Building a document genre corpus: a profile of the KRYS I corpus

V. F. Berninger, Yunhyong Kim[1] and Seamus Ross[2]
Digital Curation Centre (DCC)
&
Humanities Advanced Technology and Information Institute(HATII)
University of Glasgow, Glasgow, UK.
{y.kim, v.berninger, s.ross}@hatii.arts.gla.ac.uk

This paper describes the KRYS I corpus, consisting of documents classified into 70 genre classes. It has been constructed as part of an effort to automate document genre classification as distinct from topic detection. Previously there has been very little work on building corpora of texts which have been classified using a non-topical genre palette. The reason for this is partly due to the fact that genre as a concept, is rooted in philosophy, rhetoric and literature, and highly complex and domain dependent in its interpretation ([11]). The usefulness of genre in everyday information search is only now starting to be recognised and there is no genre classification schema that has been consolidated to have applicable value in this direction. By presenting here our experiences in constructing the KRYS I corpus, we hope to shed light on the information gathering and seeking behaviour and the role of genre in these activities, as well as a way forward for creating a better corpus for testing automated genre classification tasks and the application of these tasks to other domains.

Keywords: genre, corpus, classification, metadata, agreement analysis, document structure, information management, search

1. INTRODUCTION

Digital resources as a source of information are ubiquitous in our everyday life and this tendency is increasing at an exponential rate. The task of managing this information is becoming increasingly demanding. In particular, the representation of digital objects by making transparent their core technical requirements, administrative function and content has become crucial to the efficient and effective management and use of materials in digital repositories (cf. [14]). The manual collection of such information, known as metadata in some domains, is costly and labour-intensive and a collaborative effort to automate the extraction of such information has become an immediate concern.[3] Past efforts in automated metadata extraction (e.g. [4], [6], [16], dc-dot metadata editor;4 [1], [7]) employ methods that often rely on structural elements or presentation styles found to be common among the documents. These structural elements or styles are closely bound to the genre of the document, hence, it seems reasonable that a better understanding of the genre of documents and how they are used in information search would be a key step in developing a broadly effective metadata extraction tool.

The recognition of genre as an informative characterisation of documents is also currently being awakened in information retrieval (e.g. see [3] and [15]). It is becoming increasingly apparent that the topic of a document has limitations in conveying the relevance of a document to a pre-defined purpose or objective (for example, looking for a fictional piece about Cleopatra rather than an academic article). Despite the recognition that genre plays a strong role (going back even to its educational role in philosophy and rhetoric – consider the genre theories of Plato and Aristotle) in the effective management of information and the understanding of social actions (cf. studies of [12] and [17]), there is a severe lack of consolidated genre schema and labelled data ([11], [13]) to assist the examination of genre and its value to information retrieval, rhetoric, social organisation and corpus linguistics. To address this gap, we have built a corpus (KRYS I) consisting of documents labelled with genres. In the present paper we would like to describe our experiences in building KRYS I, consisting of documents and their classification into 70 genre classes.

[1] Research Fellow, School of Computing, Robert Gordon University, Aberdeen, UK.
[2] Dean, Faculty of Information, University of Toronto, Toronto, Canada.
[3] Observed also by the Cedar Project (2002) at the University of Leeds: http://www.leeds.ac.uk/cedars/guideto/collmanagement/guidetocolman.pdf
[4] dc-dot, UKOLN Dublin Core Metadata Editor, http://www.ukoln.ac.uk/metadata/dcdot/

The compilation provides observations on human classification behaviour and is expected to serve as a valuable resource for profiling textual stylistics with respect to genre and the realisation of the automated classification process of the same. In particular, we present an analysis of classification agreement between labellers and the rationales they have provided for submitting documents as samples of genres within the collection. This is intended to be part of a pilot study in designing a genre corpus (cf. [2]) and will allow us to assess the viability of genre classification as an automated process and the way forward to refining the corpus.

TABLE 1: Scope of observed genres

Genre Group	**Genre**	**Genre Group**	**Genre**
Book	Academic Monograph Poetry Book Book of Fiction Other Book Handbook	**Information Structure**	List Catalogue Raw Data Table/Calendar Menu Form Programme Questionnaire FAQ
Article	Abstract Magazine Article Scientific Article Other Research Article News Report	**Evidential Document**	Minutes Legal Proceedings Financial Record Receipt Slips Contract
Short Composition	Poem Fictional Piece Dramatic Script Essay Short Biographical Sketch Review	**Visually Dominant Document**	Artwork Card Chart Graph Diagram Sheet Music Poster Comics
Serial	Periodicals (Newspaper, Magazine) Journals Conference Proceedings Newsletter	**Other Functional Document**	Guideline Regulations Manual Grant or Project Proposal Legal Appeal, Proposal or Order Job, Course or Project Description Product or Application Description Advertisement Announcement Appeal or Propaganda Exam or Worksheet Fact Sheet Forum Discussion Interview Notice Resume/CV Slides Speech Transcript
Correspondence	Email Letter Memo Telegram	**Treatise**	Thesis Business or Operational Rpt. Technical Rpt. Miscellaneous Rpt. Technical Manual

2. CORPUS DESCRIPTION

The KRYS I Corpus was created as part of an effort to automate document genre classification and to develop its role in the automated extraction of metadata from digital documents to be ingested into a repository. The corpus is organised into a schema of 70 genres in 10 genre groups (Table 1).

Students of the University of Glasgow were assigned genres from the schema and given the task to find a maximum of 100 documents belonging to the corresponding genres. Some students were assigned a single genre while others were assigned more but the documents in each genre studied in the Phase I analysis (Section 3.1) were submitted by a single student. These documents were required to be in English and in PDF format. The students had no other definition of the genre other than the name and were not allowed to confer. Furthermore they were asked to give a rationale describing the reasons for submitting each document. By providing no definition we expected to gain insight into whether there is some commonality of genre vocabulary across a broad community and by asking them to provide rationales for submission we hoped to determine whether a universal definition and schema of genres in tune with human information seeking behaviour could be devised. The exercise left us with a corpus of 6,494 documents, gathered during two independent collection phases (Phase I and II). The data from Phase I (5544 documents) were reclassified independently by two secretaries without prior knowledge of the initial classification. This resulted in 5305 documents[5] which are provided with at least three labels. There were five other labellers who later volunteered to classify documents. This resulted in 1016 documents in the database with exactly one label, 105 documents with exactly 2 labels, 5249 documents with exactly 3 labels, 123 documents with exactly 4 labels and one document with five labels.

Several experiments have been carried out already to develop a automated classification method with the collected files (e.g. [9], [10]) and, some preliminary attempts have also been made to examine the human agreement with respect to different genre classes ([8]). These results have shown that the classes easily detected by human labellers are also those detected with some success by automated processing of frequent words, images, and bag of words ([9]) as well as the gap between re-occurring symbols ([10]). However, these analyses were conducted at a stage when the collection process was still in progress and did not include an analysis of the entire collection nor the rationales submitted by the students. This paper is a summary of our analyses of these to provide a starting framework for mining features for genre and other genre related studies.

We noticed three constantly re-emerging error patterns in the initial document retrieval conducted by students: they submitted

1. documents which were not examples of the genre but topic related to the genre (e.g. instead of actual emails, research articles about email were found labelled as Email) [Error type I];
2. empty templates as examples of the genre (e.g. instead of selecting ‘actual’ receipts, empty receipt forms were found labelled as Receipts) [Error type II];
3. entire magazines, conference proceedings or journals as research articles, and vice versa [Error type III] (cf. [8]).

Some may find the phrase “error patterns” inappropriate in a study of the labelling agreement, as we are taking the judgement of the classifiers at face value. However, it must be noted that, just as an error in the transmission of data should not be confused with variation in interpreting the received data, the “errors” due to the misinterpretation of the task should not be equated with the “confusion” arising from the subjective nature of the task. If we had asked several students to retrieve the documents to be included as samples of a single genre, then the errors may be interpreted as common classification behaviour, but, as the documents initially included in each genre have all been retrieved by only one person, such an interpretation would be premature. These errors could just as well have been a result of the students' lack of interest in trying to achieve high quality, only concentrating on quantity, in their work. The briefness of introduction into the work may also have contributed to a misunderstanding of the task. We will discuss this further in the next section.

3. AGREEMENT ANALYSIS

In the previous examination ([8]), involving the three labels acquired during Phase I of the document collection, we concentrated on the agreement between selected labellers. After Phase II was complete, disparate identities of labeller have emerged as labellers of each document (see Section 2). We felt the best way to measure agreement at this stage would be with respect to agreement of labels given to each document (regardless of the labeller identity). It should be noted, however, that analysis based on the set of labels given to documents is expected to result in a higher percentage of agreement compared to the analysis given in Phase I which takes the identity of the labeller into consideration.

3.1 “Phase I” analysis

There are 5305 documents with three labels, one each given by Secretary I, II and the initial student classification. The agreement between all possible pairs of labellers as well the total agreement is displayed in Table 2 (Table 3.1 from [8]). We find the following patterns: although it was predicted that the secretaries would agree more with each other on the documents due to their training, both of them agreed more with the student labelling (Secretary I agreed with about 52% of labels that the students assigned to the documents while Secretary II agreed with about 54%); the difference is much smaller between these two pairs of labellers (2%) when compared to the difference

[5] The number did not reach 5544 because of a technical error in the reclassification interface.

between the agreement resulting from either pair and the agreement resulting from the labels of the pair of secretaries (46%).

TABLE 2: Human Agreement

Labeller Group	Agreement
Student & Secretary I	2,745*
Student & Secretary II	2,852*
Secretary I & II	2,422*
All three	2,008*
*out of 5305	

The disagreement between Secretary and Student may be partially due to the three initial errors (or disagreement on instances of these errors) of students mentioned at the end of Section 2. Also, we do not expect Secretaries to be well practised in the classification of very specific research domain genres (e.g. Scientific Research Article). Secretaries are further trained to recognise limited schema of very domain specific genre classes which are defined by internal policies: confined to a schema consisting of several similar genres with no specified definition (e.g. Handbook and Manual; Memo and Email; Scientific Article and Other Research Article; Poetry Book and Poems) they may be expected to disagree often.

We have previously examined the agreement between these labellers with respect to each genre ([8]). We examined this using the average percentage of agreement and the deviation of agreement across pairs of labellers. The result from this work is presented again in Table 3. The numbers on the left hand side indicates average percentage of agreement and the numbers in the top row indicate the deviation of agreement. That is, the square in the top left hand corner (darkest square) contains genres with the best agreement (on the basis of the two metrics) and the square in the bottom right hand corner (lightest square) contains genres with the poorest agreement.

TABLE 3: Partition of documents according to human labelling agreement.

Deviation Avg. Agreement	0-0.1	0.1-0.2	0.2 - 0.3	0.3+
0.90+	Minutes Handbook CV Sheet Music			
0.80-0.90	Exam Worksheet	Speech Transcript	Email	
0.70-0.80	Poem Form		Thesis Letter Technical Report	Book of Fiction
0.50-0.70	Periodicals		Memo	Slides
0-0.50	Advertisement Academic Monograph Magazine Article		Business Report Scientific Article	Abstract Technical Manual Poster

The partition in Table 3 shows that, despite the different classification behaviours of students and secretaries observed above and the high degree of disagreement (46-54%) between classifiers, there is a high percentage of agreement with respect to the selected genres such as Minutes, CV, Sheet Music and Handbook (all greater than 90%; greater than 95% in the case of CVs, Sheet Music and Minutes). This suggests that there are genres widely recognised across domains, while Abstract, Technical Report and Poster are genres that require more definition in the form of contextualisation as domain and social actions (cf. [12]).

It is also interesting to note that some genre (Advertisement, Academic Monograph and Magazine Articles) exhibit a consistent level of disagreement, while genre classes such as Thesis and Technical Report vary widely across different pairs of labellers being considered. It may be conjectured that this is result of the latter set of classes representing a vocabulary used within selected communities (say, student community), while the former set may require a narrower community to be considered well defined (even as narrow as a community of one or two individuals).

3.2 “Phase II” analysis

To conduct the statistical analysis of agreements on the most stable, largest possible sample, in this analysis we have concentrated our attention on the documents which had exactly 3 labels (5249 documents).

A total of 4022 documents had at least two agreeing labels. 1953 of these had three agreeing labels. They are expected to disagree often. This means that, although 77% of the documents were thought to have the same genre by at least two people, a third person agreed with their decision only approximately half of the time (48.56%). There are certain genres in which the agreement rate is much higher and some in which it is much lower than in others. In the genre Conference Proceedings, for example, nearly 60% of documents were labelled the same by all three labellers. In the genre Abstract, on the other hand, only 0.78% 0f the documents were accepted by all three labellers as belong to the same genre. The greatest agreement can be found in Sheet Music and Resume/ CV as well as Minutes. These are three genres which are extremely easy to identify due to their distinctive form and content.

For a given genre G, let G1, G2 and G3 denote the number of documents that have been assigned the genre G exactly one, two and three times, respectively. The number G1 can be thought of as the number of documents exhibiting total confusion, G2 partial agreement and G3 total agreement. As such, any genre G, satisfying G1 < G2 < G3 (Group I), is conjectured to be genres well recognised across a broad range of communities and/or is distinctive in its vocabulary and presentation. Likewise, any genre G, satisfying G3 < G2 < G1 (Group VI), is likely to be a genre that is understood only within a domain specific social action (cf. [12], [17]) or those that share social action with other genres in the schema. We have partitioned the documents according to the relationship between G1, G2 and G3 (second column, Table 4). Group I (Table 4) does seem to consist of genres defined by broadly understood social activities (e.g. conference, job application), while those in Group VI are subject to domain specific interpretation (e.g. distinction between scientific article and other research articles)

To understand the cause for confusion in genre classification tasks more fully we have presented a selection of the labels with which the labels in the second column are confused (third column, Table 4). We have not listed the labels found in confusion with the genres in the last group; the confusion widely varies and did not seem to exhibit recognisable tendencies.

It is hard to make conclusions on the basis of labels from such a small number of labellers. However, speculatively speaking, genres of Group II seem to be confused with other genres associated to similar social activities (e.g. the communicative purposes shared by Email, Letter and Memo). Confusion in Group III seem to be largely the result of Error Type I (e.g. Magazine Articles about Comics). Genres confused with those of Group IV often share similar components (e.g. scientific articles with a diagram being labelled as diagram). Error Type III confusions seem to be prolific within Group V (e.g. empty forms for receipts labelled as receipts).

Further analysis based on more reclassification (perhaps after combining similar genres in the schema) is required for stronger conclusions. The examination here, however, demonstrates confusions to be often due to

- similar social actions: for example social function provided in common by Email and Memo,
- shared super- or sub-component indicative of disparate genres (e.g. Error type III),
- personal and institutional policies, (e.g. convention and domain specific lingo).

TABLE 4: Partition of genre classes: according to relationship between one, two and three label agreements.

Relation	**Genre** (# of docs given the label at least once)	**Confusion**
Group I G1 < G2 < G3 (12 classes)	Conference Proceedings(118)	Legal Proceedings, Handbook, Abstract, Essay, Other Research Article
	Dramatic Script (55)	Book of Fiction, Fictional Piece, Speech Transcript, Thesis, Abstract
	FAQ (109)	Questionnaire, Other Research Article
	Grant or Project Proposal (112)	
	Interview (100)	Magazine Article, Periodicals
	Menu (98)	Advertisement, Catalogue
	Minutes (104)	Legal Proceedings, Programme, Conference Proceedings, Forum Discussion
		Poetry Book
	Poems (58)	Raw Data, Chart, Factsheet, Form, Guideline, Other research Article
	Questionnaire (103)	
		Short Biographical Sketch, Form
	Résumé/CV (111)	Advertisement, Dramatic Script, Magazine Article, Questionnaire, Speech Transcript
	Sheet Music (42)	
		Abstract, Comic, Essay, Interview, Minutes, Slides
	Speech Transcript (103)	

Relation	**Genre** (# of docs given the label at least once)	**Confusion**
Group II G1<G3, G2<G3, G1>G2 (4 classes)	Catalogue (113)	Handbook, Advertisement, List, Menu
	Job, Course, Project Description (103)	Advertisement, Announcement, Scientific Research, Other Research.
	Letter (138)	Email, Notice, Memo
	Regulations (126)	Guideline, Handbook, Miscellaneous Rpt.
Group III G1>G3, G2<G3, G1>G2 (10 classes)	Comic (91)	Magazine Article, Academic Monograph, Abstract, Handbook, Thesis
	Email (85)	Essay, Manual, Memo, Miscellaneous Rpt., Other Research Article
	Exam or Worksheet (47)	Manual, Guidelines, Miscellaneous Rpt., Slides
		Guideline, Other Research Article, Poster
	Factsheet (192)	Manual, Technical Manual, Guideline
	Handbook (274)	Advertisement, Announcement, Letter, Form
	Notice (30)	Technical Manual, Poster, Other research Article, Manual, Handbook, Diagram
	Product or Application Description (162)	List, Raw Data, Factsheet, Chart
	Table or Calendar (87)	Scientific Research Article, Raw Data, Other Research Article, Miscellaneous Rpt.
	Technical Report (219)	*same as above,* Poetry Book, Comic
	Thesis (168)	
Group IV G1>G3, G2>G3, G1>G2 (7 classes)	Book of Fiction (17)	Dramatic Script, Poems, Scientific Research Article, Journals
	Contract (103)	Form, Receipt
	Financial Record (109)	Business or Operational Rpt., Raw Data
	Poetry Book (94)	Magazine Article, Essay, Review, Poems
	Poster (142)	Advertisement, Artwork, Factsheet
	Receipt (64)	Contract, Form
	Slides (115)	
Group V G1<G3, G2>G3, G1>G2 (4 classes)	Form (201)	Factsheet, Receipt, Slips, Letter, Contract
	Forum Discussion (109)	Other Research Article, Essay, Magazine Article
		Review, Essay
	Short Biographical Sketch (104)	Letter
	Telegram (18)	
Group VI G3 < G2 <G1 (33 classes)	Abstract (129)	*Not reported*
	Academic Monograph (44)	
	Advertisement (70)	
	Announcement (62)	
	Appeal or Propaganda (33)	
	Artwork (32)	
	Business or Operational Report (118)	
	Card (85)	
	Chart (123)	
	Diagram (103)	
	Essay (239)	
	Fictional Piece (14)	
	Graph (95)	
	Guideline (171)	
	Journals (127)	
	Legal Appeal Proposal or Order (11)	
	Legal Proceedings (113)	
	List (84)	
	Magazine Article (186)	
	Manual (180)	
	Memo (101)	
	Miscellaneous Report (239)	
	News Report (40)	
	Newsletter (99)	
	Other Book (50)	
	Other Research Article (347)	
	Periodicals (Newspaper or Magazine) (100)	
	Programme (73)	
	Raw Data (141)	
	Review (150)	
	Scientific Research Article (235)	
	Slips (18)	
	Technical Manual (219)	

4. RATIONALES

The rationales that were given by the students were of very different nature. To make the analysis clearer, we divided them into six groups.

Group A (the name of the genre)

The students had found documents, which fit into their own internal definition of the given genre and labelled them with the genre but went no further to explain the features that define the genres for them. Example rationales included:

- "abstract"
- "essay"
- "it fits the genre of 'Letter'"

Group B (description of the topical content paired with the genre)

This variation went from a one-word definition to larger pieces of text. For example,

- "*educational* poster"
- "Slide show which takes *American democratic ideals*, such as the importance of constitutional government, and attempts to prescribe its principles to a simplified vision of a 'new Iraq'"
- "*theological dualism* and poetry"

Group C (external knowledge, context, and high level stylistic pronouncement)

The student used a high level analysis of the context and style of the document or drew on external knowledge which led to the creation of the document and included this into the rationale. In the category "Book" for example, one rationale explained that it was also available in hard copy, information not available within the document itself. More examples include:

- "This is a notice of response to an opinion release. *It is a notice in advance of an action*, in this case sending an e-mail. It is suitably curt and straight to the point of what it sets out to achieve."
- "Topic covered would be current affairs *to target readership at time of publication*.
- Written in a *journalistic style*. Gives a digest and analysis of factual information."

The third example does not detail what features or part of the document might indicate journalistic style.

Group D (single word description of an aspect or part of the document)

The rationales in this group specified the parts of the document that led them to believe that it was part of a certain genre. They included single words such as "layout", "title" or "content".

Group E (further description of an aspect or part of document)

The rationales not only listed a part (e.g. title) of the documents they used (as did group D) but also described the words that were contained in the corresponding parts (such as the location and name of the genre or related words appearing in the document) or the fact that the document included an author's name or a date. This group also mentioned objects in the files such as images, graphs and tables.

Group F (discounting genres from a range of possible genres)

This group included rationales based on what the document was NOT. For example, if an article did not have the features of an abstract or a magazine article, it was concluded to be a Scientific Research Article.

Analysis

The proportion of rationales in Groups A, B, C, D, E, F were approximately 9%, 43%, 16%, 5%, 22%, 5%, respectively. The rationales in Group A are interesting only in so far as the type of information they convey are extremely different from the other groups. The students had either not understood the task or made their work easier by only giving the genre name. On the face of it, the students associated to Group B also seem not to have understood the task, but, given the high percentage of such instances, we can not discount the possibility that genre and topic classification intertwined and, perhaps, that human classification activities often rely on using one of these classifications as a support mechanism for determining their approach to performing the other classification. The rationales in Group C are highly relevant to genre in that it discusses the linguistic style of the document as well as the goals one might be trying to achieve in terms of effect or target audience by employing that style. The relevance however fails to translate into computationally viable solutions because the style is discussed at a high level, lacking the detail in how this style is detectable within the text. The rationales in Group D, gives us an interesting hint towards the different features that distinguish genres. Knowing that the title of a document indicates the nature of its genre is one step further to the solution. The descriptions in Group E, however, give us a more profound insight into the labelling process in the human mind. The students in Group E for example also found many results in looking at the title of the document. Often the name of the genre would appear in this and thus identify the document. The fields in Table 5 show the characteristics which appeared in

Newsletters, Newspapers and News Reports. The characteristics are not necessarily defining features of the genre but when marked with “YES” were found in more than one document.

TABLE 5: Characteristics in similar genres

	“News” in title	**Contents page**	**Issue number**	**“Newsletter” in title**	**“*-post” or “*-times” in the title**
Newspaper	Yes	Yes	Yes	No	Yes
Newsletter	Yes	Yes	Yes	Yes	No
News Report	Yes	No	No	No	Yes

Not only did the rationales in Group E describe the frequency and position of certain words or numbers in certain genres but also explained the appearance of certain objects and forms in the documents. Certain genres such as academic monographs for example would in many cases contain graphs and tables such as a bibliography which can also be identified by its from just as much as its title. News articles and newspapers as well as posters often contained images, while comics were defined by this feature and most likely to have it.

Some genres had such distinct features that no other genre would contain. Sheet Music for example contains clefs and notes as well as bar lines which give those documents the most individual characteristics of all. The distinctiveness of this genre recognisable across a broad spectrum of communities is reflected again in the high level of agreement (approx. 95% across all labellers) we have found with respect to that genre in the analysis of Section 3. Although not quite as obvious as Sheet Music, E-mails seem to be identified by their unique headers which contain a “From:” and “To:” line, followed by an e-mail address in addition to the distinctive “Cc:” and “Subject:” line.

The rationales of Group F the similarity and, thus, the possibility of confusion between selected document genres. Giving the information that, the document concerned is a comic book, rather than an article about comics, or a document, that contains comics, also informs us, that this is a common mistake, made by search engines. It is hence necessary to check the document for certain features that it should not contain to identify it definitely as a comic book.

Although the rationales in Group E indicate the features of document genre that are perhaps most useful to computational solutions in automated genre classification, the features are low level and any machine learning technique developed on this basis can not change with genres evolving over time and organisational objectives without expensive re-labelling and re-training exercises, It seems necessary to develop a link between high level stylistic aspects identified in the rationales of group C to those low level features in Groups D, E, and F. Further, extensive study should be conducted to understand the role of document topic in genre classification and vice versa to create a robust information management system inclusive of both classifications.

The analysis of the rationales can be summarised as follows:

- There are some low level features in a document that distinguish selected genres very well (e.g. Group E).
- The topic of a document is often confused or used in conjunction with genre to identify document class (see Group B).
- The classification of document genre might be performed by considering other possible genres (e.g. Group F).
- A great deal of indicators can be found in headers and titles of documents (e.g. Group D).
- There is a weight put on the objectives, style and readership of the document (e.g. Group C).
- Apart from descriptions related to topicality the most frequently mentioned rationales were related to style and location and type (e.g. author name, date, genre name) of low level features (e.g. Group C and Group E).

We failed to gather similar sets of rationales from the secretaries. It would perhaps provide more context to carry out the same study with respect to labellers from other backgrounds to compare the results and make a more rigorous characterisation of genre classes. As mentioned earlier, the current study is meant to form only a pilot study of building a genre corpus.

5. OVERALL CONCLUSION

In this paper we have presented agreement and behaviour analysis of humans genre classification activities within the KRYS I corpus. The results illustrate the complexities involved in building a genre corpus. It is clear from the results of this paper that successful automation of genre classification would involve a better categorisation of document structure, content, and form, in relation to their style, purpose, and the social activities they entail as communicative acts. In addition, we need to better understand the role of topical content in document genre classification. The relationship between genre, user activity, and social action identified within the rationales could be used to formulate better genre definitions.

In our process, we did not use definitions of the genres in order to gauge how established genre vocabulary might be across labellers. Now that the initial analysis has identified the genres that are immediately recognisable by name alone, the study might benefit from a new classification based on genres equipped with definitions derived from the rationale analysis presented in this paper. We are also hoping to gather some information from users of the Corpus online[6]. Further labelling performed by volunteer classifiers from other background using the classification system available online [7] may also help to understand the extensibility of the results in this paper.

Some have studied word statistics and statistics of linguistic patterns to determine bias and homogeneity of datasets (e.g. [4]). We were, however, interested in establishing human genre classification behaviour and reasons behind disagreements, before studying the relation between genre and intra-textual statistics. Without a sense of what constitutes a reasonable genre classification palette, we felt that the intra-textual statistics would not be properly understood within context. However, it is without doubt that a study of intra-textual statistics would be invaluable.

We realised while building the Corpus that there is also a significant amount of difficulty in creating such resources that can be shared due to legal issues such as copyright infringement. Copyright holders are not always clearly indicated and to find this information is often impossible. The process of contacting and receiving replies from all owners by e-mail is also hindered by spam mail filters. It is necessary to find a way of contact which avoids spam filters and draws attention to the message.

The future of online cooperation through databases depends on the legalisation process. It is the large amount of anonymity of persons on the internet that complicates the possibility of legal clearance. If it is possible to find a way around these hurdles, the development of databases and other means of academic cooperation will be vastly improved.

6. ACKNOWLEDGEMENTS

The KRYS I Corpus was constructed during a project supported by the DELOS: Network of Excellence on Digital Libraries (G038-507618)[8], funded under the European Commission's IST Sixth Framework Programme, provides a key framework and support, and the UK's Digital Curation Centre (DCC)[9], funded by the Joint Information Systems Committee (JISC)[10] and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC)[11][GR/T07374/01].

REFERENCES

[1] Bekkerman, R., McCallum, A. and Huang, G. (2004) Automatic categorization of email into folders. Benchmark experiments on enron and sri corpora. Technical Report IR-418, Centre for Intelligent Information Retrieval, UMASS.
[2] Biber, D. (1993) Representativeness in Corpus Design. Lit Linguist Computing.1993; 8: 243-257
[3] Dewdney, N., VanEss-Dykema, C., MacMillan, R. (2001) The Form is the Substance: Classification of Genres in Text. Proceedings of ACL Workshop on Human Language Technology and Knowledge Management.
[4] Giuffrida, G., Shek, E. and Yang, J. (2000) Knowledge-based metadata extraction from postscript file. In Proceedings of the 5th ACM International Conference on Digital Libraries, pp. 77–84.
[5] Anne De Roeck, Avik Sarkar and Paul Garthwaite. Frequent Term Distribution Measures for Dataset Profiling. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC). Lisbon. 2004. 1647-1650
[6] Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z. and Fox, E.A. (2003) Automatic document metadata extraction using support vector machines. In Proceedings of the 3rd ACM/IEEECS Conference on Digital Libraries, pp. 37–48.
[7] Ke, S.W. and Bowerman, C. (2006) Perc: A personal email classifier. In Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006), pp. 460–463.
[8] Kim, Y. and Ross, S. (2007) Variation of word frequencies across genre classification task. Second DELOS Conference on Digital Libraries, 5-7 December 2007 Tirrenia, Pisa (Italy), http://eprints.erpanet.org/134/01/YKSRDELOS05122007_FinalV2.pdf
[9] Kim, Y. and Ross, S. (2008) Examining variations of prominent features in genre classification. In Proceedings 41st Hawaiian International Conference on System Sciences, IEEE Computer Society Press, , ISBN-13: 978-0-7695-3075-8, ISBN-10: 0-7695-3075-3, ISSN 1530-1605. http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=4438695
[10] Kim, Y. and Ross, S. (2008) Formulating representative features with respect to genre classification tasks. LDV Forum Special Issue on Genre. To appear.

[6] http://www.krys-corpus.eu
[7] http://genre.hatii.arts.gla.ac.uk

[8] http://www.delos.info
[9] http://www.dcc.ac.uk
[10] http://www.jisc.ac.uk
[11] http://www.epsrc.ac.uk

http://valian.kgf.uni-frankfurt.de/gldv/

[11] Kwasnik, B. and Crowston, K.(2005) Introduction to the special issue: Genres of digital documents. Information Technology & People, 18 (2), 76 - 88. Emerald Group Publishing Limited. ISSN: 0959-3845. DOI:10.1108/09593840510601487.

[12] Miller (1984) “Genre as Social Action”, Quarterly Journal of Speech, Vol 70, 151-167.

[13] Santini, M. (2007) Automatic identification of genre in web pages. Thesis submitted for the degree of Doctor of Philosophy, University of Brighton, Brighton, UK.

[14] Ross, S., Hedstrom, M. (2005) Preservation research and sustainable digital libraries. International Journal of Digital Libraries. DOI: 10.1007/s00799-004-0099-3

[15] Stein, B. and Meyer zu Eissen, S. (2006) Distinguishing Topic from Genre. In Klaus Tochtermann and Hermann Maurer, editors, Proceedings of the 6th International Conference on Knowledge Management (I-KNOW 06), Graz, Journal of Universal Computer Science, pages 449-456. ISSN 0948-695x.

[16] Thoma, G. (2001) Automating the production of bibliographic records. Technical report, Lister Hill National Center for Biomedical Communication, US National Library of Medicine.

[17] Yates, J. and Orlikowski, W. (2002) “Genre Systems: Structuring Interaction through Communicative Norms.” Journal of Business Communication 39.1: 13-35.