Anchors in Context:
A corpus analysis of web pages authoring conventions

Einat Amitay

einat@mri.mq.edu.au

Language Technology Group
Microsoft Research Institute
School of MPCE
Macquarie University
NSW 2109
Australia

CSIRO
Mathematical and Information Sciences
Locked Bag 17
North Ryde, NSW 2113
Australia

Abstract

Through interaction with the web users became aware of the existence of other users and their hypertext documents. By creating their own homepages this awareness was translated into practical generalities and the formation of patterns in web authoring behaviour. This study describes the conventions with which hypertext documents for the web are being written. It is suggested in this paper that studying these conventions and applying this knowledge to existing academic work would be beneficial to both hypertext users and the research community.

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. 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. This study aims at identifying the emerging linguistic characteristics of hypertext as they are practiced on the web.

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; 1997), 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. 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 our 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 on the web by simply retrieving and extracting information from it. Two corpora of HTML files were gathered and analysed.

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 paper is 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. 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.

Collecting the Corpora

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 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. 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, we believed the results from the URL search to be clean enough to be reliable.

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. This way we could avoid “cut and paste” versions of documents which were copied from a paper document to the web. 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.

Comparing the two corpora

The issue of similarity between corpora is a very interesting one. Kilgarriff (1996; 1997; 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-occurrence 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. 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. All of the bigrams shared by the two lists concern navigation aspects and directions on the web. The seven bigrams are:

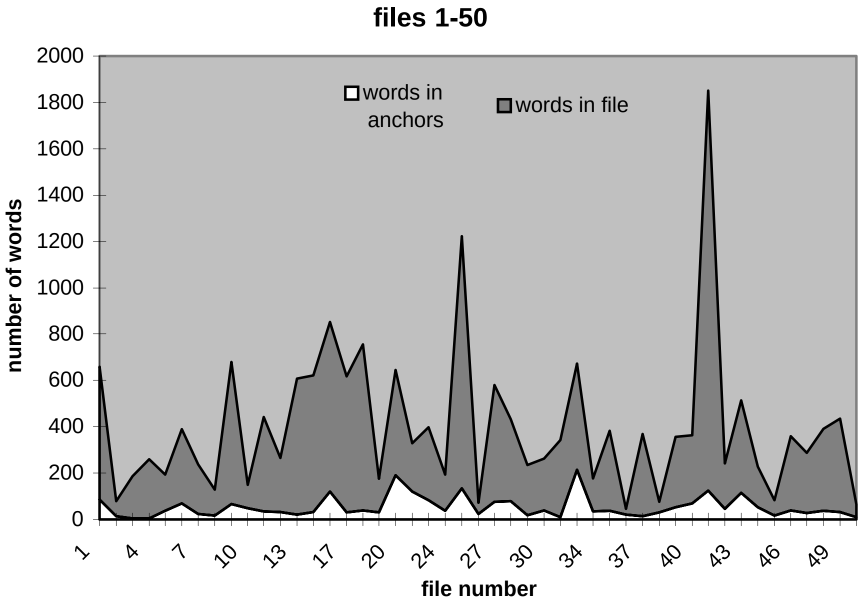
home page
back to
click here
of the
to the
on the
in the

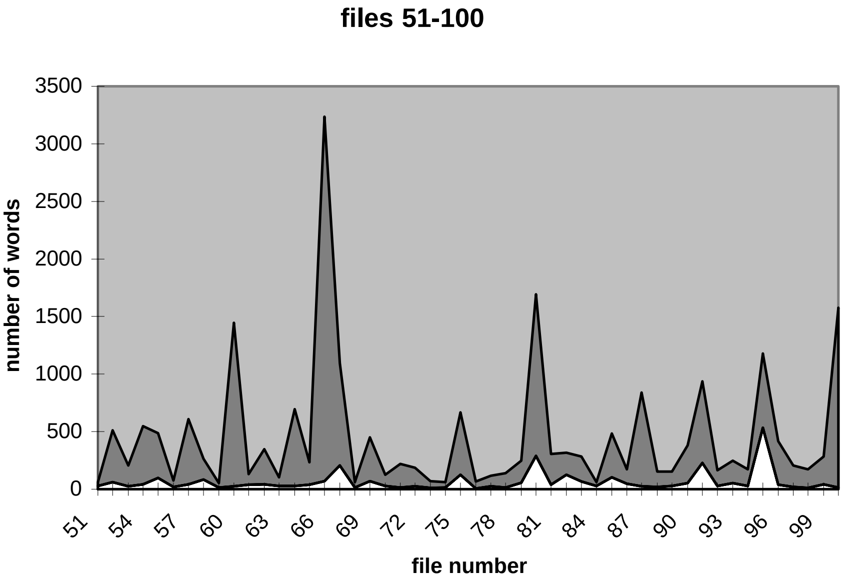
Three of the bigrams are unique to web pages and an assumption can be made that they will not appear in flat text which has nothing to do with the internet. The reason for not finding more instructional combinations in the Home corpus, as can be found in the bigram list of the Parker Corpus, is probably 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 the following section. From this point onward most of the analysis of data will be performed on the Home Corpus. 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 for the web.

Analysing the Home Corpus

In order to better understand the linguistic environment in which people choose to put their hypertext-web document, a detailed study would examine the content of each such document, trying to infer the context in which they were created. In this paper 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 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 finding reinforces the conclusion of Nygren et al. (1995) who 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. 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? If the research community could have an answer as to how many and in what context links can be created then the problem of 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 study.

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

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 paper 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 includes 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 co-occurrence 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 placing 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 their 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 controls navigation. From the hypertext corpora gathered for this study 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:

	number of words			
file	**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 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 what 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.

Conclusions

Although it seems that there is no real order and method in writing hypertext documents for the web, the discussion and findings in this study 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-web documents, 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.

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.

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

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.

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.

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.

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

Kilgarriff A. (1997). 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.

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

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.(1997). 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.

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

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

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

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.