Genre Analysis of Structured E-mails for Corpus Profiling

Malcolm Clark
The School of Computing
The Robert Gordon University
+44 (0)1224 262478
mc@comp.rgu.ac.uk

Ian Ruthven
L12.16 Livingstone Tower
Strathclyde University
+44 (0)141 5483098
Ian.Ruthven@cis.strath.ac.uk

Patrik O'Brian Holt
The School of Computing
The Robert Gordon University
+ 44 (0)1224 262708
ph@comp.rgu.ac.uk

ABSTRACT

This paper reports on our approach to the analysis of genre recognition using eyetracking. We focused on a collection of different types of email which could represent different datasets, such as, mailing lists for calls for papers, newsletters, etc. We found that genre analysis based on purpose, form and layout features is potentially effective for identifying the characteristics of these datasets and we have highlighted some of the new important features of genres. The results from a pilot study showed a clear effect, with an interaction between the email texts and the visual cues or features perceived and also the strategies employed for the processing of the texts. We found, in our small sample, that readers can determine the purpose and form of genres and that during this process some readers do skim the shape of the e-mails (form).

Keywords

Genre, perception, eyetracking, ecological, affordances constructivist, e-mail, corpus, profiling, datsets.

1. INTRODUCTION

Currently, much enterprise/web retrieval research depends on corpora containing features related to uniform resource locators (URLs), links, tags, etc, (for example, the text retrieval conference (TREC) blog). We are proposing alternatives to these content and function conceptual features by using these atomic units as composites. This approach will show the advantages of using genres as a way of profiling corpora.

According to Ingwersen and Järvelin [6], IR is divided into computer science lab experiments versus 'user-oriented' social studies. This study is concentrated on the latter and forms part of a wider human context, namely, how the context of a community gives rise to standardized information forms, which are and should be exploited within corpora. By examining the full text structures and the related features that effectively characterize corpora, we aim to develop alternative retrieval methods which will help fill a large void in IR/NLP research. The IR community, for example, TREC, and the initiative for the evaluation of XML retrieval (INEX) have recently started to understand the importance of (technologically) structured text retrieval.

However, there is also a need to examine the 'socially and naturally structured' texts called *genres*. The various corpora used by the organizations such as Wikipedia (INEX) and TREC (Enterprise, etc) do have one particular thing in common in terms of context: they contain exemplar types of documents. Up to now, the research community has largely overlooked these exemplar types (genres).

The aim of this study is to investigate these types of genres for profiling corpora and, particularly, how they are used within e-mails inside an organisation. We also looked at the human perceptual process (es) and methods which were used to identify and employ them. Within most e-mail accounts there tend to be socially constructed communicative behaviours, namely genres, which emerge to improve the efficiency of the activities in a "community of practice" [23]. Genre benefits organizations financially and administratively by filtering certain types of documents which are not needed; this allows the rapid retrieval of relevant types of information. A user in a legal department, for example, needs a 'legal report' or a 'court transcript' and by means of this filtering can directly exclude other types of structured documents.

We examined the genres in a corpus of emails that were present in a university community email exchange. For this, we decided to implement eye tracking to see how readers react when using new genres/features of documents for characterizing collections, what features or attributes they perceive and whether their perceptive processes can be modelled. This potentially shows how human categorization behaviour can be emulated by a machine for automatic retrieval. In some contexts, in particular, it is important to find out which of the two predominant processes -ecological and constructivist- are present in the subjects' genre recognition tasks. This experiment examines how these processes may be used in such information tasks. These processes will be discussed further in 2.4.

Section two contains four sub-sections which introduce the reader to the motivation for this research, offer general information about how people read, and provide an introduction to genre and two prominent perception theories. Section three contains related work and a discussion about previous work in this area. Section four spells out the research questions, the methods utilized, the references of the corpus used, the variables, task and procedure, the types of data recorded and the participants, concluding with the details of the results. Section five gives the conclusions deriving from this research.

2. BACKGROUND

2.1 Motivation

The main motivation for this study derives from the need to explore genres by the collection of data and to test, empirically, the ways in which people can exploit and use the features of genre from document collections, including e-mails, using their natural perception strategies.

2.2 How Texts are Used

The reading of texts, whether in a web domain or an e-mail, requires many complex processes for a human being. These processes can be used for comprehending the text, guiding the reading to salient properties of the text and so on. But which strategies and features do readers use?

Skimming or scanning are regularly employed for the understanding of the purpose and form of a text. Skimming [13] is a technique for reading which is used to identify the main points in a text. It is performed at a speed several times faster than conventional reading and is normally used when a reader has a large amount of text to read within a limited time and does not need to understand every word, for example, when a student has to perform a literature search for academic papers: an abstract, for example, could be skimmed to judge whether a particular article would be useful/relevant for the current research.

It is a main contention of this research that an important aspect, that is, the layout or genre, of the document structure is not consciously thought about when considering skimming or scanning to find out whether the document is relevant. For example, a user wants to locate a biography of Karl Marx. The layout of a biography is explicit and since a reader examines the results of the search by skimming the structures in the returned documents, the layout or genre is a perfect mechanism for finding the appropriate information quickly [17-19].

Of course, there are many definitions of genre so first it is appropriate to narrow down the meaning for this study.

2.3 Introduction to Genre

Genre is often treated as the classification of movies or books as, for example, “westerns”, “short stories” or “detective novels”. Although, of course, this definition has some relevance, the term “genre” embodies a much wider range of contexts. Although, there has been much discussion on the definition of genre, for the purpose of this experiment, genre was defined by its purpose and form (or layout - see Figure 1) only [25] in which we excluded the other concepts such as style, content etc.

Yates and Orlikoswki [24], in their pioneering work on the concept of genre, suggest that “Genres (e.g., the memo, the proposal, and the meeting) are typified communicative actions characterized by similar substance and form and taken in response to recurrent situations”, which can be used to identify types of organizational communication. The form (and purpose) is the set of structures and layout, which show the user the document’s form through its structure, regardless of the topical nature of the writing.

The purpose, communicative purpose, represents many attributes such as arguments, discourse structure and so on. The form (the readily observable features) contains several attributes (Figure 1); the *structural features* are text-formatting devices, such as lists and headings, and devices for structuring interactions at meetings such as an agenda and chairpersons. The *communication medium* can be pen and paper, telephone, or face to face. Lastly, the *language or symbol system***s** are formed from linguistic characteristics, such as the level of formality and the specialized vocabulary of corporate or professional jargon. The approach we took was to examine whether these categories of purpose and form were actually perceivable.

Good examples of purpose and form are shown in Tables 1 and 2, for example, the purpose of the cinema e-mail is to announce cinema shows, dates and times in order to entice people to come to the cinema.

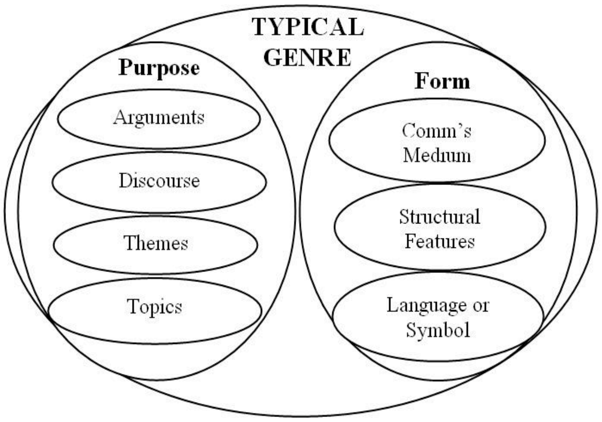
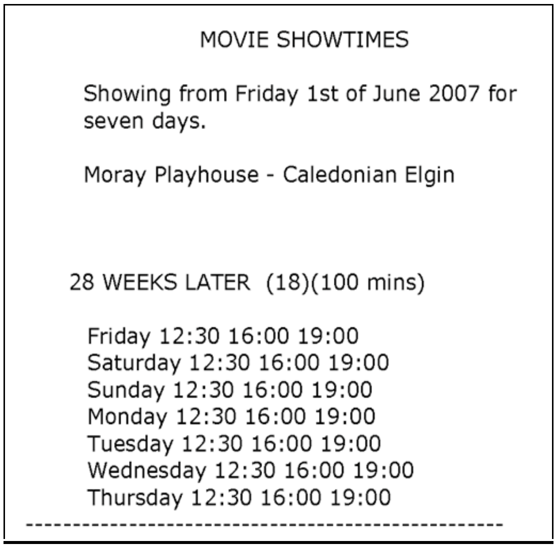


Figure 1. Yates and Orlikowski’s substance (purpose) and form [24]

This email takes the form of a well – defined text layout which enables the reader to access the information easily: the well-defined features and structure in Figure 2 show the purpose and the form, particularly in the way the title, rating and block of show times are laid out.

The experiment described in this paper examines the ways in which the textual features of documents categorize the purposes for which the text has been written and also identifies the form elements which are common to various e-mail genres.



```
MOVIE SHOWTIMES

Showing from Friday 1st of June 2007 for
seven days.

Moray Playhouse - Caledonian Elgin


28 WEEKS LATER  (18)(100 mins)

 Friday 12:30 16:00 19:00
 Saturday 12:30 16:00 19:00
 Sunday 12:30 16:00 19:00
 Monday 12:30 16:00 19:00
 Tuesday 12:30 16:00 19:00
 Wednesday 12:30 16:00 19:00
 Thursday 12:30 16:00 19:00
--------------------------------------------------
```

Figure 2. Edited example of cinema e-mail with explicit structure and formatting

2.4 Perception issues: Introduction

There are two prominent visual perception processes, amongst many others, in which the human animal perceives: ecological and constructivist.

Constructivists, such as Gregory [5], defended a top-down approach according to which perception begins with recognition, that is, the animal uses sensory information, and builds or constructs this incomplete information to make sense of it [11]. Ecological Psychologist Gibson (and others) argued for an alternative and direct (bottom-up) framework for perception and Gibson not only challenged the stages but also introduced the notion of 'affordance' as a centrepiece to his theories. An affordance guides to a property or properties of 'something' that directs how it can be used [4]. To summarize, the ecological school argues that the goal of perception is to perceive in order to act, in this context, it could be the act of directing the attention of the reader to the salient properties of the text, and the constructivists assert that the final goal is that we perceive for recognition.

Of course, locating evidence of the use of these perception processes is no easy task but this paper aims to provide pilot evidence as to how this can be done to form the basis for further study. In our research, this study takes the first tentative steps to show how people use structured texts in relation to genre and perception and will offer some useful empirical data and indicators as to whether capturing these types of interaction with texts is feasible. Following on from the limited pilot studies (section 3) by Watt [22] (ecological) and Toms and Campbell [18] (constructivist), we begin by monitoring how people use e-mail genres.

3. RELATED WORK AND DISCUSSION

With the exception of Toms and Campbell [17, 18] and Watt [22] most authors' contemporary works attempt to examine the uses of genre on the web [3, 7, 9, 14], classification [1, 2, 15] and what they are instead of how we perceive the features contained within the corpora.

Watt [22] believes that genre is present for an intentional purpose, i.e. to give guidance to readers about what to do with the text by assisting the action-making process and that it should thus be employed more widely in IR. He explains that Gibson's theory of Ecological Psychology, perceiving for action and genre work together. For example, the headline on a news story is intended to get the readers' attention, or the abstract of an academic article allows the filtering decision of whether the article is relevant or not. He thus explains that the reduction in the reader's "cognitive load" allows the reader to decide that he/she need not read a whole document as the genre provides the invariant cues in its structure to its relevance.

Gibson's affordances (see section 2.4) are meant to define how meaning and perception are related: he argues that instead of perceiving objects (such as texts) and then adding meaning later, there are visual combinations of invariant properties of objects which provide cues on how to act in relation to these objects. In the case of genre, these invariant properties or features are primarily layout cues, rather than linguistic cues; they occur in two areas and are referred to in this project as shallow (or surface), and deep features. In addition to the issue of investigating features (or invariants) there is also a case for investigating the possible actions which are afforded to the perceiver of documents; this is one of the main tenets of Gibsonian theory: 'perception for action'. The affordances of genre could be defined in terms of drawing the attention of the reader (the perceiver) to salient properties of the communication which could activate a decision that a document is relevant to the search.

Alternatively, Toms and Campbell [18], in their pilot study, leaned towards the perception for recognition process, since they aimed to contrast the function (content) and form in order to discover whether readers can perceive and process form on its own or need semantic content to identify it. Toms and Campbell [18] contend that the 'attributes' of a document's genre enable it to be specifically identified and show that genre attributes play a significant role in identifying documents. They conducted experiments using both form and function, exposing users, with backgrounds in information technology and the academic world, to digital and hard copies of web documents. They suggest that form is scanned and content is read so that possibly two processes are actually ongoing at the same time and that function provides semantic clues which demonstrate the purpose of the genres, whereas when the document structure was shown "participants had to match their sensory response with the corresponding representation stored in long-term memory" [18]. They claim that:

- To distinguish a document using form, the user scans and translates some or all of the visual cues present at the same time to locate the semantic clues.
- The participants constructed or "loaded a set of expectations" which were founded on the available visual clues on the texts.

They argue that perception is a top-down process where the readers recognize the genres through the attributes of the layout which forms the basis of document recognition, and although Toms and Campbell, like Lakoff [8], refer to the bottom-up process and suggest that genres may "act as a single gestalt" [18] they fail to explore other possibilities, such as perception for action and how a genre is perceived when the document is displayed to a reader (Watt fails to explore the perception for recognition concept). In their conclusions, however, Toms and Campbell query how the form of the document affects a user in the first few seconds and this begs the question: how does the structure of a genre aid in text comprehension and use? These are two of the questions which will form a central part of this work.

Although the research by Watt, and then Campbell and Toms, does seem to indicate a bias towards one particular process, it may emerge that they are both correct, but for different tasks and contexts:

- It is possible that documents are identified and used in differing methods depending on the context of the task, skill and expertise of the reader, reading and use.
- Maybe if the task is being performed quickly, skimming is important whereas a person with more time relies on more thorough reading.
- If a user is looking for a text seen previously, then the recognition process is important but if the search is a fresh task which involves looking for a particular unfamiliar genre then the ecological process is vital, to save time.

4. PILOT STUDY

We conducted a small pilot study to examine the questions (4.1) below. By using a well-balanced e-mail collection with eight genres, we tested and monitored five participants (who were from similar academic backgrounds and therefore familiar with some of these types of e-mails) during the timed response identification task set for them.

4.1 Research Questions

- Are purpose and form essential for the identification of genre?
- Are two kinds (ecological and/or constructivist) of processes present in genre 'recognition' tasks?
- How do humans view and utilize the invariant layout cues, such as white-space patterns or other formatting features which constitute genres?
- Do/can participants 'skim' the shape of e-mail texts?

4.2 Method

Since there is a limit to the amount of information that can be gleaned from participants through small-scale user studies and questionnaires, we decided to use eye tracking to add extra types of data (for example, durations of fixation) and some precision to the empirical data. We asked the participants to state what they believed to be the genre of the e-mail while at the same time we recorded any evidence obtained during the eyetracking process. Thus, eyetracking was used to record the form aspects which helped them to identify the genres and also to record any clues as to which perceptual processes the participants may have used during the timed-response driven task.

4.2.1 Eye Tracking

Eye tracking hardware allows the accurate capturing of the eye fixations (durations and gaze, etc) or saccades (number of, etc) to those parts of the screen which are of interest to the perceiver and helps to identify which types of strategies are important for understanding the purpose and form of the text [12]. In relation to the research questions in 4.1, we hoped to record how humans visually perceive different genres of e-mail.

4.3 Corpus

We collected e-mails which commonly occurred in several academics' personal accounts. The e-mail account holders and participants were unrelated, and the participants in the experiment did not contribute any e-mails. The typical e-mails which were found to emerge, especially in this academic environment, were personal library account updates, information technology services downtimes in the institution, calls for papers and so on.

At this stage, we decided to look only at highly visual e-mails and their features instead of 'conversational' exchange e-mails, as this type of structural content will be looked at in later studies. These e-mails are normally composed of several layers or sections, which are organized in a certain form (readily observable features) which defines the type of genre and a distinct purpose (communicative purpose). We were interested in these attributes of e-mail genres because it was necessary to discover how they are used and, in particular, perceived by the readers. This study is comprised of eight typical and repetitive genres of e-mails for the genre identification.

The e-mails collected for this task came from two domains: e-mails from sources internal to the institution of the first author and external e-mails. Although e-mails are genres in their own right which have evolved from memos, all e-mails today contain sub-genres with their own individual purpose and form such as, calls for papers, newsletters, orders and spam. Along with external e-mails (Table 2) there are also internal e-mails (Table 1) relative to particular organizations. Drawing on the purpose and form of the messages, the e-mails utilized in this collection can be defined through their coding.

Table 1. Internal e-mails

Type	**E-mail Purpose**	**E-mail Form**
Outage	Announces downtimes of servers and systems	**Structural features:** title uppercase, emboldened text items listing outage information.
Seminar	Much like call for paper but internal announcement of invited talk.	**Structural features:** uppercase titles centred, block of text about speaker, abstract, and block of text about organizer.
Library	Message from library; reminder that a book is ready for collection/return.	**Structural features:** block of centred text recipient details in uppercase. Opening salutation. Block of text (two paragraphs) terms and conditions, list of renewal item(s).

The e-mails were transformed into images, a process which was necessary for the eyetracker as the system only allowed stimuli in one image format, namely bmp. In total, 48 images of each genre of e-mail were made. There were four representations: normal structure, normal with structure but semantic content replaced with X or 9, no structure (essentially bag-of-words with original word order preserved), no structure and all semantic content replaced with X's or 9's.

Table 2. External e-mails

Type	**E-mail Purpose**	**E-mail Form**
Call for papers	Calls for submissions for conferences and workshops by announcing the requirements and important dates.	**Structural features:** large title, block of centred text (sometimes uppercased). Block of text explaining the event. Bullet lists explaining scope of subjects for conference. Important dates – titles and dates in list format
Cinema	Announces cinema listings, dates and times.	**Structural features:** uppercased cinema name/title rectangular block of text with name of film, rating, length, times per day of show.
Spam	Scam letters with the motive of deceiving people to send money for a false cause.	**Structural features:** spam uses 'letter' variation format. Top lines indicate type of spam i.e. Nigerian letter, Lottery scam etc
Newsletter	Summarizes all the weekly news from an organization, i.e. Aberdeen Football club.	**Structural features:** lists of items emboldened. Opening salutation to the recipient. Emboldened title with small summary paragraph and URL below each for the e-mail. URL at end to un-subscribe.
Orders	Confirmation from a business of an order for item(s) online i.e. Next, Tesco etc	**Structural features:** Order number and 'thank you for the order details'. Table created with format using lines consisting of symbols (- * /) with details of the order: quantity, item ordered unit cost and at very bottom total cost. Delivery address uppercased and date order being delivered.

4.4 Variables

The variables tested were as follows:

1. Purpose/type of genre: (Tables 1 and outages notices, seminars, library, call for papers, cinema, spam, newsletter and orders.
2. Form: represented in four ways (4.3 above). The same data formatting approach was previously used in Toms and Campbell [20]. Watt's [22], later e-mail work, in which he examined one genre - call for papers - was an adaptation of Toms' and Campbell's previous experiment.

4.5 Measurements

The measurements used in the experimental design were the number of *document types identified correctly* by each participant which measures the effect of the genre type. The *length of time it took for each participant to correctly identify each document* measured the effect of form on the participant.

4.6 Task and Procedure

4.6.1 Task

Each of the five participants was shown sixty-four differently structured examples of e-mails (4.6.2), and asked to identify each genre by voice after being shown the stimulus, while the eyetracking system recorded the most discriminating features observed and the strategies used by the participants which led them to identify each genre representation. Examples of strategies could be skimming the shape of the text, such as centred blocks of text [21], or using an 'F-shape' reading pattern [10].

A detailed analysis of the features used by each participant for each stimulus was evaluated by video playback some days later. The eyetracking equipment was fixed to the desk; only a simple verbal response to identify the genre was possible, because detailed discussions (head/face movements) would have interfered with the eyetracking.

4.6.2 Procedure

To control the order effects, the types of document and their allocation were randomized twice: once manually and once also by the eyetracker software facility to randomize the stimuli. We devised the following procedure:

1. The experiment consisted of 4 x 16 blocks of images. Four types of e-mail genre for first two blocks and four more for the 3rd and 4th.
2. Short break of 2 minutes between each 16-image group and repetition of calibration.
3. The bmp images were shown randomly on right hand side monitor
4. The experimenters recorded the eyetracking on the left monitor using desktop recording (Wink) for each image and the identification of each image.

4.7 Data Recorded for Genre Recognition

The several types of data identified as useful for this study were:

1. Eyetracking data which consisted of x/y location saccades and the time the pupils tracked in the location recorded by the Viewpoint Eye tracker.
2. Features and strategies of the participants used for identifying the texts.
3. Video recording of the eye trace (Figure 2) over the stimuli shown on the desktop using a screen capture package named Wink.
4. Participants' timed and vocal responses.

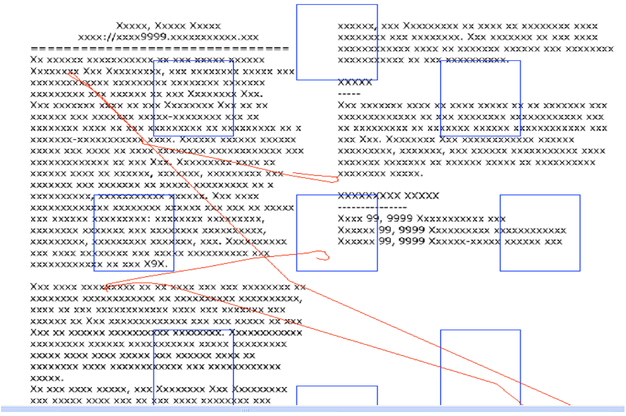


Figure 2. Screenshot shows calls for papers e-mail with semantic content removed, but structure intact. Also shows eye tracking pupil trace (in red).

4.8 Participants

This small study used five participants, all of whom were academic computer scientists. All participants were regular users of e-mail and familiar with the genres being investigated. They were all between twenty and thirty-five years of age and for all but two, English was their first language. They all held computing degrees and were post-graduates. They were very experienced web users, especially with web mail.

4.9 Results

Several types of data were observed. We state our own observations, providing an analysis and our interpretation of the results for discussion.

The response times and amount correct types of data were in direct relation to the purpose and form research question 1 (4.1). The features, strategies and any clues to the perceptual processes (questions 2-4, see 4.1) were recorded using the x/y saccade data and a recorded video of each experiment. The participants also answered a post-experiment questionnaire asking for any feedback they could provide on their strategies.

4.9.1 Amount of genres identified correctly

Although some genres were obviously easier to recognize than others, the participants were generally able to identify the genre (purpose) of the e-mails.

On average, in the course of all the experiments, 63% of the e-mails were identified correctly. This equated to just under 11.5 correct identifications of genres per 16 for each block. Most problems were encountered when the participants were trying to perceive and identify the unformatted with non-content representations. In comparison, identification of the unstructured representations averaged 41.6% whilst the proportion of structured e-mails which were identified averaged 72.9%. Breaking these scores down, we saw that identification of the *original format* averaged 87.5%, and identification of the *original format but with semantic content removed* averaged 77%. The non-structural representations: *semantic content with no structure* was identified on average 68% of the time whilst the identification of the *semantic content and structure removed* averaged 27%.

As might have been expected, the genres which were encountered by the participants on a regular basis (seminars (70%), outages (58%) and library (60%) were identified most frequently while the order (37.5%) genre was the least identified. Calls for papers and spam were identified 66.6% of the time, whilst newsletters and cinema were identified 60% and 62.5% of the time respectively.

4.9.2 Structure versus non-structure-form

The actual perceptions of the form and non-form representations were measured against each other. Participants took less time to judge the genre by form (2.22 seconds) than by semantic content only (2.72 seconds). This was not entirely unexpected, since decisions with structured text are easier to perceive than those with 'blobs' of text.

4.9.3 Identification of genre response time-form

The time response recorded was the point in time when each stimulus was first shown to the point in time when it was exchanged for the next. The eyetracker system recorded this to the thousandth of a second.

Any time recorded over a certain threshold (10 seconds) was omitted from the analysis; this, in most cases, was not due to indecision, but due to a re-calibration or adjustment to the eyetracking equipment.

Table 3. Genre identification response in seconds

Genre	Orig	No form with content	Form no content	No form no content	Avg
Cfp	2.48	2.17	1.53	3.17	2.33
Cin	1.43	1.63	1.27	1.6	1.48
ITS	2.14	2.38	3.14	3.7	2.84
Lib	2.3	2.78	2.51	2.98	2.64
Nl	1.59	2.98	1.24	2.61	2.10
Ord	3.85	4.03	3.88	3.73	3.87
Sem	2.14	2.71	2.43	3.27	2.63
Spam	1.71	1.81	2.03	2.1	1.91
Avg	2.20	2.56	2.25	2.89	

Average measures were taken for each representation of genre (eight) of e-mail. On average, *original format* was identified in 2.20 seconds, and *original format but with semantic content removed* was identified in 2.25 seconds. The non-structural representations: *semantic content with no structure* was identified on average in 2.56 seconds whilst the *semantic content and structure removed* was 2.89 seconds.

The eight genres had mixed results (Table 3). Surprisingly, the cinema genre was recognized with the most ease (1.48 seconds) while orders took the longest to identify (3.87 seconds).

4.9.4 Strategies and distinguishing features to identify genres

In relation to research questions 2-4, we were interested in looking for possible strategies and features being used by the participants:

Features. From the eyetracking data there were many features which were deemed important for each genre from the small sample collected (Table 4).

Strategies. In general, we noticed different strategies for different representations of e-mail. When participants were shown a normal or normal with semantic content replaced representation, they used a circular scanning motion or indeed a 'cross' strategy which consisted of a left-right and then up/down behaviour. When introduced to the unformatted representations, the participants very often used the right side of the block of text in an up and down scanning pattern to identify the genre. With regard to research question two, it was difficult to infer from such a small sample whether the participant was scanning and trying to recognize (constructivist) or perceiving invariant cues (ecologist) from the structure or semantic content or maybe both simultaneously.

Table 4. Important features

Genre	Feature(s) deemed important
Cfp	Dates, centred blocks of text and top title.
Cin	The rectangular shapes of the movie details (made up from large blocks of numerical content for showtimes).
ITS	Apart from uppercase text in normal version the features utilized were inconclusive.
Lib	The normal representation was identified by the book information in a block at the bottom (list format).
Nl	Mostly identified by the first six-ten lines including the title (summary) and the link after each item. Occasionally, the emboldened titles were used as cues.
Ord	The text was scanned up and down continuously. Possible that some were using the left alignment of the text and the lists of ordered items. Currency information was also a cue.
Sem	Inconclusive. Most difficult to identify. Maybe individual features not deemed prominent enough.
Spam	Normal version keywords like 'LOTTO' and address format text. Drawn to uppercased/emboldened text.

However, we obtained clues for research question four, as in some cases, the participants did indeed skim the shape of the texts in the formatted e-mail stimulus examples. When the e-mail text was heavily formatted (centred), e.g. in calls for papers, or aligned left (seminars), the shape of the text, according to the eyetracker data, did seem to play an important role in the genre identification process. In contrast, when the text had all format removed, some participants also occasionally skimmed the shape of the large 'blob' of text but this could mean that they were looking for semantic content/keywords.

Intriguingly, one important result here was that "possible" differences in reading behaviour could be observed in people from different backgrounds, but a larger sample of participants would have to be used to ascertain whether this result was simply a result of coincidence.

4.10 Summary of results

Although we only used a small sample, it was possible to detect some clues as to the importance of purpose and form. A large proportion of the genres were identified correctly, and it may therefore be inferred that form (i.e. the structural features) is indeed important for the perception and identification of genre. Thus, in line with the two previous studies, [18, 22] the data we collected seems to show that genre has a significant effect and is an important factor. The layout and formatting also seem to be important, but the importance of whitespace has been difficult to assess in our study so far.

4.11 Post Comments

All the participants agreed that they understood the task given to them, but they did experience some confusion with regard to the number of genres that they were 'being asked to remember'. It was pointed out that this was a task in which perceptual processes were being tested and which therefore required them to simply attempt to identify the genres.

There was an incidence of contradictory information during the analysis of the questionnaire and eyetrack x/y saccade data. One participant reported that, on occasion, numerical appearance (heavily prevalent, for example, in the cinema genre) was an important clue as to the purpose of the genres. Examination of the eyetrack trace behaviour, however, showed that the participant had not looked at the particular area containing the numerical data! This was an important finding since it reinforced the value of eyetracking during the experiments: although even without the eyetracking system, we are able to infer what is happening when people interact with texts, eyetracking gives us more accurate data and thus more valuable insights as to what is really happening.

5. CONCLUSIONS

Momentum is slowly building, in some circles, towards the acceptance and recognition of the importance of genre, for example, a comprehensive effort is being made by computational linguists, such as Marina Santini et al [16], to build a genre collection repository. In IR circles, unfortunately, genre is largely overlooked in most collections, such as TREC and INEX (a close examination of these various collections will reveal a multitude of genres!).

Our approach has focused on a collection of different types of email which could represent different datasets such as mailing lists for calls for papers, cinema, newsletters etc. We believe that genre analysis based on form and layout features is potentially beneficial to characterize these datsets.

For this paper, we have highlighted some of the new important features of genres, showing how they can be used to characterize corpora and how some typical genres of organization communication [24] exist in most e-mail exchanges within enterprises and organizations which help these to operate efficiently. We have also demonstrated the ways in which the recipients of emails react and use the features which constitute the genres. We have demonstrated some useful alternatives to the content and function features by showing how humans can use these small textual features collectively as genres.

Although the work described in this paper is ongoing it is already perfectly feasible to say that purpose and form are very important for explaining the interaction with textual documents so that a 'community of practice' [23] can operate efficiently. However, there are some other conceptual features which are also useful in a web context, such as functionality, which should not be entirely overlooked.

The two perception processes were difficult to detect but some clues were found in the data collected, for example, when the participant was not aware of the genre type, he/she did not know what to expect in terms of attributes or layout nor did he/she have any previous knowledge to draw from. Therefore, if during the task, the genre was identified as a 'cinema' genre, we could assume that the e-mail had possibly afforded its purpose or, indeed, gained the reader's attention and/or directed the reader to the salient properties of the particular type. For example, the blocks of numerics (layout and format features) afforded the information/action, e.g. the title of a movie, rating and list of times allowed the decision of whether to go and watch a movie or not. Now, if the reader was fully familiar with a type of e-mail genre, again, cinema, this could lead to an 'expectation of purpose/form' with the result that the reader would compare knowledge expectation to visual attributes and thus recognize the purpose and form.

Our future work will involve extending this experiment not only with a larger sample of participants from within our institution but also with others drawn from other faculties, such as engineering and pharmacy, in order to allow more extensive data to be collected to investigate our research questions.

6. REFERENCES

[1] Boese ES. Stereotyping the Web: Genre Classification of Web Documents [Masters]. Ft Collins: Colorado State University; 2005.

[2] Clark MJ, Watt SNK. Classifying XML Documents by Using Genre Features TIR-07 4th International Workshop on Text-based Information Retrieval (DEXA 2007); 2007 3-7 September 2007 Regensburg, Germany. IEEE; 2007.

[3] Dong L, Watters C, Duffy J, Shepherd M. An Examination of Genre Attributes for Web Page Classification. Hawaii International Conference on System Sciences, Proceedings of the 41st Annual; 2008 7-10 Jan. 2008; Hawaii. IEEE; 2008. p. 133-.

[4] Gibson J. The Theory of Affordances. The ecological approach to visual perception. 2nd ed. Hillsdale, New Jersey: LEA; 1986. p. 127.

[5] Gregory RL. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London Series B*. 1980 July 8 1980;290(1038):181-97.

[6] Ingwersen P, Järvelin K. The Turn: Integration of Information Seeking and Retrieval in Context. 1st ed. Croft WB, editor.: Springer; 2005.

[7] Kim Y, Ross S. Examining Variations of Prominent Features in Genre Classification. Hawaii International Conference on System Sciences, Proceedings of the 41st Annual; 2008 7-10 Jan. 2008; Hawaii. IEEE; 2008. p. 132-.

[8] Lakoff G. Women, fire, and dangerous things: what categories reveal about the mind. Chicago: University of Chicago Press; 1987.

[9] Levering R, Cutler M, Yu L. Using Visual Features for Fine-Grained Genre Classification of Web Pages. Hawaii International Conference on System Sciences, Proceedings of the 41st Annual; 2008 7-10 Jan. 2008; Hawaii. IEEE; 2008. p. 131-.

[10] Nielsen J. F-Shaped Pattern For Reading Web Content. Freemont: Nielsen Norman Group; 2006 [updated 2006 17 April 2006; cited 2007 22 September 2007]; Available from: http://www.useit.com/alertbox/reading_pattern.html.

[11] Pike G, Edgar G. Perception. In: Braisby N, Gellatly A, editors. Cognitive Psychology. Milton Keynes: In association with The Open University; 2005. p. 71-112.

[12] Poole A, Ball LJ. Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. Ghaoui C, editor. Pennsylvania: Idea Group Inc; In Press.

[13] Rayner K. Eye Movements in Reading and Information Processing. Psychological Bulletin. 1998;124(3):372-422.

[14] Rosso MA. User-Based Identification of Web Genres. Journal of the American Society for Information Science and Technology. 2008;59(7):1053-72.

[15] Santini F. Automatic Identification of Genre in Web Pages [PhD]. Brighton: University of Brighton, United Kingdom; 2007.

[16] Santini M, Mehler A, Rehm G, Sharoff S. A Wiki for Genre Research. Persistent Uniform Resource Locator(PURL); 2008 [updated 2008; cited 2008 July 19]; Available from: http://purl.org/net/webgenres.

[17] Toms EG. Recognizing Digital Genre. Bulletin of the American Society for Information Science and Technology. 2001;27(2):20-2.

[18] Toms EG, Campbell DG, editors. Genre as interface metaphor: exploiting form and function in digital environments. Proceedings of the 32nd Hawaii International Conference on System Sciences 1999; Hawaii. IEEE.

[19] Toms EG, Campbell DG, editors. Utilizing information" shape" as an interface metaphor based on genre. Proceedings of the 27th Annual Conference of the Canadian Association for Information Science Information science: where has it been, where is it going; 1999 June 1999; Sherbrooke, Quebec. QB: The Canadian Association for Information Science.

[20] Toms EG, Campbell DG, Blades R. Does Genre Define the Shape of Information? The Role of Form and Function in User Interaction with Digital Documents. Proceedings of the ASIS Annual Meeting. 1999;36:p693-704.

[21] Watt SNK. Text Categorization Without Words: Can it Possibly Work. In press 2004.

[22] Watt SNK. Text categorisation and genre in information retrieval. In: Goker A, Davies J, Graham M, editors. Information retrieval: Searching in the 21st Century: John Wiley & Sons; In Press.

[23] Wenger E. Communities of Practice and Social Learning Systems. Organization. 2000 May 1, 2000;7(2):225-46.

[24] Yates JA, Orlikowski WJ. Genres of organizational communication: a structurational approach to studying communication and media. Academy of Management Review. 1992;17(2):299-326.

[25] Yates JA, Orlikowski WJ. Genre Systems: Structuring Interaction through Communicative Norms. Journal of Business Communication. 2002;39(1):13.