these three words we can infer that people tend to write shorter sentences, with a simple, straightforward style, trying to avoid contradictions and inconsistency by omitting negative meanings. The word *and*, however, is on the affirmative and therefore used as a normal connective between two simple sentences.

From the 10 words which appear in the Home Corpus top 35 frequency list and not in the BNC list we can learn more about the stylistic preferences of the authors of the hypertext documents. These ten words can also be divided into groups. The first group includes the words *my*, *m* (short for *am* in *I'm*; this verb is mentioned in this context but it also strengthens the aspect/tense group), and *me*. The existence of this group reinforces the claim made earlier that the language used in hypertext is more direct and informal. The second group of words which appear in the top 35 most frequent words in the Home Corpus but do not appear in the top 35 words in the BNC frequency list includes two words: *university* and *research*. As mentioned before, these words are strong indicators of the content of the Home Corpus: it is a collection of homepages of people who work in universities and are involved in research. The other group, the third, is an ambiguous one: The words *language* and *information* can be associated with both the content of the documents (research and universities), and with the group which will be discussed in the following paragraphs. This ambiguity might reduce the significance of the existence of these words in the top 35 most frequent words.

The last group of words is *page*, *here* and *home*. Looking at this group in a larger context could emphasise its significance. The larger context in this case is the frequency of pairs of words within the Home Corpus. These pairs of words are called bigrams and their top 21 most frequent occurrences in all the files of the Home Corpus are listed below (21 bigrams because the last one is relevant to this analysis):

of the
in the
university of
at the
on the
home page
i am
to the
department of
for the
and the
is a
from the
the university
natural language
if you
cognitive science
to be
i have
of my
click here

The word *home* appears usually as a pair of the word *page* (231 *page*, 149 *home*, 105 *home page*). The word *page* appears more than 50% of the time on its own. Does it mean that the other occurrences of the word *page* signify different meaning? Let us examine the following examples:

Mike Barlow's parallel corpus page (017.html, Home Corpus)

Georgetown University Natural Language Processing Parser Modularity Demo Page (023.html, Home Corpus)

Return to top of page (023.html, Home Corpus)

Help pages etc. (043.html, Home Corpus)

A service for checking that WWW pages conform to the standard (051.html, Home Corpus)

Lonely Planet's Rongo-rongo page (066.html, Home Corpus)

Adam's Easter Island Page (067.html, Home Corpus)

the author's page here (080.html, Home Corpus)

The word *page* substitutes the combination home page and is therefore a metonym for both *home page* and *web site* (which does not appear in the above bigram frequency list).

The word *here* appears to be also a part of a combination of words: *click here* (161 *here*, 65 *click*, 49 *click here*). Again, the word *click* occurs more with the word *here* than on its own, while the word *here* appears almost 70% of the time on its own. This combination is very interesting because it is unique to the environment of web navigation. A very interesting behaviour of this combination can be found looking at their cooccurrence in the bigram and monogram frequency list in the anchors of the Home Corpus: only 16 *click here* bigrams; but 82 *here* on its own! Knowing that *click here* has a well defined context, and that it is in the imperative, it can be concluded that the combination *click here* appears as partially highlighted in 67.4% of the cases. The word *here* also appears as a one-word-anchor 60 times, meaning 73% of the time. Combining the number of occurrences of *here* on its own (60) with the times it appears in the combination *click here* (16) exemplifies the iconic nature of the usage of *here*. The word here serves not only as an indicator of placement in space but also as an iconic word which is used as a button.

These orientation instructions given by authors are unique to hypertext in the sense that it is unlikely and unconventional to find an author of flat text telling its readers to turn the page or to physically relate to one word or another. This role that the authors are taking on themselves is new because they no longer count on the verbal cues and linguistic conventions to guide their readers, but they also add an explicit physical instruction to control their audience. In her article Charney (1994) describes this change in the role of the author as follows:

> *".. for readers to make appropriate connections between related ideas, the sentences expressing these ideas should appear in close proximity. Thus a text is easier to read if its points are developed in coherent sequences of sentences, paragraphs, and sections and if it contains discourse cues that signal the relations among these ideas.... Since readers use high-level ideas to tie portions of text together, these concepts should be explicitly stated early in the text and should be clearly signalled so that the reader can easily recall them as the need arises.... Thus it is easier to read, comprehend, and remember a text if it contains an informative title, headings, overviews, and topic sentences introducing key concepts that are repeated and developed in successive portions of text....These strategies, however, place the burden of selecting and arranging information, and providing signals to the arrangement, primarily on the writer. Hypertexts, by shifting a large portion of this burden to the reader, by proliferating the readers' choices about what portions of a text to read and in what order, compound the difficulties of creating a coherent mental representation."*

Whalley (1993) suggests to place reader-author relations on a continuum:

Reader Control					***Author Control***
little cohesive reference	reference manual	encyclopaedia	course reader	distance teaching text	much cohesive reference

He also says that *"the amorphous 'links' provided by hypertext do not provide any true cohesive reference. If everything is related to everything then essentially no cohesion is provided."*.

These two researchers put a finger on the reason for inserting more physical instruction in the body of hypertext. These verbal directions as to where the reader should go from a certain point in the text are part of the convention of writing hypertext documents, probably because people find that there is a need for more guidance in reading such a spatially complex text. All the phenomena studied in the above paragraphs seem to indicate that authors do their best to facilitate the navigation between hypertext documents. They use direct and informal language. They avoid complex linguistic structures. They use the present and factual tense. They introduce their environment by writing where they come from and what they are doing, and by indicating that the document, for the sake of orientation, is the here and the now. And above all they try to assist the readers by supplying them guidelines as to how they would like them to read their homepage.

The unique function of determiners in hypertext

When we think of a conventional discourse it is always defined with a beginning, a body, and an end. If it is a conversation, it begins with the first human sounds or gestures and it ends when the sound or gesture stop. But, as shown before, in hypertext there is never an end, and, in most cases, the participant begins the reading *in medias res*. The hypertext visible on the computer screen is part of a larger context of documents and sites.

When we take a book from the shelf and start reading it, we can always 'jump' to its end to see what has become of the hero, or what the answer to the riddle is, but after doing that we know that there is no other end to this book. Even if we keep on reading from where we were before, the end will remain the same, and in the identical place where it was before. In other words, the size and content of the text would remain the same. In hypertext the dimensions of the whole context are not defined since the links are part of the hypertext and the links they themselves include are also part of the original hypertext etc. We can now speak of a flat text as a unit which has textual integrity, and of hypertext as many sub-units having no textual integrity. Landow (1992) relates to this issue, specifying that:

> *"Another related effect of electronic linking [is that] it disperses "the" text into other texts. As an individual lexia loses its physical and intellectual separation from others when linked electronically to them, it finds itself dispersed into them. The necessary contextuality and intertextuality produced by situating individual reading units within a network of easily navigable pathways weaves texts, including those by different authors and those in non-verbal media, tightly*

together. One effect of this process is to weaken and perhaps destroy any sense of textual uniqueness" (p.53; see also pages 57-59).

This observation about the spatial nature of hypertext might mislead readers because it seems as if there is no order in hypertext and that only chaos control navigation. From the hypertext corpora gathered for this dissertation there is strong evidence contradicting this observation, at least from the linguistic point of view. Authors make use of the linguistic tools they know in order to help the reader navigate. As mentioned earlier they indicate both place and context whenever this is possible. Furthermore, they use definite and indefinite articles (*a* and *the*) to make explicit what the reader ought to assume as basic world knowledge and what they are not expected to know. Akmajian et al. (1995) describe the role of definite and indefinite articles as the indicating cues for presuppositions and given-new information. Since hypertext is accessed from many directions and many different contexts, the authors of hypertext documents introduce information in a very cautious way. They use determiners to characterise the information given, and to let the reader know the immediate context of this information.

From the first 55 files of the Home Corpus six contain once or not at all the word *the* (files 2 (once), 10 (once), 23, 36, 38, and 50), and ten contain once or not at all the word *a* (2, 10, 24 (once), 26, 36, 38, 42, 50, 51 (once), 53 (once)). This is a very interesting phenomenon because there are files which do not contain articles at all or once. The table below shows the number of determiners and verbs in these files:

file	**number of words in the file**	**verbs used**	**the**	**a**
002.html	64	0	1	0
010.html	100	0	1	0
036.html	25	last *modified…*	0	0
038.html	46	0	0	0
050.html	54	0	1	0

Although it is obvious that the number of words in these files is relatively small, it is not common to encounter texts without determiners and verbs. From this data it can be understood that some of the hypertext documents contain incomplete sentences and even incomplete noun phrases. The use of determiners is therefore limited to cases where they can add information to the environment apart from their linguistic function. This is not to say that determiners and verbs are not used in a similar manner they would normally be used in writing text, but that there is an additional linguistic role to their appearance or lack of appearance in hypertext. Let us examine the occurrence and usage of the article *the* in the anchors. The following list includes all the anchors in the Home Corpus which begin with the definite article *the*:

The Shepheardes Calender
The Beggar's Opera
The Humanist Web
The Brown University Home Page
The "home page" of my thesis
The "homeopathic fallacy" in learning from hypertext"

The (almost) Complete Guide to WWW in Israel
The Adaptive Hypertext and Hypermedia Homepage
The Argus Clearing House
The Back 40: Archaeology
The Book of the Courtier
The Center For Cognitive Science
The Centre for the Easily Amused (C*E*A)
The Chesapeake Bay Bolide:
The Chronicles of England
The Clickable Anthony
The Coconut Veranda
The Data Mine
The Day the Universe Went All Funny:
The Duke of Edinburgh's Award International Association

The Dukes of Hazzard. Yeeha!
The Electronic Neanderthal Woodworker
The Electrotechnical Laboratory
The English-Norwegian parallel corpus project
The Faerie Queene
The Fine Print
The Fortune 500 firms, 1996
The Fountain of Moravec
The Fowre Hymnes
The Free On-line Dictionary of Computing
The Gaelic college
The Garden of the World Project
The Geological Society of America
The Gernsback Continuum
The Goodies
The Guardian newspaper on-line
The HCI Research school in Stockholm
The HPSG Workshop
The Hazardous Materials Sheet for Women
The Integration of AM/FM and Work Management
The Interdisciplinary Weekly Tea Seminar
The Irish Chess Archive
The JRR Tolkien Information Page- Info about my favorite author.

The Java Repository
The Java programming Language
The Koine Greek Verb: Tense and Aspect.
The Korin Richmond Repository.
The LINGUIST Network
The LTG crew
The Lady of May
The Language Software Helpdesk
The Language Software Helpdesk
The Language Technology Group
The Legal Stuff (how you can/cannot use this)
The Living Room: HOT LINKS
The MainStay BBS ADDRESS BOOK
The Mammoth Saga
The Migraine Project
The Modern English Collection
The Music Room: Claire & Her Music.
The NLP Software Registry Homepage
The Natural Language Processing Group in the Department of AI.

The Net, BBC TV
The New Age
The Official Homepage of Toad The Wet Sprocket- one of the best bands ever!

The Pearl Dive
The Pixel Forge: Hand Hammered Special Effects (slow load; lotta pictures)

The Poetry Corner
The Press Room: Media lies and distortions. NEW!!!
The Program, ITV
The SFEP-ED-L homepage
The Sacred Chao
The Sacred Chao
The Scotland index
The Semantics and Pragmatics of Lexical Aspect Features.

The Simpsons;
The Sinclair Archive
The Sinclair ZX Spectrum Switchboard
The Skylight: Life On Mars?
The Spam Filter
The Sunsite Gaelic homepage
The Syntax and Semantics of Predication
The Tony Godwin Memorial Trust homepage
The Tree of Life Home Page
The University of Berkeley Museum of Paleontology
The University of Cambridge
The University of Edinburgh
The Unofficial Haitian Home Page
The Vicarious Learner Project
The Virtual Earth
The Voyager Cd-Rom
The WWW Virtual Library
The WWW Virtual Library
The WWW yellow pages of Israel
The Windows 95 FAQs
The Zero Point Knowledge Unit
The comp.fonts Home Page
The history of king Richard the Thirde
The last of the greats - Alice and Peter's 50th Birthday Party

The lunar calendar of Tablet Mamari", Journal de la Societe des Oceanistes, Paris, 1990
The normal home page - only for local users
the
the
the
the Centre for Cognitive Science
the DEFACTO project
the DRAFTER project
the Department of AI
the GIST project
the Human Communication Research Centre
the NLP group
the University of Edinburgh
the author's page here
the list of publications
the release announcement
the webmaster

This list consists mostly of stand-alone noun phrases which are preceded by the definite article. If we examine it thoroughly a very interesting pattern appears. The definite article is capitalised in most anchors (when it is in initial position) and it refers to an external existing object. The reason for saying that these objects are external is that anchors, by nature, are doors to other places. Their name is sometimes an indication to what can be found behind them. The interesting fact is that if before in this chapter a somewhat simple and iconic language appeared to be used in order to facilitate navigation, here we see what seems to be the opposite: a reference to an object which was never mentioned before as if its existence is a well known fact. This seems to be contradictory: Why would authors bother with explaining the context of their hypertext documents in such an explicit manner and then assume knowledge that is obviously missing in many cases?

These anchors appear to be token-reflexive (Reichenbach (1947); Burks, (1949); Sellars, (1956)). Token-reflexive expressions are expressions which serve to connect the circumstances in which a statement is made with its sense. Such a definition would help in explaining why people use definite article when they refer to external and previously unseen or unmentioned objects. The additional information provided by the use of the definite article is a valuable one. It means that in the local context there is only one object by that name and that this is a fact known to the author. It also coins a new term or proper name to be used in this given environment, in order to use it as a verbal button. Sellars (1956) claims, in the context of basic world knowledge and conversation, that such token-reflexive expressions add both authority and credibility to the text. The text becomes more authoritative because the author’s assertions are definitive and therefore assumed to be true. Thus the very same text becomes more credible because it is, within context, factually true. Inserting new knowledge, presented as if it was known to both the author and the reader, contributes to the factual environment people want to create in their web pages.

To strengthen this credibility people tend to use the indefinite article to stress that the links given are only a sample from might be a large collection of occurrences of similar documents or sites. This usage of the indefinite article adds to the accuracy of facts given by authors because assuming that some of the objects might have more than one related site on the web seems reasonably true in the list given below. The following examples are all the anchors in the Home

Corpus which begin with the indefinite article *a*. Six of the anchors are actually names of articles which are cited by the authors (marked with *):

*A Cognitively Relevant Lexical Semantics
*A Defence of Poesie
*A GRASP for job-shop scheduling, presented at INFORMS San Diego Spring 1997 Meeting, San Diego, May 1997.
A Gardener's Page
A Japanese Easter Island site
A Sentimental Journey through France and Italy
A View on the Present Condition of Ireland
A clickable map of Wales
A clue
*A comparison of lexicon-building methods for subword-based speech recognisers
A cool trivia page
A corpus for teaching Portuguese
*A joint segmentation and labelling scheme for use in acoustic subword based speech recognition
A map of Israel
A map of Israel & its neighbouring states
A page dedicated to my second favorite author: David Eddings
A picture of Bangor
A service for checking that WWW pages conform to the standard
*A window on lexical density in speech
An NC definition
An internet search

When added to the previous findings from the Home Corpus, namely the use of present tense and the simplified syntax, this new observation seems to explain the contradiction between the simple and explicit language used and the unexplained new introduced facts: The explicit language explains to the reader the context of the hypertext document and by the time the new terms are introduced the reader understands them in the local world which is limited by physical devices such as screen and document length. The local context created enables the author to introduce new information as if it was factual in the context of the hypertext document. This new information is not only accepted as true but adds to the credibility of the document.

The linguistic devices used in hypertext can thus be explained in the following way: The authors introduce themselves and their environment to the reader, describing the context of their document with the most explicit information, giving name, place and sometime even date. This well described context allows them to refer to their document as being *here* and other hypertext documents as being *there* or elsewhere. When the author wants to introduce new information, or refer to an outside document they simply assert its existence by naming it and thus defining and inserting new facts into the document. The only knowledge needed in order navigate between documents, then, is the understanding that each hypertext document has a local environment and if the reader wants to know more about one of the facts introduced by the author, all they have to do is click this new fact and jump to its local environment. Of course such a jump might show the fact and its explained context but it would probably introduce more factual objects in the new local context etc. Iconic words such as *here*, *this, back to* and *home page* are therefore, within the hypertext writing convention, cues and road signs as to where a starting point can be found. These words appear to be used in similar contexts and syntactic structure, facilitating the orientation within and between documents.

This illustration of the hypertext document local context can also support and explain the findings, retrieved from the Home Corpus, that there is no proportion between the length of documents and the number of links they contain: Too many unfamiliar objects in one local context can create distraction and incoherent text (Charney , 1994). Since the local environment of hypertext documents appears to be the whole HTML file, then the number of new inserted linked-objects is limited within their physical surroundings. It seems that the structure of paragraphs has no affect on the number of inserted links and that the latter is restricted to fit coherence limitations.

Lists

Lists of clickable items are very common in hypertext and HTML includes many different list formats in markup commands. They are usually bulleted or numbered and can be embedded or nested one within the other. Reading such lists, especially when they are partially highlighted, might affect reading speed and comprehension. Johnson-Eilola (1994) claims that hypertext documents are *not read top to bottom across a page* but according to navigation patterns. This seems to agree with the nature of lists because these can be read in a non-linear manner if desired. Information arranged in lists might be more accessible on the screen because of its indexical nature. There are lists which contain only highlighted anchors and others which include both highlighted and non-highlighted items (highlighted items are underlined):

Map of Easter Island
"Mysterious Places"-Easter Island
Adam's Easter Island Page
Noe's Easter Island photos
Dr. Ruggles' Easter Island Images
Photos by Massimo Casulini
Outer Realms
Cordell Easter Island Expedition
Rapa Nui Outrigger Club
Rapa Nui En Español
A Japanese Easter Island site
South Seas Symposium
Chile Information Project
Local Time on Easter Island
Rapa Nui News Updates

Topographical Map
Easter Island Woman
Art of Lorenzo Domínguez
Rapa Nui at the "Wilds Scene"
Rapa Nui Puzzle
Get Your Moai Here!
"Weird Stuff on Rapa Nui"
Easter Island Virtual Reality
(067.html)

Ashtanga Vinyasa Yoga
Ayuravedic Massage
UNIX software development - see CV
Photography
(098.html)

EDINFO - The University of Edinburgh's Information Service
Cognitive Science Home Page
Cognitive Science Bone Page
Cognitive Science Bonus Page
Jobsites
Computer Documentation
Travel Sources
Sources
Bookmarks
(133.html)

It is very difficult to define exactly what makes people highlight one phrase in a list and not the other. Sometimes it is because of simple lack of information and sometimes it is motivated by document layout constraints. However reading such highlighted lists can affect both reading speed and comprehension. Nygren et al. (1995a) found that both search time and visual processing are significantly and substantially shorter for highlighted items in texts. They suggest that "*highlighting the critical items causes sequential search to be replaced by pattern recognition*". This observation also supports the claim that too many highlighted items in one hypertext document could interfere with the anchors' highlighting effect. In the list in file number 133.html, shown above, there is an interesting phenomenon: the author is highlighting the same information three times with only a slight change in each item. This over-highlighting seems to be consciously done in order to create a rhythmic effect. At first glance it seems that all the items are the same and attention is drawn to these items. It appears to be that this list was designed to create this "more attention" effect and that the author is using their knowledge about how people read hypertext in order to manipulate their reading.

Although it seems that there is no real order and method in writing hypertext documents, the discussion and findings in this chapter suggest that there is pragmatic reason behind hypertext writing and that this is manipulated and controlled by a simple usage of linguistic devices. Generating or converting flat text to such hypertext documents automatically can be a very difficult task then. However, if linguistic devices will be used in as similar as possible manner, as they are used in manually authored hypertext, then at least the hypertext environment would be familiar to the user and they could apply the basic knowledge to manually and automatic generated hypertext alike.

Chapter Six

Conclusions

Dillon et al. (1993) observed, when the hypertext authoring on the web was just beginning to become popular in the non-academic world, that there is a problem of schemata, or genre conception, in hypertext. They stated that "*since schemata are generic abstractions representing typicality in entities or events, the increased variance of hypertext implies that any similarities that are perceived must be at a higher level or must be more numerous than the schemata that exist for paper texts…. The current lack of standards in the electronic domain compared to the rather traditional structure of many paper documents is a farther problem for schemata development*". Today, almost five years later, the web is used by many people and the standards mentioned in Dillon et al. are not yet official as suggested, but, instead, there are conventions which evolved from usage and experience. In the years that passed since then, users became aware to the existence of other users by interacting with their hypertext documents and by creating their own homepages. This study describes the conventions with which hypertext documents are being written. It is suggested in this dissertation that studying these conventions and applying this knowledge to existing academic work, would be beneficial to both hypertext users and the research community.

The Idea

This part of the conclusions, the motivation for writing this dissertation, will be summarising questions which were raised before as well as introducing new problems The idea behind this dissertation is that hypertext is a new genre of expression and that it is systematically different from other communicative verbal means of expression like flat hierarchical text or speech. The underlying assumption of this dissertation is that there are many characteristics of hypertext which are ignored by the academic research community simply because hypertext is assumed to be very similar to flat text. This ignorance or avoidance can create many mismatches between what is studied today and what should be studied in order to advance the research done on hypertext in the near future.

One of the tasks of NLP community (Natural Language Processing), for example, is to be able to identify different parts of speech (POS) within a given string of language. This is usually done by learning the statistical regularities of a POS in many strings in which POS are already tagged and then applying the learned knowledge to a new encountered string, assuming that the regularities are similar to the ones learned before. Most of the POS taggers today are trained on very large corpora. These corpora are tagged for POS and comprise a very large amount of words taken from written flat text or a transcription of speech. It is claimed here that since hypertext is not represented in these corpora they are not effective as training corpora for tagging hypertext. Hypertext has different grammatical conventions which are distinct from the grammatical conventions used in speech or flat text writing. These differences in grammar conventions, if ignored, can affect the development of natural language understanding techniques applied to hypertext. We have tried one tagging technique, the XeroxTag one, on the larger corpus collected (the Parker Corpus) and since there was no sufficiently large body of text tagged for POS, we can only give our human impression that the tagging did not perform as well as it is reported to perform on flat text. It is true that many of the documents on the web are converted from flat text but their arrangement is different from the one exhibited by flat texts and therefore lacks the continuity effect achieved by flat text writing. This broken continuity alone can cause differences in the automatic tagging of hypertext because it isolates chunks of text from their immediate context which might indicate a cue or regularity pattern learned from flat text (i.e. frequency of adjacent nouns and adjectives etc.).

Since the usage of lists, for example, is very frequent in hypertext documents, the regularities of POS change in a list-like environment, especially (as it is often the case) if the list consists of anchors. Amitay and Oberlander (1997) found, in the context of anchors in isolation, that determiners are omitted from the anchor when they are not part of the proper name or a proper-noun-like entity, or when they might distract the reader from the core idea . For example:

Go to the <A HREF="1">Sample Chapter</A> (file 537, Parker Corpus)

For new material- check out the Library Fall Programs (file 127, Parker Corpus)

Next: The Titans ride the <A HREF="3">Highway to Hell</A> (file 230, Parker Corpus)

Advertising on these pages is subject to the <A HREF="4">FederalFair Housing Act.</A> (file 624, Parker Corpus)

Verb usage is also intriguing and there is a significant tendency to use imperatives and present participle morphology such as '*climbing*', '*jogging*', '*reading*', '*downloading*', etc. These forms appear very often in vertical lists of links, apparently to shorten the message and to make it as compact as possible. For example:

<A HREF="1">Sailing on the Internet</A> (file 613, Parker Corpus)

Historical Exploring (file 104, Parker Corpus)

<A HREF="3">Backpacking</A> (file 724, Parker Corpus)

This kind of linguistic difference between what is expected of a text and what hypertext is, should be examined more thoroughly, if only because of the POS tagging. This linguistic understanding is also important in producing hypertext automatically (hypertext generation). The Natural Language Generation community (NLG) is currently applying the knowledge it has about flat text to generating hypertext. This, again, presents a problem because of the linguistic and structural differences between text and hypertext. While generating flat text requires one type of syntactic and structural constructions, hypertext generation might require somewhat simpler linguistic constructions. The linguistic emphasis should probably shift from sentence generation in flat text to anchors-in-context generation in hypertext. The shift might also be necessary because of the change in environment and layout that is distinctive in flat text and in hypertext.

Studying the layout of hypertext and the context in which people choose to put their anchored-links can also be used as an aid to the Human Computer Interaction (HCI) community to better understand the needs of the users and their choice of linking between documents. This understanding can be applied to interface design techniques and online help design for commercial software (which are actually a close-system hypertext and only slightly different from the open-system hypertext - or as known by its name - the web).

The cognitive side of reading is also different between text and hypertext. While the layout of text and its influence on reading is an interesting question, in hypertext there are many more aspects and additions which do not exist in text. The first one is of course the fact that hypertext can only be viewed on a computer screen which brings to the fore the problem of screen layout and the effect of scrolling. Hypertext also has a spatial dimension which is marked by highlighted words in it. The highlighted words can both interfere with and enhance local understanding of context, and the ability to skim through the written words and sentences. These questions and problems should be considered when addressing hypertext issues in research. Some of the more simple questions were discussed in this dissertation but others require more investigation and thinking.

Concluding Remarks and Comments

Creating hypertext manually seems to be a task that requires a lot of thought and linguistic awareness. Nevertheless there are many people who write personal homepages and maintain topical web sites. It was found in this study that although homepages appear to contain different subjects, and although information is arranged in varied ways, there is a general pragmatic-linguistic consistency between hypertext documents. This consistency can be used to understand the way people navigate in hyperspace, shifting the emphasis from the referential space i.e. the connecting links between documents, to the documents themselves. This kind of shift might aid in improving hypertext documents' coherence and reducing the "lost in hyperspace" effect (Conklin, 1987).

Earlier in this document it was suggested that in order to facilitate access to and manipulation of HTML files for research purposes, the academic community should treat HTML as a "legitimate" child of SGML. The current academic attitude assumes that HTML is an unruly mutation of SGML and therefore it can not be treated as an equal and "serious" markup language. This new attitude towards HTML, if taken, would enable researchers to study the web conveniently and with less constraining computer environment such as the SGML tools for UNIX.

In the introduction to this dissertation several questions were presented. Some of these questions were answered in this study and some of them require further research. By restricting the **domain** of the Home Corpus we were able to extract information about the language used specifically in homepages, but also about the general linguistic context in which hypertext documents are displayed, and the way this context is created by the users of the web. We also observed that the single, isolated hypertext **document** presents a coherent environment which defines external and internal objects. It was found that anchors are not **paragraph** dependent and that their distribution conform with the pragmatic conditions laid out by authors. **Lists** and **anchor wording** were studied by example. It appears that there is an additional depth to the usage of determiners and that **sentence** structure, in general, is broken. Omission of verbs and prepositions is allowed as long as the message is conveyed. This study shows that there are discrepancies between what hypertext is assumed to be for research purposes and what it really is. It is suggested that hypertext is a new genre with its own, already defined, linguistic patterns and conventions, and further research to strengthen this assertion is encouraged.

References

Akmajian A., Demers R. A., Farmer A.K., and Harnish R.M. (1995). Linguistics: An introduction to language and communication. Forth Edition. MIT Press.

Amitay E., and Oberlander J.(1997).Convention Says... In proceedings of the Flexible Hypertext Workshop held in conjunction with The Eighth ACM International Hypertext Conference (Hypertext'97). also online: http://www.cogsci.ed.ac.uk/~einat/FHTW97.html

Anderson T., Berre A., Mallison M., Porter H., and Schneider B. (1990). The hypermodal benchmark. in F. Bancilhon, C. Thanos, and D. Tsichritzis (eds.), Advances in database technology-EDBT '90. London: Springer-Verlag.

Baron L., Tague-Sutcliffe J., Kinnucan M. T., and Carey T. (1996). Labelled, Typed Links as Cues when Reading Hypertext Documents. Journal of the American Society for Information Science. 47(12):896-908.

Bernstein M. (1990). An apprentice that discovers hypertext links. In Proceedings of the European Conference on Hypertext. Cambridge University Press. UK.

Blustein W., J. (1994). An Evaluation of Tools for Converting Text to Hypertext. MSc Thesis, Department of Computer Science, The University of Westren Ontario, Canada.

Burks A. (1949). Icon, Index and Symbol. Philosophical and Phenomenological Research, vol. IX, no. 4, June 1949: 673-689.

Burnard L. (1995). What is SGML and How Does It Help?. Computers and the Humanities, 29:41-50. and also: http://sable.ox.ac.uk/ota/teiedw25/

Bush, V. (1945). As We May Think. The Atlantic Monthly. Also online: http://www.isg.sfu.ca/~duchier/misc/vbush/

Carlson P.A. 1989. Hypertext and Intelligent Interfaces for Text Retrieval. in The Society of Text. ed.: E. Barrett. Massachusetts: MIT Press.

Carter E. (1996). Quantitative Analysis of Hypertext Generation and Organisation Techniques. PhD Thesis, Department of Artificial Intelligence, The University of Edinburgh, Scotland.

Charney D. (1994). The Effect of Hypertext on Processes of Reading and Writing in Literacy and Computers: The Complications of Teaching and Learning with Technology. ed.: C. L. Selfe & S. Hilligoss, New York: The Modern Language Association of America.

Chen C., and Rada R. (1996). Interacting With Hypertext: A Meta-Analysis of Experimental Studies. Human-Computer Interaction. 11(2):125-156.

Church K., and Gale W.(1995). Poisson mixtures. Journal of Natural Language Engineering, 1:2:163-190.

Church K., and Hanks P. (1989). Word association norms, mutual information and lexicography. In ACL Proceedings, 27th Annual Meeting, Vancouver.

Conklin J. (1987). Hypertext: An Introduction and Survey. IEEE Computer, 20:9:17-41.

Dale R., Milosavljevic M. and Oberlander J. (1997) The Web as Dialogue: The Role of Natural Language Generation in Hypertext. In Proceedings of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web.

Dillon A., McKnight C. and Richardson J. (1993). Space - The Final Chapter or Why Physical Representations are not Semantic Intentions. In Hypertext: a psychological perspective. Dillon A., McKnight C. and Richardson J. eds. Ellis Horwood Ltd. UK.

Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19:1:61-74.

Golovchinsky G. (1997). What the Query Told the Link: The Integration of Hypertext and Information Retrieval. In Proceedings of Hypertext 97, Southampton, UK.

Graham I. (1996). Web/HTML Documentation and Developer's Resource. http://www.utoronto.ca/webdocs/webinfo.html

Hearst M., A. (1994). Multi-paragraph segmentation of expository text. In ACL '94. Las Cruces, NM, USA.

Johnson-Eilola J. (1994). Overview: Reading and Writing in Hypertext: Vertigo and Euphoria. In Literacy and Computers: The Complications of Teaching and Learning with Technology. C.L. Selfe and S. Hilligoss eds. The Modern Language Association of America, NY.

Katz S. M. (1996). Distribution of context words and phrases in text and language modelling. Natural Language Engineering, 2:1:15-59.

Kilgarriff A. (in press). Using word frequency lists to measure corpus homogeneity and similarity between corpora. http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/#papers

Kilgarriff A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. In Language Engineering for Document Analysis and Recognition. Proceedings, AISB Workshop, Falmer, Sussex.

Kilgarriff A., and Salkie R. (1996). Corpus similarity and homogeneity via word frequency. Euralex '96, Gothenberg, Sweden.

King K. S., Boling E., Annelli J., Bray M., Cardenas D., and Frick T. (1996). Relative Perceptibility of HyperCard Buttons Using Pictorial Symbols and Text Labels. Journal of Educational Computing Research. 14(1):67-81.

Landow G.,P. (1992). Hypertext. The Johns Hopkins University Press, Baltimore, USA.

Leibowitz-Taylor S., Dahl D.A., Lipshutz M., Weir C., Norton L.M., Weidner-Nilson R., and Linebarger M.C. (1994). Integrating Natural Language Understanding with Document Structure Analysis. Artificial Intelligence Review, 8:255-276.

Maioli C., and Vitali F. (1993). Anchors and Paths in Hypertext Publishing System. Technical Report UBLCS-93-3.

McDonald S., Stevenson R. J.(1996). Disorientation in hypertext: the effects of three text structures on navigation performance. Applied Ergonomics. 27(1):61-68.

McDonald S., Stevenson R. J.(in press). Hypertext, navigation and cognitive maps: the effects of a map and a contents list on navigation performance as a function of prior knowledge. in D. Harris (ed.) Engineering Psychology and Cognitive Ergonomics: Interaction of theory and application (in press). Avebury Technical.

Milosavljevic M., Tulloch A., and Dale R. (1996) Text Generation in a Dynamic Hypertext Environment. Pages 417--426 in Proceedings of the Nineteenth Australasian Computer Science Conference (ACSC'96), Melbourne, Australia. 31 January – 2 February 1996.

Moore, J. (1995) Participating in Explanatory Dialogues. MIT Press, Cambridge, MA.

Myka A., Argenton H., and Güntzer U.(1996). Towards Automatic Hypertextual Representation of Linear Texts. Technical Report No. WSI-96-12, ISSN 0946-3852, Arbeitsbereich Datenbanken und Informationssyteme, Wilhelm-Schickard-Institut, Universität Tübingen, Germany.

Myka A., and Güntzer U. (1994). Automatic Hypertext Conversion of Paper Document Collections. Lecture Notes in Computer Science, No. 916:65-90.

Nanard J., and Nanard M. (1995). Adding macroscopic semantics to anchors in knowledge-based hypertext. International Journal of Human-Computer Studies. 43:363-382.

Nelson, T. Dream Machines. Chicago Il: Hugo's Book Service 1974 (Nachdruck Microsoft Press 1988)

Norman L. K. (1991). The Psychology of Menu Selection: Designing Cognitive Control at the Human/Computer Interface. Ablex Publishing Corporation, New Jersey.

Nygren E. (1996). From Paper to Computer Screen - Human Information-Processing and User Interface Design. CMD, Uppsala University. http://delfi.cmd.uu.se/papers/diss188/summary.html

Nygren E., Allard A., Lind M. (1995a). Effects of patterns of highlighted items on list search. Report no. 55, CMD, Uppsala University. http://delfi.cmd.uu.se/papers/55/

Nygren E., Allard A., Lind M. (1995b). Effects of figural patterns on trend assessment. Report no. 56, CMD, Uppsala University. http://delfi.cmd.uu.se/papers/56/

Recker M., and Ram A.(1994). Cognitive Media Types as Indices for Hypermedia Learning Environments. Proceedings of the AAAI-94 Workshop on Indexing and Reuse in Multimedia Systems.

Reichenbach H. (1947). Token-reflexive words. In Elements of Symbolic Logic. New York: The Free Press

Reiter E., Mellish C. and Levine J. (1992) Automatic Generation of On-Line Documentation in the IDAS Project. In Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy, pp. 64-71.

Richmond K., Smith A. and Amitay E.(1997). Detecting Subject Boundaries Within Text: A Language Independent Statistical Approach. In proceedings of The Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2). Brown University, Providence, Rhode Island, USA

Salminen A., Tague-Sutcliffe J., and McClellan Ch. (1995). From Text to Hypertext by Indexing. ACM Transactions on Information Systems, 3:1:69-99.

Salton G., Allan J., and Buckley C. (1994). Automatic structuring and retrieval of large text files. Communications of the ACM, 37:2:97-108.

Schiffrin.D. (1994). Approaches to Discourse. Oxford: Blackwell.

Sellars W. (1956). "Empiricism and the Philosophy of Mind," in Herbert Feigl and Michael Scriven, eds., Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis (University of Minnesota Press, 1956), pp.253-329. http://csmaclab-www.uchicago.edu/philosophyProject/sellars/epm.html

Shneiderman B. 1989. "Reflections on Authoring, Editing, and Managing Hypertext" in The Society of Text. ed.: E. Barrett. Massachusetts: MIT Press.

Shneiderman B. (1987). Designing the user interface: Strategies for effective human computer interaction. Reading, MA: Addison-Wesley.

Sperberg-McQueen C. M., and Burnard L. - eds. (1993). Guidelines for Electronic Text Encoding and Interchange (TEI P3). Ch. 2: "A Gentle Introduction to SGML". http://www-tei.uic.edu/orgs/tei/sgml/teip3sg/index.html

Stanton N., A. (1994). Explorations into Hypertext: Spatial Metaphor Considered Harmful. Education & Training Technology International. 31(4):276-294.

Suchman L., A. (1987). Plans and Situated Actions: The problem of human machine communication. Cambridge University Press: GB.

Trigg R. H. (1983). A network-based approach to text handling for the online scientific community. PhD Thesis, University of Maryland. Also technical report TR-1346.

Van Herwijnen E. (1994). Practical SGML: Second Edition. Kluwer Academic Publishers. US.

Wenger M., J., and Payne D., G.(1996). Comprehension and retention of nonlinear text: Considerations of working memory and material-appropriate processing. American Journal of Psychology. 109(1):93-130.

Whalley P. (1993). "An Alternative Rhetoric for Hypertext" in Hypertext: a psychological perspective. ed.: Dillon A., McKnight C. & Richardson J. UK: Ellis Horwood

Wright P. (1993). To Jump or Not to Jump: Strategy Selection While Reading Electronic Texts. in C. McKnight, A. Dillon, and J. Richardson (ed.), Hypertext: A Psychological Perspective. (1993). Ellis Horwood LTD. Chichester.

Glossary

Although this dissertation is not a guidebook to any of the technological gadgetry of hypertext we feel that there is a need to explain some of the terms used. This glossary comes only as first aid to people who have not heard one term or another used before. This information is not given in the body of the text because we assume that most of the readers would have some previous knowledge.

In this glossary we will try to give a simple explanation to terms and concepts referred to in the dissertation. The terms are listed alphabetically. Some of the explanation used here are adapted from online texts and guidebooks. Following each entry is a list of online addresses and hardcopy books in which more information can be found.

Browser

A browser is a program that provides a way to look at, read, and even hear all the information on the World Wide Web. The word "browser" seems to have originated prior to the Web as a generic term for user interfaces that let you browse text files online. By the time the first Web browser with a graphical user interface was invented (it was called Mosaic), the term seemed to apply to Web content, too.

Technically, a Web browser is a client program that uses the Hypertext Transport Control Protocol (HTTP) to make requests of Web servers throughout the Internet on behalf of the browser user. Currently, the most popular browser is Netscape Navigator. Microsoft's Internet Explorer is gaining usage as Windows 95 installations grow. A commercial version of the original browser, Mosaic, is still widely used. Other browsers include the browsers for the online services, America Online, CompuServe, and Prodigy, but these are beginning to offer Netscape or Internet Explorer in addition to or as a replacement for their own. Lynx is a text-only browser for UNIX shell and VMS users.

(http://whatis.com/browser.htm)

DTD (Document Type Definition)

A Document Type Definition (DTD) is a formal grammar, designed to be used in electronic marked-up documents. It defines what are the legal productions of a given markup language. The designer of a DTD generally has to trade off generality of use with accuracy of error detection. The simplest kind of DTD would be one which did no more than specify a set of tag names, requiring only that every element tagged in a document use one of them. A strict DTD would specify both the tag names and their legal syntactic order.

A DTD is a specific definition that follows the rules a given markup language. A DTD is a specification that accompanies a document and identifies what markup tags separate paragraphs, identify topic headings, and so forth and how each is to be processed. By mailing a DTD with a document, any location that has a DTD "reader" (or "markup compiler") will be able to process the document and display or print it as intended. This means that a single standard SGML compiler can serve many different kinds of documents that use a range of different markup codes and related meanings (assuming these are compatible with the SGML requirements). The compiler looks at the DTD and then prints or displays the document accordingly.

IBM and many large and small corporations are converting documents to SGML, each with its own company Document Type Definition or set of definitions. For corporate intranets and extranets, the Document Type Definition of HTML provides one new "language" that everyone can format documents in and read universally.

(http://whatis.com/dtd.htm)
(http://sable.ox.ac.uk/ota/teiedw25/W25F.htm)

HTML (Hypertext Markup Language)

HTML is the set of markup symbols or codes inserted in a file intended for display on a Web browser. The markup tells the Web browser how to display the information for the user. HTML is defined both by a standards committee and by proprietary extenders of the language such as Netscape and Microsoft. Currently, the latest official version of HTML is HTML 3.2, but the HTML 4.0 is being introduced while we write these lines.

The main difference between HTML and SGML is that HTML is mostly a procedural markup, which means that the browser is told by the user how to display the document, rather than what the document contains both semantically and syntactically, as done in SGML. Another difference is the actual users. While HTML is used by the web community as a mean to an end, i.e. presenting documents' layout on the web, SGML is a part of an NLP holistic approach which helps computers understand natural language by encoding an agreed upon grammar and describing the attached text. This attitude is taken by most NLP researchers and is also useful where indexing techniques are used (such as publishing companies and online libraries).

NCSA's HTML Primer (http://www.ncsa.uiuc.edu/General/Internet/WWW/HTMLPrimer.html)

The Barebones Guide to HTML (http://www.werbach.com/barebones/)
Ian Graham's Introduction to HTML
(http://www.utoronto.ca/webdocs/HTMLdocs/NewHTML/htmlindex.html)

The World Wide Consortium's Web site includes an authoritative versions of the HTML 3.2 and HTML4.0 Reference Specifications. The W3C also describes the next HTML enhancements that are being developed. (http://www.w3.org/)

(http://whatis.com/html.htm)

markup

Markup refers to the sequence of characters or other symbols that are inserted at certain places in a text file to indicate how the file should look when it is printed or displayed. The markup indicators are often called "tags." For example, a paragraph is preceded by a:

<p> (or paragraph tag)

so that it will be separated by an empty line from the preceding line.

Markup can be inserted by the document's author manually by typing the symbols in, by using an editor and selecting prepackaged markup symbols (to save keystrokes), or by using a more sophisticated editor that lets the author create the document as they want it to appear (this is called a WYSIWYG editor).

There is now a standard markup definition (or really a description of how you can define markup) in the Standard Generalised Markup Language (SGML).

(http://whatis.com/markup.htm)
Burnard, (1995): (http://sable.ox.ac.uk/ota/teiedw25/W25C.htm)

SGML (Standard Generalised Markup Language)

SGML is a standard for how to specify a document markup language or tag set. The specification itself is called a Document Type Definition (DTD). SGML is not in itself a document language, but a description of how to specify one.

SGML is based on the idea that documents have structural and other semantic elements that can be described without reference to how such elements should be displayed. The actual display of such a document may vary, depending on the output medium and style preferences. SGML is based somewhat on earlier generalised markup languages developed at IBM, including General Markup Language (GML) and ISIL. Some advantages of documents based on SGML are:

- They can be created by thinking in terms of document structure rather than appearance characteristics (which may change over time).

- They will be more portable because an SGML compiler can interpret any document by reference to its DTD

- Documents originally intended for the print medium can easily be re-adapted for other media, such as the computer display screen.

The main difference between HTML and SGML is that HTML is mostly a procedural markup, which means that the browser is told by the user how to display the document, rather than what the document contains both semantically and syntactically, as done in SGML. Another difference is the actual users. While HTML is used by the web community as a mean to an end, i.e. presenting documents' layout on the web, SGML is a part of an NLP holistic approach which helps computers understand natural language by encoding an agreed upon grammar and describing the attached text. This attitude is taken by most NLP researchers and is also useful where indexing techniques are used (such as publishing companies and online libraries).

(http://whatis.com/sgml.htm)
(http://www.sil.org/sgml/sgml.html)
(http://sable.ox.ac.uk/ota/teip3sg/)
Van Herwijnen (1994)

XML (eXtensible Markup Language)

XML is a proposed markup language for use on the world wide web. It is unlike HTML in the fact that it is a subset of SGML, but it retains the power of SGML's structural coherence. XML drops all the more esoteric parts of the SGML definition (e.g. CONCUR, LINK, RANK...) that are very rarely used in SGML applications. XML could be viewed as "diet SGML" (the SGML definition document is almost 500 pages long the XML definition is 26 pages).

The parts of SGML that have been dropped are those that make SGML documents hard to process automatically. XML introduces the new concept of "well formed" documents, these are documents that be parsed without reference to a DTD, e.g. documents can be constructed according to a DTD, however when they may be delivered to (and processed by) systems that do not have access to the DTD. This will greatly ease the construction of browsers able to process XML and thus allow its delivery over the web.

Apart from its use on the web, XML is a good standard to use in normal SGML applications, it is increasingly rare to author "raw" SGML documents so many of the SGML features included to support this are redundant. Unless there are good reasons to use the elements of SGML not included in XML, we would recommend any new SGML project to adhere to the XML definition. XML is new idea and its development will be watched with interest, more information is available from the w3 organisation

adapted from: The SGML FAQ - http://www.jorvic.demon.co.uk/sgmlfaq.htm
http://www.w3.org/TR/WD-xml-961114.html