**University of Brighton**

*ITRI-04-03*

# State-of-the-Art on Automatic Genre Identification

Marina Santini

**January, 2004**

Information Technology Research Institute Technical Report Series

ITRI, Univ. of Brighton, Lewes Road, Brighton BN2 4GJ, UK
TEL: +44 1273 642900 EMAIL: firstname.lastname@itri.brighton.ac.uk
FAX: +44 1273 642908 NET: http://www.itri.brighton.ac.uk

# *State-of-the-Art*

# *on*

# *Automatic Genre Identification*

**Marina Santini**

*1st version*

*January 2004*

Please, feel free to send comments, suggestions or hints for amendments and improvements. I will be glad if you point out inaccuracies (from typos to misinterpretations). Thank you.

Marina.Santini@itri.brighton.ac.uk

TABLE OF CONTENTS

# 1 Introduction

The aim of this survey is to collect in one single document a list and a short description of all the published works on automatic genre identification/classification.

Genre is an important categorical concept and many research communities are trying to exploit it for a range of purposes: document management; the automatic generation of genre specific documents; information filtering; and so on. This survey has been carried out only in a restricted research field, automatic genre identification/classification.

We begin by stressing that the terminology to refer to the "type of a document" to be detected automatically is confusing, overlapping and fuzzy (Section 2). There is no agreement either on the genre labels, or on the genre systems and taxonomy. A broad range of terms has been used: "genre", "text type", "style", "stylistic genre", "functional role", etc. In addition, the list of detectable genres is different from one work to another. The genre labels range from the traditional "editorial", "reportage", "academic prose", etc., to "fact" and "opinion", or to "subjective" and "objective", and so on.

An overview of the various projects on automatic genre identification/classification is then presented (Sections 3 and 4). The description of BIBER (1988) is not included, but considered as a pioneering work and a milestone in this field of research. Section 5 contains our reflections on automatic genre identification/classification, and Section 6 draws some conclusions.

No attempt is made to define the theoretical notion of "genre" and more specifically what are the genres of digital or Web documents. These issues will be addressed elsewhere. It is worth noting that most projects on automatic genre identification/classification do not bother very much with these theoretical issues: in many cases what they aim to achieve is a classification of documents not based on the "content", but on other features. Most works use linguistic features (but here, too, different sets of linguistic features have been used), and for Web documents also, sometimes, layout features.

# 2 Genre Identification

Genre[1] is a multifaceted concept, and definitions of what genre is remain fuzzy and slippery. In the broadest sense, it can be used "to refer to a distinctive category of discourse of any type, spoken or written, with or without literary aspirations" (SWALES 1990: 33). Although a wide range of interpretations may be found, the main goal of genre identification is to identify groups of texts that share a common form of transmission, purpose and discourse properties[2]. Basically, genre categorization can be applied to most forms of communication, and genre information unconsciously models

[1] Genre is an ancient literary concept that goes back to Aristotle. Traditional definitions of genre focus on textual regularities. In traditional literary studies, genres, like sonnet, tragedy, ode, etc., are defined by conventions of form and content.

[2] A genre is a class of communicative events: the main criterion that turns a collection of communicative events into a genre is represented by some shared set of communicative purposes (SWALES 1990: 45).

one of the most distinguishing features in information searches. While a widely accepted interpretation of genre refers to culturally-established categories of texts, such as novels, letters, manuals, etc., automatic genre categorization is based on a quantitative approach, leveraging on extractable and computable features (i.e. observable properties in a text) to discriminate among different classes of documents. More recently, the computer-Internet combination has resulted in the emergence of web genres, and the triple <content, form, functionality> has been suggested to characterise them more conveniently (SHEPHERD AND WATTERS 1999). This triple has become a quadruple <content, form, functionality, positioning> when applied to the framework for identifying genre characteristics and to describe the current Swedish online newspaper genre (IHLSTRÖM AND ÅKESSON 2004).

Since the publication of Biber's work on the linguistic variation across speech and writing (BIBER 1988), a traditional term *text types* has entered corpus linguistics. Text types traditionally refer to the four rhetorical categories of *narrative, description, exposition* and *argumentation*. This rhetorical partition has been adopted by BEAUGRANDE AND DRESSLER (1981)[3] and WERLICH (1976)[4]. Biber's text types, instead, derive from the linguistic features found in the texts themselves[5]. His work is by now a classic of the data-driven approach, based on multivariate statistics, and has influenced also European standards for large language resources, such as the EAGLES guidelines on text typology (EAGLES 1996). Biber (1988) extracted 6 factors/dimensions from the LOB Corpus and from the London-Lund Corpus, and labelled the resulting 10 text types with new descriptive (and rather subjective) labels, such as *Informational vs. Involved Production*, *Overt Expression of Persuation*, etc[6]. Biber's claim was to make a distinction between genres and text types. While genre is influenced by cultural and external criteria, text types elicit from the internal linguistic nature of the texts themselves, irrespective of their genre classifications. But this clear-cut distinction has

[3] To the traditional text types descriptive, narrative, argumentative, they add literary texts, poetic texts, scientific texts, didactic texts. Conversation seems to play a role apart. They state that these additional types of text "contain various constellations of description, narration, argumentation" (BEAUGRANDE AND DRESSLER 1981: 185).

[4] WERLICH (1976: 39) lists five text types: description, narration, exposition, argumentation and instruction.

[5] While external criteria follow distinctions and classifications that are already available in the culture, Biber establishes a typology of texts based on internal linguistic criteria only, and then interprets the results with reference to external "functions". Biber's internal criteria are taken from published studies of language variation.

[6] Unexpectedly, BIBER (1989) presents "a typology of texts in English with respect to a five-dimensional model of variation. […] Eight text types are identified with respect to these dimensions". Some pages later, he states: "To date, five major dimensions of variation have been identified in English. Biber (1988) presents a unified description of genre variation in English in terms of this five-dimensional model." Where has the 6th dimension disappeared? It was still alive and well in the last chapter of BIBER (1988): "The present studied has identified six dimensions of variation among texts in English, […]. Biber (forthcoming) uses a cluster analysis to develop a typology of English texts in terms of these same dimensions. […] In all, eight text types are identified." BIBER (1988: 207).

proved too artificial and counter-intuitive[7]: the boundaries between the two terms remain fuzzy and blurred, and in recent corpus-analysis literature, "genre" and "text types" are used most of the time as synonyms. For example, STUBBS (1998: 11) and, WALLER (1987: CHAPTER 9) use the two terms interchangeably, similarly to with many other linguists and researchers. Even in the research field, where Biber (1988)'s approach has been followed extensively and has inspired a number of interesting projects, the distinction between the two terms is not applied. And this is not all: other terms, such as "style", or "functional roles", have come into play to make the maze even more intricate.

Biber's terminology is extremely tricky. In BIBER (1988: 68) he states: "I use the term "genre" to refer to text categorization made on the basis of external criteria relating to author/speaker purpose. The genre categories in the present study are adopted from the distinctions used in the corpora." A couple of pages later, he describes his use of the term text type: "I used the term "text types" to refer to groupings of texts that are similar with respect to their linguistic form, irrespective of genre categories. For example, a science fiction text represents a genre of fiction, but it might represent an abstract and technical text type (in terms of linguistic form), similar to some types of academic exposition and different from most other fictional texts. In a fully developed typology of texts, genres and text types must be distinguished, and the relations among them identified and explained." (BIBER 1988: 70). Some years later, he realizes that the distinction between register and genre tends to be quite abstract and vague (BIBER 1995). Therefore he makes the decision to use the term register "as the general cover term associated with all aspects of variation in use" (BIBER 1995: 9). In contrast, the term text type is consistent with his previous work and refers to text categories defined in strictly linguistic terms. Consistently with this definition of register, in BIBER ET AL. (1999) we read that "Register distinctions are defined in non-linguistic terms, with respect to situational characteristics such as mode, interactiveness, domain, communicative purpose, and topic" (BIBER ET AL. 1999: 15). Unfortunately, a few lines below, we get a little bit confused, because register and text types are used as synonyms: "In many cases, registers are institutionalized varieties or text types within a culture, such as novels, letters, editorials, sermons, and debates" (IBIDEM).

LEE (2001) makes an remarkable effort to shed some light and suggests interpretations for the long list of terms (*genres, registers, text types, domains, styles*) that appear to be used sometimes subjectively, sometimes interchangeably and synonymically by the researchers involved in computational and corpus-based projects.

LEE (2001) gives interesting opinions on the use of these terms. He states that as text types cannot yet be established explicitly in terms of linguistic features (Biber's taxonomy has not taken ground among linguists), perhaps the looser sense of the term text types could be handier: *text types* can be used in the sense of the traditional four-part rhetorical categories of narrative, description, exposition and argumentation. He suggests that *style* has essentially to do with an individual's use of language, while the two terms, *register* and *genre* should be seen as two different points of view covering the

[7] One objection is that this classification of texts based purely on internal criteria does not give prominence to the sociological environment of the text, thus obscuring the relationship between the linguistic and non-linguistic criteria (EAGLES 1996).

same ground: register and genre are in essence two different ways of looking at the same object. Register is used when we view a text as language: as the instantiation of a conventionalized, functional configuration of language tied to certain broad social situations (a variety according to use). Genre is used when we view the text as a member of a category: a culturally recognized artefact belonging to a grouping of texts according to some conventionally recognised criteria. Thus we talk about the existence of a "legal register", but the instantiations will be the genres of "will", "testaments", and so forth. We talk about a "formal register", where "official documents" and "academic prose" are possible exemplar genres. In contrast, there is no literary register, but rather, there are "literary styles" and "literary genres", because the essence of literature is creativity and originality. Genres are instantiations of registers (each genre may invoke more than one register). Genres come and go, or change, being cultural constructs which vary with the time in a society. Confusion comes from the fact that often we label with the same word both language (register or style) and category (genre), as in the case of "conversation".

Regardless this effort of clarification, the overlapping and the boundary fuzziness between those concepts are still extensive. In quantitative linguistics research, for example, KARLGREN (2000) swings between *genre, text types* and *functional style* and uses also terms such as *stylistic experiments, stylistic analysis, stylistics*. STAMATATOS ET AL. (2001) state that "text *genre* detection concerns the identification of the kind (or *functional style*) of the text". DEWDNEY ET AL. (2001) use the wording "format style, i.e. 'genre' ". JOHANNESSON AND WALLSTRÖM (1999) use the term *stylistic genres,* for KLAVANS AND KAN (1998) *text types* and *genres* are synonyms, etc. In a few words, we could say that all research projects after BIBER (1988)'s experiment use *genre* as an umbrella term to define what in a text is not *topic*. Rather, *genre* is proposed as a kind of assisting concept complementary to topic (WOLTERS AND KIRSTEN 1999).

## 3 *Automatic Genre Identification of Digital Documents*

Several studies have been carried out on automatic genre identification using quantitative-statistical techniques. Most of them admit being indebted to Biber's work (mainly BIBER 1988, 1989, 1995), but they implement a wide range of variants. The list of research projects below collects the major achievements after BIBER (1988) and follows an ascending chronological order that allows a better understanding of the evolution in the field.

NAKAMURA (1993) applies a statistical procedure known as Hayashi's Quantification Method Type III (HT3) to the textual data stored in the Bank of English[8] and works on specific verbs, grouping them according to their semantics. He shows how the various sub-corpora of the Bank of English are related to one another on the basis of the distribution of these verbs and how the verbs are related to one another in the databank. The categories used as input to the statistical method are three semantic classes of verbs[9]. The raw frequencies are adjusted to the corpus size (about 10 million

[8] For a description of this corpus, see NAKAMURA (1993: 295).

[9] For the list of verbs used by the author, see NAKAMURA (1993: 297-300).

words) and the normalised figures are submitted to HT3, which allows the quantification of qualitative categories and samples simultaneously. The distinctive feature of the method is that it can classify or quantify both categories and samples, thus evaluating responses qualitatively, i.e. assessing whether each sample reacts positively or negatively to several categories. HT3 offers the added advantage of showing the relationships among linguistic categories, and the relationships among texts, because it enables the comparison of the two relations directly and explains the distribution of texts in terms of the distribution of linguistic categories and vice versa. This statistical method assumes that there is mutual dependency between types of texts and the linguistic categories used in them. This assumption contrasts with factor analysis used by Biber, where inter-dependency among variables is attained as a result.

The results of this study have been plotted into charts, which graphically show patterns of distributions. A first chart displays the distribution along two axes of the four corpora used in the study, and a second chart shows the distribution of the verbs. Interestingly, the patterns of distribution relating to corpora are reflected in the distribution of verbs, and vice versa. For example, the location of the Spoken Corpus in the first chart coincides more or less with the location of the verbs "mean", "think", and "know" (which are extremely frequent in this corpus) in the second chart. Specific charts have been plotted for public, private, and suasive verbs.

WIKBERG (1993) studies the connection between verbs and text categories, as well as the use of verbs as style markers within the tagged LOB corpus. He uses two common tools, WordCruncher[10] and TACT[11] in his study. His detailed analysis and commentary is restricted to "verbs of communication", contact, hedging, and cognition. He provides specific evidence on the role that verbs play in different text categories and claims that a corpus of one million words is enough to group specific classes of lexical verbs.

MICHOS ET AL. (1996) present an approach to categorization of texts in terms of functional style. They suggest a three-level description of functional style, which includes: 5 basic style categories (public affairs, everyday communication, scientific, journalistic, literary styles), their stylistic features (formality, elegance, syntactic complexity, verbal complexity), and style markers for the identification of those features (such as idiomatic expressions, formal words, poetic words, abbreviations, scientific terminology, etc.). Their model includes a morphological analyser, a syntactic parser, a stylistic analyser and a final evaluation/estimation step. The classification results are based on 5 text samples in Modern Greek, belonging to the 5 style categories mentioned above. The model could identify correctly 3 categories. They claim that the model is language independent, and suitable for a range of application. But the findings shown in this article seems to be very preliminary.

TAKAHASHI 1997 attempts to detect text typology in the LOB corpus using an extended version of HT3, the Extended Hayashi's Quantification Method Type III

[10] WordCruncher is a tool for creating and analyzing electronic textbases. For further information, visit http://www.itl.nist.gov/div898/handbook/pmc/section5/pmc5.htm

[11] TACT (Text Analysis Computing Tools) is designed to do text-retrieval and analysis on literary works. For further details, visit http://www.chass.utoronto.ca/cch/tact.html

(EHT3). The main goal in Takahashi's research is the same as Biber's, i.e. the identification of hidden constructs through factors, but he follows NAKAMURA (1993)'s tradition to unveil them. EHT3 is supposed to make the interpretation easier because the relationships between features and categories along each axis of coordinates should facilitate the interpretation of the factors. Results are not extremely encouraging and sometime weird. For example, the author states " 'singular proper nouns' and 'plural proper nouns' are located separately in axis 1, but it is impossible to tell the structural difference between them in a sentence" (TAKAHASHI 1997: 33). Probably, this approach needs further investigations. In his evaluation, where he tries an interpretation of Axis 1 and Axis 2, he himself states "it seems to me that this approach does not allow me to describe the characteristics of each dimension properly" (TAKAHASHI 1997: 22) and he suggests a "macroscopic interpretation", which does not seem extremely convincing.

SIGLEY (1997) uses Principal Component Analysis (PCA) to re-interpret pre-existing text categories in corpora for the analysis of linguistic variation. He observes that, if one wishes to explain observed patterns and predict new ones, it is not safe to use pre-existing text categories as natural groups, rather more consistent text groupings should be created. His research is based on the claim that English relativiser choice varies with formality. For example, the relative 'that' and especially null relatives are more characteristic of informal language, while wh-relativiser are more characteristic of formal language. He builds an index of formality to test relativiser choice. His research is confirmatory, not exploratory, and aims at explaining behaviours rather than describing texts. He uses three corpora of New Zealand English and applies PCA on simple wordform-based counts to create the formality index. Internal consistency of the major categories was evaluated by non-parametric analysis of variance (Mann-Whitney U or Kruskal-Wallis H). Where this test returned a significant result, individual Mann-Whitney tests were also carried out comparing each subcategory in turn with the reminder of its category. One interesting finding from this study is that there is enormous variation in academic texts according to subject fields, from the most "formal" writings on Law, History and Natural Science to fields more characterised by abstraction and argument, such as Philosophy, Education, Linguistics, and so on. He also tries to evaluate text categories in purely linguistic terms, independently from pre-existing categories, using cluster analysis. He illustrates the cluster analysis of Commentary, which reveals a situational difference with strong linguistic effects. However, he realizes that for two-dimensional comparison, cluster analysis offers little advantage over straightforward inspection. In conclusion, he acknowledges that his approach is time-consuming (it requires manual disambiguation of selected word-forms), and that PCA makes the interpretation of factors extremely difficult, but in the end as formality is a general construct subsuming many situational dimensions, he claims that it is worthwhile to spend time in refining the index.

KESSLER ET AL. (1997) acknowledge that genre is a necessarily heterogeneous classificatory principle. They propose the theory of genres as bundles of generic facets, which correlate with surface cues. Generic cues are the observable properties of a text that are associated with facets. Cues can be structural, lexical, character-level and derivative. A facet is simply a property which distinguishes a class of texts that answers to certain practical interests, and which is associated with a characteristics set of computable structural or linguistic properties, the 'generic cues'. In principle, a given

text can be described in terms of an indefinitely large number of facets[12]. Unfortunately, this clean theorical framework becomes more obscure when they put their categorization scheme into practice. They manually classify the Brown corpus according to their scheme, and eliminate texts that do not fall unequivocally into one of their categories. They end up using 499 out of the 802 original texts in the corpus. For their experiments, they analyze the texts in terms of three categorial facets: *brow, narrative, genre*. Brow characterizes a text with respect to the intellectual background required to understand the text itself. A brow facet has several degrees, it can be popular, middle, upper-middle and high. Narrative is a binary facet and decides whether a text is written in a narrative style. Genre classifies a text either as reportage, editorial, scitech, legal, non-fiction, or fiction. A set of 55 lexical, character-level and derivative cues were used to describe the documents. Their corpus of 499 texts was divided into a training subcorpus (402) and an evaluation subcorpus (97). They use Logistic Regression (LR) and neural networks. Results are exploratory. They evaluate their approach against a baseline and compare the results achieved with their generic cues (lexical, character-level and derivative cues) against the results achieved by of KARLGREN AND CUTTING (1994) with "structural cues" (KESSLER ET AL. (1997) call KARLGREN AND CUTTING (1994)'s features "structural cues" because some of them are POS tags frequencies. It is interesting to see that what is "simple metrics" for KARLGREN AND CUTTING (1994), becomes difficult for KESSLER ET AL. (1997), just because a tagger is required). As KESSLER ET AL. (1997) themselves state, their results are only fair. For the facet level Genre, which lists six categories, they set a baseline[13] of 33% and the best accuracy they achieve is 79% (i.e. 46% more than the baseline) with a selected set of features (the most discriminating) and with a 3-layer perceptron. As for the comparison with the performance using KARLGREN AND CUTTING (1994)'s cues, the best accuracy achieved on the Genre facet is 66%. They report other accuracy results, but these are based on binary level classification, so they are less interesting because they do not give an full picture of the classification performance. Overall, on the theoretical side, it is intriguing the idea of considering genres in "compositional" terms, as bundles of facets, but the categorial facets of *brow, narrative* and *genre* are not satisfactorily justified, and add more candidates into an already confused terminology and classification schemes. Their claim that there is only a small difference between surface and structural cues has proved to be rather subjective, because subsequent research has proved the contrary (for example, cf. STAMATATOS ET AL. (2001)).

KARLGREN (2000) includes a collection of several experiments (published individually from 1994 to 2000) on genre identification and describes the use of stylistics for information retrieval tasks. His aim is not to get an accurate genre classification, but he wishes to find out a practical application for information retrieval, that could help if content-based methods select a too large set of documents (KARLGREN 1994). One of the most important experiments is a simple method for categorizing texts into pre-determined text genre categories using the statistical standard technique of

[12] For example a newspaper story could be seen an example of *narrative* as opposed to *directive* (as in a manual), or *suasive* (as in an editorial), or *descriptive* (as in a market survey).

[13] Their baseline is based on the guess of the most populated category: NONFICT for Genre, MIDDLE for Brow, NO for Narrative.

discriminant analysis applied to the documents included in the Brown corpus (KARLGREN AND CUTTING 1994, KARLGREN 2000: 49 FF.). In this experiment, discriminant analysis is used to classify documents into 2 (Experiment 1), 4 (Experiment 2) and 10-15 (Experiment 3) text categories, according to 20 features, which are "easy to compute assuming we have a part of speech tagger" (KARLGREN AND CUTTING 1994). From their results, it emerges that the error rate increases with the number of the categories tested in the corpus. This fact may be a consequence of how genres are defined in the Brown corpus. Here an old issue emerges: can we classify texts into externally (i.e. culturally) defined categories, such as genres, using internal (i.e. linguistic and stylistic) parameters? This is still an open question, and we rely on future work to get a reliable answer.

In other experiments (KARLGREN 1999 AND KARLGREN 1996B), stylistic variations are used to weed out non-relevant documents from the TREC material. TREC documents come from various sources, with different stylistic preferences, and varying usefulness for any given topic. In these experiments, several types of simple stylistic items are measured and combined using non-parametric multivariate techniques, such as decision tree learning techniques. These experiments also show that stylistically determined genres, or functional styles, are different as regards potential usefulness for the queries tested, and that the distinctions between relevant and non-relevant differ between genres.

STAMATATOS ET AL. (2000) present a method for detecting the text genre quickly and easily following an approach originally proposed in authorship attribution studies, which uses as style markers the frequencies of occurrence of the most frequent words. They claim that the most frequent words of the BNC represent the most frequent words of written English and are reliable discriminators among genres. They try this approach on the Wall Street Journal Corpus. They compare their results with Burrows[14]' results based on the 55 most frequent words in their training corpus. STAMATATOS ET AL. (2000)'s approach is based on the 30 most frequent words of the BNC and returns a 2.5 error rate, while Borrows' returns 6.25 error rate. Additionally, they note that the frequencies of occurrence of the most common punctuation marks play an important role in accurate text categorization. The combination of most frequent words plus punctuation marks achieves >97% accuracy.

TyPText (a project financed by ELRA) develops a methodology and a toolkit aiming at testing and extending Biber's work using the French language (ILLOUZ ET AL. 2000, FOLCH ET AL. 2000). This approach aims at classifying a new text with respect to already-formed groups. It is an "inductive" approach to text typology ("une demarche typologique inductive")[15]. One of the main interests of the project is 'text profiling' ("profilage de textes"), i.e. the 'calibration' of a corpus based on a quantitative evaluation of vocabulary and grammatical categories (morpho-syntax, syntax, semantics, etc.). The

[14] Burrows, J. (1987), Word-patterns and Story-shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2(2), pp. 61-70.

[15] "L'optique, inductive, dans laquelle nous nous inscrivons consiste a faire émerger *a posteriori* les types de texts – considérés comme des agglomérats fonctionnellement cohérents de traits linguistiques – grace à un traitement statistique multidimensionnel des textes annotés. Cette optique constitue la ligne directrice des travaux de D. Biber (Biber 1988, Biber 1995)."(HABERT ET AL. 2000).

aim of calibration is to produce measures of homogeneity within the different parts of a corpus in terms of one or more parameters. Corpus profiling (ILLOUZ ET AL. 1999) is also necessary to profile texts in order to check the correspondence between the language uses they represent, and the ones that the NLP tools being developed are intended to tackle. TypTex has been tested successfully on four architectures for text processing: TIPSTER, GATE, IMS, LT XML (ILLOUZ ET AL. 2000).

RAUBER AND MÜLLER-KÖGLER (2001) present a way to provide automatic grouping of documents based on text structure. Their include a combination of various surface level features of texts, such as word statistics, punctuation information, the occurrence of special characters and keywords, as well as mark-up tags capturing image, equation, hyperlink and similar information. More specifically, they use surface level cues and distinguish among four distinct types of features: text complexity information and text statistics, special character and punctuation counts, characteristic keywords, and format-specific mark-ups (about 200 different features). Based on these structural descriptions of documents, the self-organizing map (SOM), an unsupervised neural network, is used to cluster documents according to their structural similarities. Information is incorporated into SOMLib digital library system, which provides an automatic, topic-based organization of documents according to their content. By clustering documents according to the similarity conveyed by their structural features, the types of documents, or genres, should be revealed. Various series of experiments have been performed in different settings, including technical documents and website analysis. While the content-based SOM provides an organization of articles by their subject, the genre SOM analyses the structural features of the documents and groups the documents accordingly. Instead of assigning every unit to a specific genre label, SOM genres are mapped into an RGB-colour space. The units in between are automatically assigned to intermediate colours. Interestingly, this study present a visual approach to genre variation and mixture: even though the actual genre of a document cannot be intuitively told by the colour in which it is represented, this coloured approach allows gradual changes between various genres. As the colour metaphor is not intuitive, the evaluation of their results is based on the feedback of users. The outcome of the evaluation seems to be encouraging because most people had a feeling of what to expect from a document in a specific colour after visiting a few documents on their respective areas of interest.

BAGDANOV AND WORRING. (2001A AND 2001B) approach the general problem of machine-printed document genre classification using content-free layout structure analysis. The genre of a document is determined from the layout structure conveyed from scanned binary images of the document pages, using no OCR and minimal a priori knowledge of the logical structure of the document. Their approach uses attributed relational graphs to represent the layout structure of document instances, and a first-order random graph (FORG) to represent document genres. The test data used in all experiments is a set of 150 documents sampled from the Océ Competitive Business Archive. The collection contains sample documents from trade journals and product brochures. The sample consists of five genres, two of which have four subgenres. For their experiment they collapse the subgenres into top-level genre categories, creating a total of 11 fine-grained genres. Each document has been scanned as a binary image file at 300 dpi. A variety of statistical classification techniques were

evaluated in order to compare the effectiveness of genre classification by first order random graphs with traditional techniques. For the statistical, feature-based classifiers evaluated, global page-level features were extracted from the first page of each document. The classification accuracy for each method was estimated using leave-one-out cross-validation. On the collection sample, the FORG classifier significantly outperforms purely statistical classification methods.

# *4 Automatic Genre Identification of Web Documents*

STAMATATOS ET AL. (2001) present an approach to text categorization in terms of stylistically homogeneous categories, text genres and authors, in Modern Greek, and build an *ad hoc* corpus of Web texts downloaded from various Greek Web sites, belonging to 10 different genres. They do not include any Web-specific genres, but only traditional genres such as "Press Editorial", "Press Reportage", "Academic Prose", "Literature", etc. One of their main interests is to demonstrate the advantage of using complex linguistic features derived from NLP tools. In fact, in contrast to previous stylometric approaches, they attempt to take full advantage of existing NLP tools, exploiting the sophisticated linguistic knowledge they output. The entire corpus is analyzed by the SCBD (an existing NLP tool for Greek), which automatically provides a vector of 22 parameters, or style markers, including information derived from chunking and parsing. The corpus is divided into a training set and a test set of equal size. The vectors of the training corpus have been used to extract the classification model using multiple regression and discriminant analysis. These classification models were then applied to the vectors of the test corpus for cross-validating their performance on unseen cases. In order to evaluate their approach, they set a baseline using two previous stylometric approaches that are based on distributional lexical measures: a multivariate model of function of vocabulary richness[16], and the frequency of occurrence of the most frequent words[17]. On average, both discriminant analysis and logistic regression perform well. Comparatively, the performance of vocabulary richness is quite poor, and the frequencies of the most frequent words are less accurate than those achieved with their method. They claim that their methodology outperforms existing lexically based methods. What they suggest in the end is a combination of their approach with lexically based methods in order to get a very reliable text categorization system.

TyPWEB[18] (ongoing) is an extension of the approach developed for TyPText (see previous section). TyPWEB (BEAUDOUIN ET AL. 2001A, BEAUDOUIN ET AL. 2001B, ILLOUZ AND HABERT 2002) aims at adapting the TyPTex architecture to the processing of Web sites. The aim is to provide a methological and practical framework for website profiling and the development of a fine-grained typology of these sites. The approach consists in characterising each site by a set of markers concerning both content and

---

[16] Holmes D. (1992), "A stylometric analysis of Mormon scripture and related texts, *Juornal of the Royal Statistical Society, Series A*, 155(1), pp. 91-120

[17] Burrows J. (1992), "Not unless you ask nicely: The interpretative nexus between analysis and information", *Literary and Linguistic Computing*, 7(2), pp. 91-109.

[18] **TyPWEB –** TYpologie et Profilage de sites WEB (LOT 1), available at: http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/typWeb.htm

structure. This project aims at exploiting textual, structural and presentational features in order to identify the typology of personal and commercial websites. Corpora containing commercial websites and personal home pages have been created for the project. Websites have been converted into XML (the conversion to the XML format can help overcome all the inconsistencies of the bad quality of the HTML pages crawled from the Web) and statistics has been computed. For instance, they count words, HTML tags, etc. Multivariate statistics and NLP tools, like taggers and parsers, have also been used. The goal of the project is to extract text typologies, on the basis of observed correlations among grammatical and lexical markers, textual and hypertextual structure and the multimedia aspect[19]. So far, they have accomplished two experiments on personal pronouns and link structures. More specifically, they analyse the distribution of personal pronouns, links, grammatical words, the lexicon of the host and lexical opposition (French words vs. English words). Their findings show that commercial websites have a more complex structure and more links. The qualitative exploration of personal and commercial websites confirms these stylistic differences. The final aim of the project is not to detect predefined styles (narrative, descriptive, argumentative, poetic and so on), but to group (cluster) documents or parts of documents, according to their use of linguistic features, some lexical tags and specific vocabulary, for example, different types of semantic adverbs, such as negation, possibility, time and place adverbs and subordinating conjunctions.

## *4.1 Genre-based Classification for Information Retrieval*

Automated recognition of genre can dramatically leverage information seeking on the Web, thus addressing the real and growing problem of information overload. Automatic Web genre identification is one of the key factors for improving the often inadequate results of search engines, as the user would be able to specify the desired Web genre along with a set of keywords. One common limitation of Web genre detection projects is the restricted approach to Web genres. This means that researchers in the field focus only on a very limited number of genres defined *a priori*. For instance BEAUDOUIN ET AL. (2001) mainly aim at discriminating between personal and commercial Web sites, ROUSSINOV ET AL. (2001) try to single out only five major groups of documents genres that might be used in a interactive search tool that would allow genre-based navigation, REHM (2002) concentrates only on Academic's Personal Homepages, and so forth.

Genres are useful because they make communications more easily recognizable and understandable by recipients. Researchers in the area of hypermedia and Web design have noted that user orientation and navigation is contingent on the user's perception of genre (DILLON AND VAUGHAN 1997). An established genre provides expectations for its readers, because it carries with it a whole set of prescriptions and restrictions. YATES AND SUMNER (1997) claim that a "new burden for providing fixity" in the unstable Web communications world is being met by increased reliance on genre.

[19]"Il s'agit plutot de faire émérger, de manière inductive, des typologies sur la base des corrélations observée entre des indicaterus portant sur l'outillage grammatical et de lexique, sur la structurations textuelle et hypertextuelle, et sur l'aspect multimédia. La construction de ces typologies se nourrit des entretiens menés auprès des concepteurs de sites marchands et de sites personnels comme de l'observation fine de tels sites." (BEAUDOUIN ET AL. 2001A).

Indeed, the concept of genre is proving its value as an analytical tool in the whole Information System research (SCHMID-ISLER 1997, BROWN AND SMEATON 1998).

Not only retrieval tools could be improved if genre is one of the selection criteria that users can exploit. Also interfaces can take huge advantage from the use of genre. BRETAN ET AL. (1998), for instance, propose a richer representation of retrieval results in the search interface by focussing on the notion of document grouping, using both stylistic genre-based document categorisation and statistical content-based clustering.

ARGAMON ET AL. (1998) are interested in categorizing texts according to style for information filtering. For example, they would like to determine if a document is promotional or informative, whether it is written by a native English speaker or not, and so on. They try to measure the discriminating power of two different types of features, function words and POS trigrams (which they call pseudo-syntactic). The rationale behind the use of the frequencies of function words (a method used in stylometrics) is that presumably it is not driven by content and hence might be expected to remain invariant for a given author over different topics. They expect such invariance within classes of similar documents. The rationale behind the use of POS trigrams, instead, is that POS trigrams are large enough to encode useful syntactic information, and small enough to be computationally manageable. They use a machine learning technique, the Ripper classifier, with pairwise distinguishability and 5-fold cross-validation. They realize that other learning algorithm could be more suitable for style-classification tasks (for example, Winnow or Naïve Bayes) and other classes of features could also be tried. No evaluation is provided but only accuracy results. The best accuracy rate is achieved by the combined use of functional words and POS trigrams.

Similarly to ARGAMON ET AL. (1998), also JOHANNESSON AND WALLSTRÖM (1999) are interested in information filtering. They present a method that both visualizes and determines similarity between stylistic genres, both in Swedish and in English. The novelty in their approach is that they aim at making comparisons and detecting similarities among different genres. Even though they see genres as open classes constantly evolving, they believe that it is meaningful to analyze and define genres as if they were static entities. They choose to implement only 29 lexically-based linguistic features out of the 67 linguistic features used by BIBER (1988), with the debatable claim that when genre categorization is oriented towards Information Retrieval one cannot assume that text can be tagged or parsed. In order to evaluate their method, they compare two approaches, one based on linear algebra, the other on Kohonen self-organizing maps. With the first method, they want to find out how well a certain genre correlates to a predefined genre. They use three genres as a base for the comparison: FAQs, news articles, research papers, in Swedish and English, and they use five representative documents (downloaded from the Internet) for each genre to build the templates that will serve for the comparison. Then they calculate a template vector. Next they select 10 representative documents (downloaded from the Internet), different from those used for the templates, to be compared with the template vector. They use the cosine of the angle (a linear algebraic method) to compare the template vector with the other vectors. The second method they use is the Kohonen self-organizing map, which does not need any predefined genres or group of documents. A Kohonen self-organizing map is a two-dimensional neural net consisting of a number of nodes linked in a 2-dimensional structure much like a regular map. Each of the nodes has a template

vector of the same dimensionality as the vectors the map is supposed to organize. When the Kohonen map is finished, it sorts every document into the node that is most similar to its stylistic vector. The documents from this experiments were taken from 10 diverse categories. These two methods have been applied to English and Swedish documents separately. Results from the first method show that both news articles and research papers are very similar, while FAQs are more easily distinguishable from the other two genres. Results from the second method show that Kohonen self-organizing map separates very well among the 10 categories, using the 29 features. Findings appear to be extremely preliminary, because the number of documents used is very low.

ROUSSINOV ET AL. (2001) are involved in an ongoing study focused on the use of Web page genres to facilitate information exploration. Their study includes three phases. The first phase is the identification of which genres most/least frequently meet searchers' information needs. They identify 116 different genres (74 are included in CROWSTON AND WILLIAMS (1997)). The second phase is the identification of five major groups of document genres that might be used in an interactive search tool that would allow genre-based navigation. Among the major genres included in the five groups are: home pages, articles, product lists, instructional materials, FAQs, etc. The third phase is focused on the development of a novel user interface for Web searches allowing genre-based navigation. Their initial results suggest that certain genres are better suited for certain types of needs. A second conclusion is that the users do not always agree on a genre, that's why they suggest that an interface supporting fuzzy genre definitions will be more profitable. The study is still in its early stages and the preliminary results have still numerous limitations that can be addressed in future, such as the size of the sample, which is too small.

DEWDNEY ET AL. (2001) investigate text classification in terms of "format style". They claim that genre complements topic classification and can significantly improve retrieval of information. The genre of a document here is a label, which denotes a set of conventions relating to the way in which information is presented. These conventions cover both formatting and style. Their work investigates the use of two different feature sets: a set based on words (traditional bag-of-words technique, weighted with TF*IDF and feature reduction achieved by applying the Information Gain algorithm), and a set of features that reflect the way in which the text is presented, which varies from linguistic features (such as frequency of adjectives, past tense verbs, sentence complexity, etc), to layout features such as line-spacing, tabulation, mark-up tags, etc. They used 323 word features and 89 presentational features. The corpus is provided by CMU (Carnegie Mellon University) and includes seven genres. They compare the performances of Naïve Bayes, C4.5 and SVM classifiers. Presentational feature set yields a significant advantage over use of word frequencies, except when using Naïve Bayes. The best accuracy was achieved by combined features using SVM (92.1%). Overall, results show that the use of combined feature sets gives good classification accuracy in sorting seven genres. More interestingly, these sets of experiments also show that presentational features alone, when used with a suitable classifier, are discriminating enough for genre classification without the need for word features.

The HYPPIA system[20] is a web service that will monitor websites for news articles, classify them using a number of classifiers and create personalized digital libraries from the content. New classifiers can be added as the system grows. This system is currently under development and preliminary results are presented in FINN ET AL. (2001). The core of the system is the article database. A set of classification agents classifies articles in the database. Each classifier is responsible for a particular class. The opinion classifier recognizes articles that express the opinion of the author. These articles contrast with those which report facts. When developing the opinion classifier they exclude domain specific features in order to easily migrate the classifier to new domains. They work exclusively on text, excluding other elements of web pages, such as graphics. In fact, they extract the body of the article and base the classifier on features that occur within this area. The reason for this choice is that text from outside the main body often misleads classifiers. As a baseline of their opinion classifier, they use the results given by the occurrence of words in a text as features and a Naïve Bayes classifier. Their second approach examines the type of language in the document. Intuitively, they expect that the kind of language used in opinion documents will be different from factual articles. They process documents using Brill's POS tagger, and then represent a document as the fraction of words for each POS. Then they use C4.5. In their experiments, they use two domains, football and politics, in order to evaluate accuracy as well as domain transfer. A corpus of documents was spidered from the web and then manually classified as being opinion or non-opinion. Their corpus includes 350 football articles and 230 politics articles. Their experiment uses tenfold cross-validation. Results show that the classifier based on POS statistics (C4.5) performs better than the classifier based on word occurrence statistics (Naive Bayes) in all cases. They conclude that the "kind of language" (that they will call "genre" in the papers below) used in a document is a better indication of subjectivity than the content of the document. Moreover, they conclude that POS approach is better suited to generalize to new domains.

Also the two following papers have been written within HYPPIA project. DIMITROVA ET AL. (2001) describe how shallow text classification techniques can be used to sort the documents returned by web search engines according to genre dimensions, such as: the degree of expertise assumed by the document (EXPERTISE DIMENSION); the amount of detail presented (DETAIL DIMENSION); whether the document reports mainly facts or opinions (SUBJECTIVITY DIMENSION). They claim that their taxonomy is consistent with other works in this area, but their prototype uses genre-labels, such as "brief.technical", "detailed.technical", "brief. semi-technical", "detailed.non-technical", which have not been used before. The features they use vary according to the genre dimension they want to extract. For the EXPERTISE DIMENSION, they rely on: frequency and length of words in a document, frequency of technical HTML tags, such as <SUB> and <SUP>. The choice of these tags is motivated by evidence that word length and frequency determine the complexity of cognitive processing and sentence structure. Features to extract DETAIL DIMENSION are the physical size of a document, the number of lines of rendered HTML, and the frequency of long words in the document. For SUBJECTIVITY DIMENSION they seem to rely mainly on the distribution of POS tags. They use machine learning techniques to

[20] The homepage of this project is at: http://www.smi.ucd.ie/hyppia/

achieve their classification, but they do not mention any specific algorithm. Their final goal is to create a Web Genre Visualizer, i.e. a user interface that replaces conventional ranked document lists with a graphical, genre-oriented depiction of the retrieved documents. The current prototype allows user to visualize the EXPERTISE and DETAIL genre dimensions. SUBJECTIVITY and other dimensions will be incorporated in future.

FINN ET AL. (2002) identify document genre as an important factor in retrieving useful documents and focus on the novel document genre dimension of subjectivity. They focus on information filtering services for the personalised retrieval of online news articles. The genre class they investigate is whether a document presents the opinion of its author or reports facts. This is a common distinction in newspaper articles and other media. For example, articles of genre class *fact* may be reporting the latest stock prices. Articles of genre *opinion* may give the opinions of various financial analysts. They investigate three approaches to automatically classifying documents by genre: traditional bag of words technique (and this document representation has been used as baseline), POS statistics (36 POS features, expressed as a percentage of the total number of word in a document), hand-crafted shallow linguistic features (76 features including stop-words, counts of punctuation, average sentence length, number of long words, etc.). Their experiments demonstrate that the POS approach is better than the traditional bag of words approach, particularly under the domain transfer conditions. By measuring domain transfer, they aim at identifying feature sets that generalize well to unseen subject domains. Their experiments use documents from three domains (football, politics, finance). Their corpus of documents has been spidered from the Web and each document has been manually classified as being either opinion or fact. They use C4.5 as classification tool and tenfold cross-validation. They evaluate the three feature sets in a single domain and in domain transfer. In the single domain experiments (each domain, football, politics, finance, is taken separately) the hand-crafted features perform better. In the domain transfer (where the classifier is trained on documents belonging to a domain, and tested on documents belonging to another domain), the POS features perform significantly better than bag-of-word features, which means that a model built using frequencies of words is more closely tied to the document collection used for training, and its generalization power is more restricted.

With regard to the heterogeneous diversity of Web pages, REHM (2002) considers all-genre inclusive approaches rather coarse and incomplete when considering the set of distinct features that constitutes a certain genre with respect to a group of genres. Therefore, he proposes to concentrate only on one genre, i.e. Academic's Personal Homepages: a restricted domain, but which is broad enough to identify a Web genre hierarchy. A corpus of 3,000,000 Web pages from German universities is being constructed. He proposes an approach that is predominantly founded on the feature-based analysis of the HTML structure of a document, or group of documents. His list of features includes: metadata, HTML structure, document-spanning features, linguistic and structural cues, language issues. The overall goal is not only the automatic identification of Web genres, but also the extraction of the content contained in genres modules and its integration into a structured XML document. A central point of REHM (2002)'s approach is the breaking of physical document boundaries, leveraging on the knowledge about the typical hypertextual structure. He assumes that generalized Web genre types exist which constitute the basic framework of a certain Web genre at its

most abstract level. This framework includes one or more compulsory and optional modules. Optional and compulsory web genre modules need not exist in the homepage's physical file, but might be represented by links to other files. He provides an analysis of a random sample containing 200 documents resulted in an initial version of an academic Web genre hierarchy.

LEE AND MYAENG (2002 AND 2004) present a methodology for genre classification using word statistics. They construct their own Web document collections (reportage, editorial, technical paper, critical review, personal home page, Q&A, and product specification) in English and Korean. A term is a good feature depending on three factors: a) the number of documents belonging to the genre containing the term; b) how evenly the term is distributed among the subject classes that divide the genre class; c) how discriminating the term is among different genre classes. They test two approaches: document frequency ratios (df) and term frequency ratios (tf). They use the Naïve Bayesian approach and the similarity approach. Results show that in the similarity approach, the use of df ratios gives significantly better performance that tf ratios across all cases. It seems that the deviation formula makes use of both genre-classified documents and subject-classified documents to eliminate terms that are more subject-related than genre-related. On the other hand, the overall performance is quite different when the Bayesian approach is used. The best results were obtained when the ordinary tf values only were used. The performance difference between the new method and the best naïve Bayesian approach is greater in the Korean collection than in the English collection. More specifically, LEE AND MYAENG (2004) show that while subject classes clearly help improving the genre-based classification, it is not clear whether using the genre class information for documents in the same way helps subject-based classification.

FINN AND KUSHMERICK (2003) expand FINN ET AL. (2002) and cover also the dimension of positive and negative. They try this dimension on film and restaurant reviews. They use 3 feature sets as in FINN ET AL. (2002) and the results indicate that the POS approach is not suitable for the task of classifying reviews as being either positive or negative. The bag-of-word approach can achieve good performance in a single subject domain, but cannot transfer to new subject domains. An interesting new idea is to use domain transfer performance as a means of evaluating features set performance. For example, in this case, even though the traditional methods of evaluating a classifier indicate that the bag-of-words approach achieves good performance, their experiments indicate that it performs poorly when their extra domain transfer condition is evaluated. The review classification task is more difficult than the subjectivity classification task. All feature sets achieve good single domain accuracy on the latter task, while the POS feature set also achieves good domain transfer. On the review classification task, the bag-of-words approach achieves good single domain accuracy, but none of the feature sets achieve good domain transfer. They try also to combine all feature sets and see if they get a better performance. More precisely, they combine all the models based on their different feature sets. They call this combination a "multi-view-ensemble" learning approach (MVE). A majority vote is taken to classify a new instance. This approach to classification exploits the fact that the three different feature sets do not make mistakes on the same documents. Therefore, a mistake made by the model based on one feature set can be corrected by the model based on another feature

sets. When each feature set performs well, they are more likely to correct each other mistakes. In cases where some of the feature sets perform poorly, this approach will achieve performance that is proportional to the relative performance of the individual feature sets. It seems that for genre classification tasks where it is not clear which feature set is most suitable for the task, this approach could increase the likelihood of the classifier to perform better. The results achieved by MVE are encouraging.

# 5 *Discussion*

Keeping on using the word "genre" as a cover term, we can say that automatic genre identification/classification has a fairly long tradition by now, since 1988 when Biber's work on the variation across spoken and written (BIBER 1988)[21].

We have seen that two main approaches have singled out: bottom-up/unsupervised and top-down/supervised. The bottom-up approach allows us to detect similarities and differences among types of documents belonging to the same or to different genres. This approach overrides the social-cultural genre labels, and provides a deep linguistic insight into texts. On the contrary, the top-down approach relies on a set of prototypical texts, correctly classified into genres by humans (the training set), to derive the classification on unclassified documents (the test set).

Apart from the statistical weaknesses of both these approaches[22], they point at different targets. The bottom-up approach is mainly linguistic-oriented, it sheds new light on the linguistic nature of texts, but Biber's taxonomy of text types has not gained ground among linguists and his text typology is too fuzzy and subjective to be generalized. The top-down approach is extremely coarse because being based on prototypical documents; it performs reasonably well only with texts very similar to the ones included in the training set.

The very weak point with both of these approaches and with automatic genre identification in general is that they rely on a restricted number of linguistic and (for more recent research) layout features. The 67 linguistic features selected by Biber (BIBER 1988: 221-245) more than 15 years ago are derived from previous sociolinguistic studies, but NLP tools were quite limited at that time. Even if Biber claims that his features can be helpful in detecting the type of a text, most features are mainly lexico-grammatical, and syntactic and discourse structures are under-represented or misrepresented. For

---

[21] His work derives from his PhD dissertation ended in 1984: Douglas Biber (1984), *A model of textual relations within the written and spoken modes*, Unpublished PhD dissertation, University of Southern California. But text genre classification had also a European tradition. PHILLIPS (1985) suggests that a typology of texts could be established on the basis of differing tendencies to collocational patterns in different kinds of texts. PIRRELLI (1985) applies Multidimensional Scaling (MDS) to the Italian corpus and builds on the experience of the French tradition in this field (see PIRRELLI (1985)'s bibliography).

[22] In Lee (2003), the statistical validity and the empirical stability of the bottom-up approach is challenged. A common problem with the top-down approach is *overfitting*. "Classifiers which overfit the training date tend to be extremely good at classifying the data they have been trained on, but are remakably worse at classifying the other data" (SEBASTIANI 2002: 15). Cf. also WULFEKUHLER M., PUNCH W. (1997): "the standard approach of classifying training documents can be done with high accuracy, but is not generalizable for classification of documents not in the training set, and is not useful for discovering new data".

example, he only includes *because* among the causative adverbial subordinators, because other forms, such as *as, for* and *since* can be causative, but as they have also a range of other functions, they are too ambiguous to be included among causative features. Why should we blame Biber if 15 years ago NLP tools were not so advanced as they are today? Nowadays we can extract sophisticated linguistic knowledge from parsers' output, not only morpho-syntactic, but also macro-syntactic[23], with grammatical relations and dependency functions[24]. His features were mostly lexical, because searching for specific words is much easier and straightforward than trying to extract more complex syntax or information about the textual structure. But this was 15 years ago. Now Biber's features are almost completely inadequate if we want to upgrade from traditional ASCII corpora to the Web. The Web is a huge reservoir of textual resources[25]. It is not only huge; it is unpredictable, untamed and mostly unclassified. Layout plays a major role in Web pages: it is extremely important because the Web has a strong visual nature. This is why the most recent projects on automatic genre identification/classification are complementing Biber's features with layout features (TyPWEB (BEAUDOUIN ET AL. 2001A, BEAUDOUIN ET AL. 2001B, ILLOUZ AND HABERT 2002), RAUBER AND MÜLLER-KÖGLER 2001) or using predominantly HTML tags (REHM 2002). Layout has a strong impact on syntax, punctuation and on the textual structure of Web documents. The concept of extra-grammaticality[26] triggered by layout is an interesting one, and deserves more investigations.

Web documents also entails additional issues. 1) What is a "Web document"? Is it a single Web page? Is it the bunch of Web Pages connected together by hyperlinks? REHM (2002) proposes a model where the all sets of documents linked together should be taken into account in order to detect the Web genre "Academic's Personal Homepage". Also the TyPWEB project (BEAUDOUIN ET AL. 2001A, BEAUDOUIN ET AL. 2001B, ILLOUZ AND HABERT 2002) goes beyond the single physical document by aiming at discriminating commercial website and personal website. Most of the other projects, instead, consider the individual Web page as a Web document.

---

[23] Syntax has strong discriminating power. Also stylometric researchers start (though reluctantly) yielding before the facts. AARONSON S. (1999) states: "I'm somewhat surprised that grammar rule frequency performed at least competitively with data-driven features. I conjecture that if I test at least frequent grammar rules and vary the number of rules used […], I might see grammar rule frequency definitively outperforms the data-driven features. At the very least, I think the results suggested that fully automatic parsers hold promise for stylometric applications".

[24] For example, Connexor Machinese Syntax for English is a syntactic parser, which produces part-of-speech classes, noun phrase syntax and syntactic relations (subject, object, complements, verb chains, adverbial functions, etc.) (http://www.connexor.com/demos/syntax_en.html). Another interesting parser is RASP, which outputs grammatical relations (GRs) associated with a particular analysis (http://www.cogs.susx.ac.uk/lab/nlp/rasp/).

[25] In the Information Retrieval area, research is being carried on Web page classification. Approaches seem to be quite tentative and exploratory (KOVACEVIC ET AL. 2002), Web page categories appear to be arbitrary (ASIRVATHAM AND RAVI 2001), or simply topical (WONG AND FU 2000, RIBONI 2002).

[26] By extra-grammaticality we mean unprecedented uses of grammar not yet included and listed into traditional grammars. Interesting findings come out in the Natural Language Generation field: "laid-out texts abound with sentences which should be considered to be grammatically ill-formed in standard, running texts but which are "extra-grammatical" in that they conform to a grammar acceptable to the genre" (BOUAYAD-AGHA N. ET AL. 2000).

2) What kind of taxonomy could we adopt for Web genres? Some studies are trying to classify Web pages using new genre labels (CROWSTON AND WILLIAMS 1997, SHEPHERD AND WATTERS 1998, 1999). However, when text categories are culturally and socially-based, they are unstable, they change over time, following social and cultural trends and usages. Just think at the genre "novel": *Le Roman de Thèbes* and *Eugene Onegin* are novels in verse, but for us a novel is something opposed to poetry: a novel must be written in prose to be a novel, all over the world. Similarly, as far as we currently know, a "home page", which for us can be a "personal home page", a "corporate home page", an "academic home page", an "e-shop home page", a "e-bank home page" etc. in a few months it can just refer to a "personal home page", as ROBERTS (1998) seems to suggest. The title of his paper is "The *Home Page* as a genre: a narrative approach", but what he means by home page is "the personal home page" and he proposes the acronym PHP. What is interesting in ROBERTS (1998)'s paper is also the approach he wants to suggest: *a narrative approach*. Similarly, SMOLIAR AND BAKER (1997) attempt to apply the discipline of narratology to hypermedia to classify documents as instances of different text types. In their paper, they consider the following text types: description, argument, and narrative. They analyse and give an example of each as a hypermedia document: the preparation of a recipe (description), a report of a group meeting (argument), and recounting one of the versions of the Rashomon story (narrative). They then argue that this classification provides an organizational framework that facilitates the construction of the reader's understanding of the content that the writer intended to convey.

In the quagmire of different interpretations of what is a genre, we think that a classification based on traditional and intuitive text categories, such as *narrative*, could really give a hand to genre classification activities, especially for Information Retrieval tasks. If we state that a ***technical manual*** is a kind of *objective description*, a ***report*** is a kind of *objective narration*, a ***comment*** is a kind of *subjective argumentation*, a ***recipe*** is a kind of *objective instruction*, etc., the set of features we need to extract in order to achieve this kind of classification is quite self-evident, and there is a bulk of linguistic and typographic[27] studies that suggest the most representative features for these classes of texts.

There are two important things we need to keep in mind. First, the Web is something still new, still evolving, still fluid, and consequently changeable. If we try to make a very rigid classification now, we are bound to make a very bad investment, because the Web uses communicative forms that have not been standardized yet, and technology allows us new forms of communication every single day. Second, we must not forget that a text (on any media) almost never belongs to only one class. This mono-oriented classification is one of the major limitations of traditional text categorization

[27] The role of typography/layout and punctuation is becoming more and more fully-acknowledged by NLP community. A specialized ATALA workshop was held in November 2003 in Paris, with the title *Role of typography and punctuation in natural language processing*. New and more proactive approaches were proposed: "The pre-processing of a text must exploit the formal structure (titles and sub-titles localisation; text fragmentation in sentences, paragraphs, utterances, propositions, words; quotation identification; item list identification; spatial disposition consideration; images, diagrams, captions, boxes localisation....), before executing other tasks […] Without complete control of the exploitation of formal structure, text processing will not really be operational".

(categorization based on topic): if we classify an article on Wimbledon tennis competition only as a sport article, forgetting about its nature of event for society people, we are making a very arbitrary classification. This is one of the inadequacies of the traditional text categorization based on topical categories.

A text is a mixture of different forms of expressions and different communicative acts; it almost never corresponds to an ideal or idealized type (BEAUGRANDE AND DRESSLER 1981: 181 FF.). If we want to be realistic, we must give a ranking, listing all the probabilities for a text to belong to several different classes. If we could say to a search engine user looking for touristic information about Cornwall that a Web page is 80% descriptive (because it contains a list of places to visit with a short description), 50% instructional (because it contains a list of restaurants, a bus timetable, and a list of museums), and it is for 30% a comment (because it contains the welcome of the mayor of Penzance, who praises local attractions), may be our user could be more satisfied than he currently is. The main evaluation measure for such a system would be the user feedback on the proposed ranking.

# 6 *Conclusions*

This survey on automatic genre identification has highlighted a few points:

- The terminology is fuzzy. Many different terms have been used to refer to a classification principle that is not based on the "content" of a document.
- The notion of genre in most projects is extremely vague (with some exceptions, for instance REHM 2002 and IHLSTRÖM AND ÅKESSON 2004). There is no general agreement on what is to be considered a "genre". As extreme cases, facts and opinions have been considered genres.
- From the multivariate statistical techniques used in the projects that follow Biber's approach, the tendency is now to use machine learning techniques, with a supervised approach, where a classifier learns the characteristics of different genres from a set of pre-classified examples.

We can say that, at least so far, almost everything in the automatic genre identification research field is fuzzy, slippery, unstable, flexible (especially the notion of "genre" and the terminology), and conditioned by the computational cost of extracting relevant features. In addition, it seems that not much attention has been paid to how to deal with documents that are mixed. Especially with supervised machine learning techniques, documents are forced into one single genre, while we know that real life documents are mostly mixed, showing combination of genres. As pointed out earlier, KESSLER ET AL. (1997) have proposed a multi-faceted approach to genres, but their practical approach is not convincing, nor the facets proposed. Very recently, CROWSTON AND KWASNIK (2004) have suggested a more comprehensive framework for a faceted classification of genres. A faceted classification is particularly appropriate because not only are genres complex entities, but also dynamic, new ones emerge and old ones can collapse into the new ones, creating a mixture. The problem is to design intuitive facets, easily derived from features extractable from the documents. An attempt to address this issue (which is especially significant for Web documents) is made by two current projects, KWASNIK ET AL. (2001) and SANTINI (2003).

# 7 References

AARONSON S. (1999), *Stylometric Clustering: A comparative analysis of data-driven and syntactic features*, Project report available at http://www.cs.berkeley.edu/~aaronson/sc/report.doc

ARGAMON S., KOPPEL M., AVNERI G. (1998), "Routing documents according to style", *Proceedings of the First International Workshop on Innovative Internet Information Systems* (IIIS-98).

ASIRVATHAM A., RAVI K. (2001), "Web Page Classification based on Document Structure", available at http://www.iiit.net/stud_pdfs/kranthi1.pdf.

BAAYEN H., HALTEREN H. VAN, TWEEDIE F. (1996), "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution", *Literary and Linguistic Computing*, 11.

BAGDANOV A., WORRING M. (2001A), "Content-free Document Genre Classification using First Order Random Graphs", *Proceedings of the seventh annual conference of the Advanced School for Computing and Imaging (ASCI '01)*, Heijen, The Netherlands, June 2001

BAGDANOV A., WORRING M. (2001B), "Fine-Grained Document Genre Classification Using First Order Random Graphs", *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR '01)*, Seattle, Washington, September 2001.

BEAUDOUIN V., FLEURY S., HABERT B., ILLOUZ G., LICOPPE C., PASQUIER M. (2001A), "TyPWeb: décrire la Toile pour mieux comprendre les parcours", *CIUST'01, Colloque International sur les Usages et les Services des Télécommunications, e-Usages*, Paris, 12-14 juin 2001.

BEAUDOUIN V., FLEURY S., HABERT B., ILLOUZ G., LICOPPE C., PASQUIER M. (2001B), "Traits textuels, structurels et présentationnels pour typer les sites web personnels et marchands", available at http://www.atala.org/je/010428/TyPWeb.ppt

BEAUGRANDE R. DE, DRESSLER W. (1981), *Introduction to Text Linguistics*, Longaman, London-New York, 1981.

BIBER D., JOHANSSON S., LEECH G., CONRAD S., FINEGAN E. (1999), *Longman Grammar of Spoken and Written English*, Longman, Harlow.

BIBER, D. (1988), *Variations across speech and writing*, Cambridge University Press, Cambridge.

BIBER, D. (1989), "A typology of English texts", in *Linguistics*, Vol. 27, 3-43.

BIBER, D. (1995), *Dimensions of register variation*, Cambridge University Press, Cambridge.

BOUAYAD-AGHA N., SCOTT D., POWER R. (2000), "Integrating content and style in documents: a case study of patient information leaflets" *Information Design Journal 9:2-3, pp. 161-176.*

BRETAN I., DEWE J., HALLBERG A., WOLKERT N., KARLGREN J. (1998A), "Web-Specific Genre Visualization", *WebNet '98*, Orlando, Florida, November 1998.

BRETAN I., DEWE J., HALLBERG A., WOLKERT N., KARLGREN J. (1998B), "Web-Specific Genre Visualization", in *Proceedings of the 3rd World Conference on the WWW and Internet*, Orlando, Florida, November. AACE.

BROWN E., SMEATON A. (1998), "Hypertext Information Retrieval for the Web", *SIGIR'98 Forum*.

CROWSTON K., WILLIAMS M. (1997), "Reproduced and Emergent Genres of Communication on the World-Wide Web", *Proceedings of the 30th Hawaii International Conference on System Sciences (HICSS-30)*.

CROWSTON K., KWASNIK B. (2004), "A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality", *Proceedings of the 37th Hawaii International Conference on System Science (HICSS '04).*

DEWDNEY N., VANESS-DIKEMA C., MACMILLAN R. (2001), "The form is the Substance: Classification of Genres in Text", *ACL '2001 Conference*, Toulouse, France, also available at http://www.elsnet.org/km2001/dewdney.pdf

DILLON A., VAUGHAN M. (1997), "It's the journey and the destination: Shape and the emergent property of genre in evaluating digital documents", *New Review of Multimedia and Hypermedia*, Vol. 3, 91-106, availabe at http://www.gslis.utexas.edu/~adillon/publications/journey.html

DIMITROVA M., FINN A., KUSHMERICK N., SMYTH B. (2002), "Web Genre Visualization", submitted to the *Conference on Human Factors in Computing Systems* (Minneapolis).

EAGLES 1996, *EAGLES Preliminary Recommendations on Text Typology*, EAGLES Document EAG-TCWG-TTYP/P, Version of June, 1996, available at: http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html

FINN A. AND KUSHMERICK N. (2003) "Learning to classify documents according to genre" *IJCAI-2003 Workshop on Computational Approaches to Text Style and Synthesis* (Acapulco).

FINN A., KUSHMERICK N., SMYTH B. (2001), "Fact or fiction: Content classification for digital libraries", *Joint DELOS-NSF Workshop on Personalisation and Recommener Systems in Digital Libraries* (Dublin).

FINN A., KUSHMERICK N., SMYTH B. (2002), "Genre classification and domain transfer for information filtering", *Proceedings of European Colloquium for Information Retrieval Research* (Glasgow).

FOLCH H., HEIDEN S., HABER B., FLEURY S., ILLOUZ G., LAFON P., NIOCHE J., PRÉVOST, S. (2000), "TyPText: Inductive typological text classification by multivariate statistical analysis for NLP sistems tuning/evaluation", *LREC 2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May - 2 June 2000.

HABERT B., ILLOUZ G., FLEURY S., FOLCH H., HEIDEN S., PRÉVOST S. (2000), "Profilage de textes: cadre de travail et expérience", *JADS 2000: 5ES Journées Internationales d'Analyse Statistique des Données Textuelles*, available at http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/56/56.pdf

ILLOUZ G., HABERT B. (2002), "TyPWeb: Typologie et Profilage de sites Web", available at http://www.limsi.fr/RS2002FF/CHM2002FF/LIR2002FF/lir4.html

ILLOUZ G., HABERT B., FLEURY, S., FOLCH H., HEIDEN S., LAFON S. (1999), "Maîtriser les déluges de données hétérogènes", *Atelier Thématique TALN 1999, Cargèse, 12-17 juillet 1999*.

ILLOUZ G., HABERT B., FOLCH H., HEIDEN S., FLEURY, SERGE, LAFON S., PRÉVOST S. (2000), "TyPText: Generic features for Text Profiler", *Content-Based Multimedia Information Access (RIAO'2000)*, Paris.

IHLSTRÖM C., ÅKESSON M. (2004), "Genre Characteristics – a Front Page Analysis of 85 Swedish Online Newspapers", *Proceedings of the 37th Hawaii International Conference on System Science (HICSS '04)*.

JOHANNESSON E., WALLSTRÖM C. (1999), "Automatic Analysis and Visualization of Stylistic Genres", paper presented at *The Twenty Second IRIS Conference (Information Systems Research Seminar In Scandinavia)*, 7-10 August, Keuruu, Finland, available at http://iris22.it.jyu.fi/iris22/pub/Wallstr%F6m_Johannesson_R315.pdf

KARLGREN J. (1996A), "Non-Topic Information Retrieval using Computational Stylistics", *ERCIM News* No. 26 – July 1996 – SICS.

KARLGREN J. (1996B), "Stylistic Variation in an Information Retrieval Experiment", *Proceedings of The Second International Conference on New Methods in Language Processing – NeMLaP 2*, Bilkent, September 1996. Ankara: Bilkent University.

KARLGREN J. (1999), "Stylistic Experiments in Information Retrieval", *Natural Language Information Retrieval*, Strzalkowski T. (ed.), Kluwer.

KARLGREN J. (2000), *Stylistic Experiments for Information Retrieval*, Thesis submitted for the degree of Doctor of Philosophy, Department of Linguistics, Stockholm University.

KARLGREN J., BRETAN I., DEWE J., HALLBERG A., WOLKERT N. (1998), "Iterative Information Retrieval Using Fast Clustering and Usage-Specific Genres". In *Proceedings of the 8th DELOS Workshop on User Interfaces in Digital Libraries*, pp. 85-92, Stockholm, Sweden, October. ERCIM.

KARLGREN J., CUTTING D. (1994), "Recognizing Text Genre with Simple Metrics Using Discriminant Analysis", *Proceedings of COLING 94*, Kyoto.

KESSLER B., NUMBERG G., SHÜTZE H. (1997), "Automatic Detection of Text Genre", *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*.

KLAVANS J., KAN M. (1998), "Role of Verbs in Document Analysis", *Proceeding of the Conference*, COLING-ACL 1998, PP. 680-686.

KOVACEVIC M., DILLIGENTI M., GORI M., MILUTINOVIC V. (2002), "Recognition of Common Areas in a Web Page Using a Visualization Approach", *Artificial Intelligence: Methodology, Systems, and Application*, Scott D. (ed). pages 203 ff.

KWASNIK B., CROWSTON K., NILAN M., ROUSSINOV D. (2001), "Identifying Document Genre to Improve Web Search Effectiveness", *Bulletin of The American Society for Information and Technology*, Vol 27, No. 2.

LEE D. (2001), "Genres, Registers, Text types, Domains, and Styles: Clarifying the concepts and navigating a path through the BNC Jungle", *Language Learning and Technology*, Vol. 5, Num. 3, pp. 37-72, also available at http://llt.msu.edu/vol5num3/lee/

LEE D. (2003), *Modelling variation in spoken and written language: The multi-dimensional approach revisited*, London , Routledge.

LEE Y., MYAENG S. (2002), "Text Genre Classification with Genre-Revealing and Subject-Revealing Features", *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR 2002: 145-150

LEE Y., MYAENG S. (2004), "Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization ", *Proceedings of the 37th Hawaii International Conference on System Science (HICSS '04).*

MICHOS S., STAMATATOS E., FAKOTAKIS N., KOKKINAKIS G. (1996), "An Empirical Text Categorizing Computational Model Based on Stylistic Aspects", *Proceedings of the 8th International Conference on Tools with Artificial Intelligence (TAI'96)*.

NAKAMURA, J. (1993), "Statistical Methods and Large Corpora – A New Tool for Describing Text Type", in *Text and Technology*, Baker M., Francis G., Tognini-Bonelli E. (eds.), J. Benjamins Publishing Company, Philadelphia - Amsterdam, pp. 291-312.

PHILLIPS, MARTIN (1985), *Aspects of text structure*, North-Holland, Amsterdam–New York–Oxford.

PIRRELLI, V. (1985), *La statistica multivariata nell'analisi linguistica del testo. Un esempio: lo Scaling Multidimensionale*, Unpublished dissertation, Università degli Studi di Pisa, supervisor: Antonio Zampolli.

RAUBER A., MÜLLER-KÖGLER A. (2001), "Integrating Automatic Genre Analysis into Digital Libraries", ACM/IEEE joint Conference on Digital Libraries 2001.

REHM G. (2002), "Towards Automatic Web Genre Identification. A corpus-based approach in the Domain of Academia by Example of the Academic's Personal Homepage", *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*.

RIBONI D. (2002), "Feature Selection for Web Page Classification", available at http://students.silab.dsi.unimi.it/~dr548986/riboni02.pdf

ROBERTS G. (1998), "The Home Page as Genre: A Narrative Approach", *Proceedings of the 31st Hawaii International Conference on System Sciences (HICSS-31)*.

ROUSSINOV D., CROWSTON K., NILAN M., KWASNIK B., CAI J., LIU X. (2001), "Genre Based Navigation on the Web", *Proceedings of the 34th Hawaii International Conference on System Sciences (HICSS-34)*.

SANTINI M. (2003), *Identifying Genres on the Web: PhD Thesis Outline*, Technical Report ITRI-03-06, available at http://www.itri.brighton.ac.uk/techindex.html

SCHMID-ISLER S. (1997), "The Language of Digital Genres. A Semiotic Investigation of Style and Iconology on the World Wide Web", *Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS-33)*.

SEBASTIANI F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, 34(1):1-47.

SHEPHERD M. AND WATTERS C. (1998), "The Evolution of Cybergenre", *Proceedings of the 31st Hawaii International Conference on System Sciences (HICSS-31).*

SHEPHERD M. AND WATTERS C. (1999), "The Functionality Attribute of Cybergenres", *Proceedings of the 32nd Hawaii International Conference on System Sciences (HICSS-32)*.

SIGLEY R. (1997), "Text Categories and Where You can Stick Them: A Crude Formality Index", *International Journal of Corpus Linguistics*, Vol. 2 No. 2, pp. 199-237.

SMOLIAR S., BAKER J. (1997), "Text Types in Hypermedia", *Proceedings of the 30th Hawaii International Conference on System Sciences (HICSS-30)*.

STAMATATOS E., FAKOTAKIS N., KOKKINAKIS G. (2000), "Text Genre Detection Using Common Word Frequencies", *Proceedings of the 18th International Conference on Computational Linguistics (COLING2000)*.

STAMATATOS E., FAKOTAKIS N., KOKKINAKIS G. (2001), "Automatic Text Categorization in Terms of Genre and Autour", *Computational Linguistics* 26, 4, pages 471-495.

STUBBS M. (1998), *Text and Corpus Analysis*, Blackwell Publishers, Oxford, Reprinted 1998 (first published 1996).

SWALES J. (1990), *Genre Analysis. English in academic and research settings*, Cambridge University Press, Cambridge.

TAKAHASHI K. (1997), *A Study of Text Typology: Multi-feature and multi-dimensonal analyses*, UCREL Technical Paper, University of Lancaster.

WALLER R. (1987), *The typographic contribution to language. Towards a model of typographic genres and their underlying structures*, PhD thesis submitted to the Department of Typography & Graphic Communication, University of Reading.

WERLICH E. (1976), *A Text Grammar of English*, Quelle & Meyer, Heidelberg (Germany).

WIKBERG, K. (1993), "Verbs as indicators of text type and/or style: Some observations on the LOB corpus", in *Corpus-based Computational Linguistics*, Souter C., Atwell E. (eds.), Rodopi, Amsterdam–Atlanta, pp. 127-145.

WOLTERS M., KIRSTEN M. (1999), "Exploring the Use of Linguistic Features in Domain and Genre Classification", *Proceedings of EACL '99*, pages 142-149.

WONG W., FU A (2000), "Incremental Document Clustering for Web Page Classification" in *Proceedings of 2000 International Conference on Information Society in the 21st Century: Emerging Technologies and New Challenges (IS2000)*, Aizu-Wakamatsu City, Fukushima, Japan November 5-8.

WULFEKUHLER M., PUNCH W. (1997), "Finding Salient Features for Personal Web Page Categories", available at
http://decweb.ethz.ch/WWW6/Technical/Paper118/Paper118.html