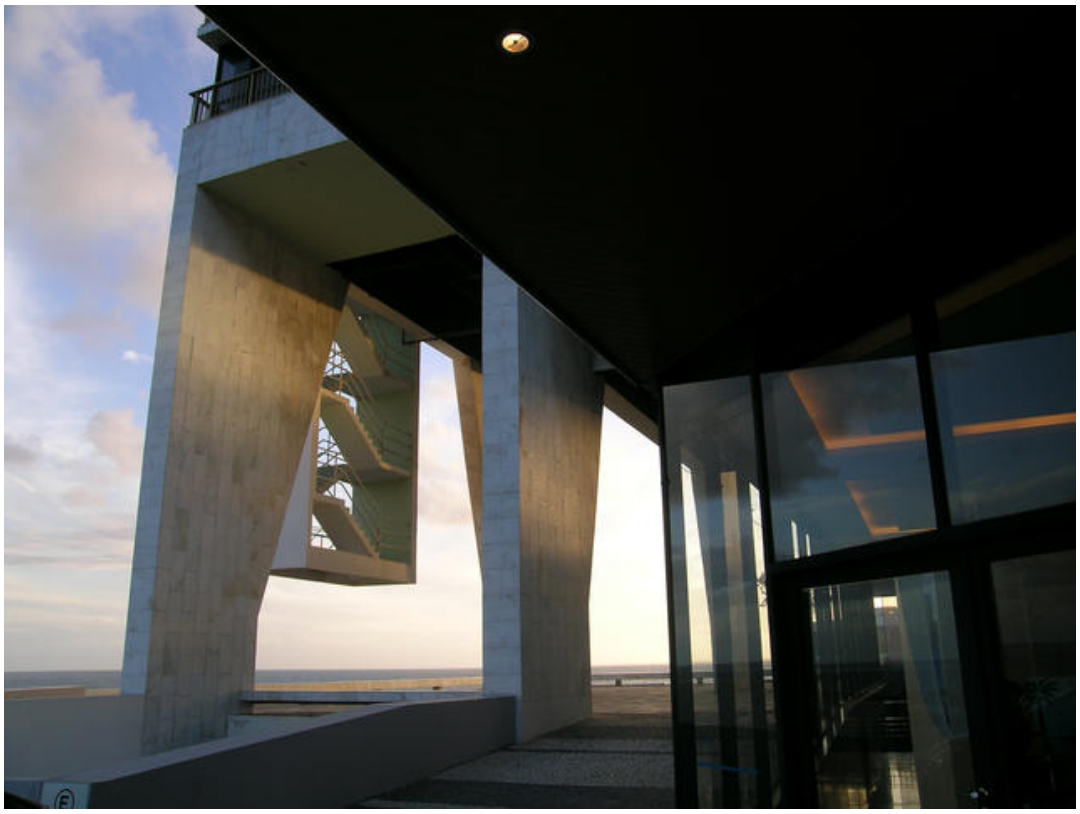
Stylistic Analysis Of Text For Information Access

Shlomo Argamon Jussi Karlgren James G. Shanahan

August 19, 2005



Stylistic Analysis Of Text For Information Access

Shlomo Argamon Jussi Karlgren James G. Shanahan

August 19, 2005

Abstract

Papers from the workshop held in conjunction with the 28th Annual International ACM Conference on Research and Development in Information Retrieval, August 13-19, 2005, Salvador, Bahia, Brazil

Keywords

Stylistic analysis, Genre, Text, Information Access, Language Technology

Swedish Institute of Computer Science
Jussi Karlgren
jussi@sics.se
Box 1263, S–164 29 Kista
Sweden

SICS Technical Report T2005:14
ISSN 1100-3154
ISRN SICS-T–2005/14-SE

Table of Contents

Theme and goals of the workshop

Shlomo Argamon, Jussi Karlgren, James G. Shanahan

Information management systems have typically focused on the "factual" aspect of content analysis. Other aspects, including pragmatics, opinion, and style, have received much less attention. However, to achieve an adequate understanding of a text, these aspects cannot be ignored.

This workshop, held on the day following the 2005 SIGIR conference, was the first ever to specifically address the automatic analysis and extraction of stylistic aspects of natural language texts for purposes of improving information access.

Stylistic Analysis

The goal of improving the textual analysis of information access systems is a motivating factor for stylistic research. In addition, readers, authors, and information specialists of whatever persuasion are aware of stylistic variation. This provides us with the added philological motivation for research: that of understanding text, readers, and authors better.

Style may be roughly defined as the "manner" in which something is expressed, as opposed to the "content" of a message. Stylistic variation depends on author preferences and competence, familiarity, genre, communicative context, expected characteristics of the intended audience and untold other factors, and it is expressed through subtle variation in frequencies of otherwise insignificant features of a text that, taken together, are understood as stylistic indicators by a particular reader community. Modeling, representing, and utilizing this variation is the business of stylistic analysis.

Applications

Useful applications of stylistic analysis abound, including systems for genre-based information retrieval, authorship attribution, plagiarism detection, context-sensitive text or speech generation systems, organizing and retrieving documents based on their writing style, attitude, or sentiment, quality or appropriateness filters for messaging systems, detecting abusive or threatening language, and more.

Challenges

Style work to date has been stymied by two obstacles. Given the subtlety and complexity of the phenomena, automated learning systems need a considerable amount of (tagged) text before achieving reliable performance. As a result, few theories have been specified and few linguistic resources have been developed to a level where reliable tagging is easy and reliable.

Our purpose with the workshop, therefore, was to bring together people from various areas of intellectual endeavour to explore core issues regarding the annotation, modeling, mining, and classification of style in

text, across a range of text information management applications. The goal was to address a rather wide range of issues, from theoretical questions and models about style, through annotation standards and methods, to algorithms for recognizing, clustering, and displaying these aspects.

This objective was at partially met: for future meetings express invitations should be extended to practitioners and parties in the business of information production and dissemination.

Challenge Questions

We invited contributions to address challenges such as the following:

Style in Theory:

- What is style?
- How can it be defined?
- How does it differ from other types of non-topical variation?
- What are its social characteristics and interpretations?
- What dimensions of variations do you assume?
- What is the appropriate level of abstraction for best explanatory power?
- What linguistic universals of style may be identified?

Style in Engineering:

- How is style analyzable?
- What is the appropriate level of abstraction for useful computational purposes?
- What features are valuable for analysis?
- How could stylistic information be used for generation or modification of existing information?
- What issues and solutions exist for cross-lingual style analysis and synthesis?

Style in Applications:

- What tasks can stylistic information be used for?
- How do people understand style?
- Can stylistic information be used profitably e.g. in information access interfaces?

Style in Research:

- What tools and resources do you use, and can we use them too?

Program

The program for this workshop was tight and full of presentations: this was an exploratory meeting, with presentations ranging extensively across various examples of non-topical analysis of text. The data sets used, the features extracted, the target dimensions aimed at, and the computational schemes employed varied widely, attendant to the impressive variation in application.

Speaking generally, the participants agreed that taking first steps in stylistic analysis of text is quite easy:

- select computable textual features;
- combine them judiciously;
- model the choice space;
- compare results from measurements on texts under consideration to some norm or norms.

This is a process which is familiar to any practicioner of information access research. The challenge, returning to the motivations mentioned above, is to ensure that the analysis has reliable predictive power for the application under consideration, and that the results have adequate explanatory altitude to provide purchase for further study and generalization.

Evaluation

Evaluation was naturally at the forefront of the presentations. The various application areas motivated several different approaches to evaluation, from the relatively clear case of authorship attribution and forensic applications to the less clear cut ones one of mood classification of blog posts. For any information access application, the evaluation must be both operationally quantifiable and related to some formalization of user needs — one of the projects presented explicitly gathered user opinions for an information retrieval system which utilized stylistic analysis for presentation of results.

Feature Rally

The crucial methodological difference between stylistic analysis and topical information retrieval is that of feature extraction. The features studied are different than those studied in topical analysis of text – in the workshop we addressed this in a Feature Rally session, where participants were invited to present their favourite feature in a few minutes.

Common Resources

To better synchronize the efforts of resources, the workshop decided to establish a clearinghouse for common resources, a mailing list, and a bibliography of previously published research. First and foremost, the proceedings of this first workshop on textual stylistics in information access will be made publicly available.

Any interested parties are welcomed to contact the organizers for further information!

Future Meeting

At the end of the proceedings, the workshop ended with the consensus that future meetings are in order, including the possibility of requiring a common task addressing a common data set for participants to provide an embryo of a common evaluation scheme. Contact the organizers of this years workshop to find out more!

A Stylometry-Based Method to Measure Intra and Inter-Authorial faithfulness for Forensic Applications

Marta Sánchez Pol
Institut Universitari de Lingüística Aplicada (IULA)
Universitat Pompeu Fabra
La Rambla, 30-32
08002 Barcelona
+34 699 300 818
martaspol@gmx.atT

ABSTRACT
This article presents a method to measure author stylistic faithfulness, the so-called idiolect. Given the issue raised within the stylometry domain on the diverse possibilities of measuring style, we base our work on stylistics and statistics to measure an author's internal variability. Results are applied to text comparison for authorship attribution.

Keywords
Stylometry, authorship attribution, lexicometry, statistics applied to linguistics, stylistics.

1. Introduction

From its origins, stylistics has been used to analyze style from two major perspectives (Ullmann, 1964), i.e. style either of a given language or of a text or author. The main objectives of style research based on text or author level have aimed at describing a text from a rather formal perspective (number of words, repetitions, sentence length) so that such text can be attributed to an author, a time or a geographical zone (or dialect, for instance). But stylistics has also been applied to other subject fields such as the study of language disorders, e.g. aphasia (Holmes, 1996), or genre-oriented text categorization (Stamatatos, 2000).

In this study, we assume that language is a sequence of options and choices (Halliday, 1978) and that a writer or speaker tends to be recursive when selecting language units from a range of possibilities (options). Such selection constrains the writer's options so that when a new choice is made an option-choice sequence is created and a writer's trace is marked within the language set. We believe that if such trace can be measured, key information about style for authorship attribution it could be determined if it actually distinguishes an author from other authors. Thus, according to this theoretical assumptions, the main objective of this work is to measure intra and inter-authorial variation by quantifying stylistic variables through different language levels, particularly lexical, syntactic and semantic.

The main questions underlying this research correspond to those raised by Sanders (1977) in regard to what he calls *Stiltheorie* (theory of style) and *Stilistik* (stylistics):

> What is style?
>
> Is style measurable?
>
> How could it be measured?

We would add one more question that could be interesting for style analysis:

> To what extent is an author faithful to his or her own style?

In this article we attempt to propose a new approach for style analysis. First we describe the theoretical assumptions within which this experiment is framed; then the methodology of the study is presented and the first results obtained from the initial experiments are shown. Finally, we conclude with some perspectives for future work.

1.1 Domains of analysis

This study is framed within *a)* stylistics, since its main purpose is to find out if style is measurable, *b)* forensic linguistics, because its immediate application is authorship attribution, *c)* computational linguistics, given that an automatic extraction of stylistic variables is carried out, and *d)* lexical statistics or lexicometry, because our methods for the analyses are mostly statistical.

2. Proposal for style measurement

2.1 Theoretical framework

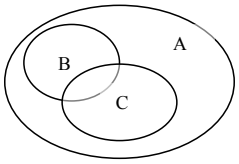
The theoretical framework for this study combines Halliday's (1978) language theory and Sanders' (1977) theories of stylistics. We base on Halliday's assumption of language as a series of options and choices and agree with his notion of text as a representation of choices: "A text is what is meant, selected from the total set of options that constitute what can be meant".

On the other hand, Sanders' (1977) definition of style complements this notion of text: "Dann wird Stil aufgefasst als das Resultat aus der Auswahl des Autors aus den konkurrierenden Möglichkeiten des Sprachsystems".[1] That is, within a wider set understood as language (options), each author makes choices that in turn constitute both text and style.

However, although we share most of Sanders' ideas, we do not agree with his statement about the impossibility of generalizing within stylistics: "Diese Stilistik will kein Patentrezept der Stilanalyse bieten (das es im übrigen gar nicht geben kann, da jeder konkrete Text pontentiell neue, unvorgesehene Probleme stellt)".[2] We do believe that there exists a method to carry out a stylistic study of the text as a whole. In this study we propose a method able to stylistically analyze any text with independence of length, genre and other text variations.

The first experiments underteaken have validated the hypothesis of recursiveness in the author's choices. In spite of this fact, as we expected, some variation has been obtained which makes style attribution difficult. Therefore, we focused on measuring what we have called stylistic faithfulness that represents the style scale within which an author keeps his or her texts.

The following sets illustrate the way this study conceives language and the manner an author uses it (and how such use reflects in style). If set *A* represents all the options offered by a language, *B* and *C* are subsets representing the options selected by two different authors to express an idea.



That is what Sanders calls *principle of choice*[3]. As an example of this theory of language, let me say the following: right now I am writing this article and carefully deciding which language structures best define what I want to transmit to you and which word is the most suitable to make this study understandable. In spite of being a linguist, if I wanted to manipulate my own style, I would not be able to (beyond a certain degree) because the set of options offered by language is determined by personal circumstances. I could certainly use synonyms distant from my usual choices, but there are other style variables that it would not be possible to manipulate such as sentence length, punctuation distribution, etc. "C'est peut-être dans la zone souterraine de sa conscience linguistique, et dans les ramifications enfouies de ses habitudes, qu'un auteur cache les traits originaux de son style" (Dugast, 1980).[4]

Going back to Sanders (1977), it is worth mentioning that he considers text as "the transfer from thought structures to language structures", since it means that the detection of intra- and inter-authorial variability (by measuring style) would represent the extraction of an author's thought structure.

2.2 Objectives

The main goal of this article is to measure style which has to be previously defined through a formula that allows to measure the style of any text by quantifying each one of its variables. The results obtained are applied to text comparison for forensic linguistic purposes (especially authorship attribution).

The study of stylistic variability will allow to measure the faithfulness of an author with his or her own style. Such variability will be determined on a horizontal plane (the variables with the highest or the lowest variability) as well as on a vertical plane (the degree of variability of each variable).

The variation in an author is important considering our purpose of initially comparing two texts through statistical analyses, in order to discover if were written for the same author or independently. The final result of the text comparisons will be given by weighting each variable (about 40). That is the reason why we want to keep away from what has already been shown in some forensic linguistics studies, i.e. one only style variable (n-grams, function words, etc) can be enough for author-oriented text classification. In this study, all the style variables that provide some information about the author will be included.

2.3 Methodology

2.3.1 Corpus

The corpus for this study consists of 20 texts (opinion articles) written by 6 different authors with a total of 120 texts downloaded from online newspapers, between March 2004 and April 2005. All the texts are written in Spanish, and although some geographical variants from Latin America were included so that the analysis is not limited to the single variant from peninsular Spanish.

The first linguistic analyses of this corpus have been carried out on rough text (non-lemmatized and without morphological annotation). At present (since June 2005) the corpus under use is lemmatized and annotated.

The whole corpus is compiled in complete texts, due to we conceive text as an indivisible unit. Furthermore, our objective is to create a measure to analyze the style of any text independently from its length.

---

[1] Style is the result of choices made by an author from a range of possibilities offered by the language system.

[2] This stylistics does not pretend to offer a patented recipe (which, in fact, can not exist, since each text presents unforeseen problems).

[3] "from the total potential offered by a language, only a fragment is selected. Such selection from the language environment differentiates authors and texts from other ones. This is, precisely, the basis of the style and stylistic theory: everything can be expressed in many different ways."

[4] It is, maybe, in the subterranean zone of linguistic consciousness and in the buried ramification of his/her habits where an author hides the original features of his/her style.

2.3.2 Variables of style measurement

The following are some of the variables we are planning to analyze so that useful stylometric information for describing the text style can be obtained:

Total of tokens and lemmas, type/token ratio, Yule's K, total of part of speech categories (nouns, verbs, adjectives, etc), total of content words, total of function words, letter distribution, sentence and paragraph length, sentence type (simple, complex), first level syntactic structures (chunks), discourse connectors, punctuation distribution, etc. Similarly, we are planning to measure some less quantitative and more linguistic features such as vagueness, modalization, use of synonyms, etc.

3. First experiments

During the first stages of analysis, some variables such as type/token ratio, lexical frequency, percentage of function words, punctuation distribution, etc. were extracted from the texts. In the next section, some of these experiments and their results are described. The results are shown in the Figures below in order to illustrate variation by author.

3.1 Function words

Figure 1 presents the dispersion of the function words[5] percentage variable relative to the total of words in the text. It is a way of measuring the lexical richness, since the highest the index of function words, the lower the percentage of content words. The analysis carried out is an analysis of variance (ANOVA) in order to measure if there is any difference among the 20 texts of each author and what dispersion each author offers. At axis X are the authors and at axis Y is the percentage of function words. The box represents where most texts are placed and the lines the maximum and minimum value.

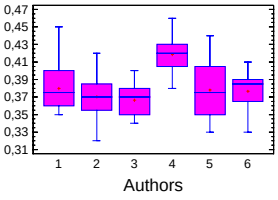


Figure 1 – Dispersion of function word frequency

As it can be noticed, there is wide variation, even though there is no author presenting variation between the highest and the lowest value, that is, our corpus contains some authors with a 32% of function words in their texts and others with a 46%, but there is no single case of an author whose texts show an index of function words ranging from 31% to 46%.

[5] Extracted from the Spanish List of function words prepared by the Real Academia de la Lengua Española.

3.2 Percentage of hapax legomena

Another example is shown in Figure 2 which presents the percentages of the words with frequency 1, the so-called hapax legomena:

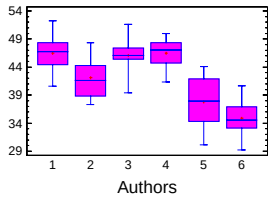


Figure 2 – Dispersion of the hapax legomena

In these variables the dispersion intra-author has decreased. In this case, the percentage of words with frequency 1 varies from 29% to 53%, and although such variation is 24 points, only one author (number 5) exceeds a variation greater than 15 points. The other authors present a variation over 10 points. Therefore, for our purpose, this is a very useful variable. It is another case of lexical richness measurement, since the greater the percentage of hapax, the lesser the repetitions and, as a consequence, more lexical variety.

3.3 Adverbs suffixed with *–mente*

Inversely, Figure 3 presents an example of a variable with excessive dispersion which from this perspective of analysis does not provide us with any interesting information about the author. This Figure shows the percentage of adverbs ending in *-mente* (*-ly*). However, as it was mentioned before, we attempt to analyze the whole text and therefore a linguistic analysis of the use of this variable will be made rather than just a quantitative study of it.

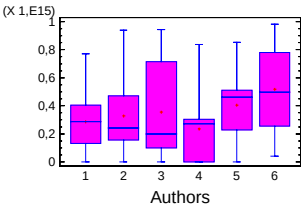


Figure 3 – Dispersion of adverbs suffixed with *–mente*

As the main interest of this work is to analyse those style variables that can provide some information about the author's idiolect, but remaining independent of the content of the text, we decide not to use this variable.

The objective of these experiments was to measure the possible variation within authors for every variables. The program StatGraphics Plus allow us to analyze if we are before different populations (p-value) and gives us the range for each author. This range value is what we use later to measure the stylistic faithfulness.

3.4 High frequency words

Another variable we have studied is that of the most frequent words. The 5 most frequent words have been extracted from each text and the following similarity measure have been determined:

$$s=((x_1,...,x_5),(y_1,...,y_5))=\sum_{i=1}^{5}0.2\ \left(1_{x_i=y_i}\right)+$$

$$\sum_{i=1}^{4}0.1\ \left(1_{x_i=y_{i+1}}+1_{y_i=x_{i+1}}\right)+\sum_{i=1}^{3}0.05\ \left(1_{x_i=y_{i+2}}+1_{y_i=x_{i+2}}\right)$$

Where *x* and *y* are the sequences of the 5 most frequent words. *s* is the sum of the score we give each word according to its position in the series (i.e. 0.2 if they are the same token and present the same position (like *de* at the example below); 0.1 if they are in the same position +/−1 (like *que*); and 0.05 if they match in position +/−2 (like *el*)) so that we get a result between 0 and 1, where 1 means total coincidence (in token and position) and 0 means null coincidence.

For example, the following sequences:

Text 1 > *de, que, y, el, la*

Text 2 > *de, el, que, y, en*

would have a value of $s = 0.45$

Once extracted the index of comparing all the texts, (about 7,000 comparisons), we assign to each comparison a value of 1 if comparisons were made between texts written by the same author and 2 if the compared text are from different authors. Through this measure it can be established whether they are different populations or whether there are only slight differences between the values of 1 and 2.

Figure 4 shows the results of the ANOVA of the similarity index outcome. The p-value representing the distance (or similarity) between the two populations is <0.00, which means that there are two well-differentiated populations. On the other axis we have the values of comparisons of texts written by same author (1), that have higher values than, comparisons of texts written by different authors (2).

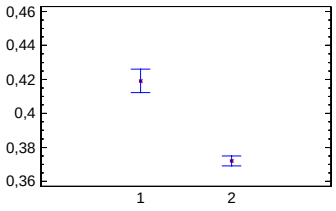


Figure 4 – Similarity index results

Once we had finished this experiment, we decided to repeat the whole operation with the 10 most frequent words to see if better results could be obtained. The results were already positive (p-value <0.00) but the distance between the two dispersions was relative smaller. That is due to the fact that we took fewer function words and more content words, so the dispersion is higher, too big to provide more accurate information.

Those results allow us to confirm that authors really tend to be recurrent at the level of use of the most frequent words. This fact is, probably, due to the functionality of the most frequent words: they can have more than one grammatical category (as *que* that can be a conjunction or a pronoun), most of them are also semantically ambiguous (as *de*). But the most critical characteristic of the most frequent words is the text content and the text length independency, that allow us to compare texts from different extensions and different contents.

3.5 Combination of all the variables

The variation scales for each author and variable have been obtained with the results shown so far. As a first conclusion it can be said that authors tend to follow similar patterns for the different variables, that is, for instance, when there is a variable with a much wide dispersion (as in the case of adverbs ending in *-mente*) it is evidenced in all the authors. Such findings facilitate our task since it proves that variation is associated not only to the authors but also to the variables selected to analyze their style. That is what we have called vertical plane variation.

4. Text comparison

When the scales for each variable have been determined, the second part of our experiment, that is the comparison of two texts for authorship attribution, starts. The calculated range for each style characteristic allow us to measure the variation intra-author so that, when comparing two text, some punctuation can be gives on depending if the range of every variable exceedes the normal dispersion or is under the limits that has been observed as the maximum dispersion of one author's style. In other words, the measure of the range allows us to measure the possible variation of the idiolect, and throw it, to decide if the text can be produced for the same author or independently.

That is the reason why, when comparing two texts, all the variables will be extracted and their value ranges calculated. These estimations will determine whether the style of the analyzed texts keeps within the faithfulness limits of an author; if it does not, the probability of attributing those texts to the same author would be low. According to Sanders' *principle of choice*, with this operation we measure each author's faithfulness in that choice. By studying all the variables in various authors, we expect to establish a maximum variability threshold of an author to determine the extent of recursiveness in the principle of choice.

As an example, taking the variable of the percentage of hapax legomena (Figure 2) we can determine that the maximum variation that an author can have is around the 10 points, although the most of the authors at the articles presented have a variation around the 5 points. Following this observation we assign to every comparison between two texts a value: we give 0 points if the difference between the values from both texts is bigger than 10 points, 1 point if the difference is between 10 and 5, and 2 points if the difference is less than 5. And we repeat this operation for the 10 variables we have assigning a different scale value for each variable. So that when comparing two texts we will have different comparison values indicating the distance between two texts.

Figure 5 shows the results of the analysis of variance (ANOVA) comparing the texts (200 comparing texts from were done, 100 from the same author and 100 from different authors) according to the variables extracted until April 2005: Type/Token Ratio, Percentage of Hapax Legomena, Percentage

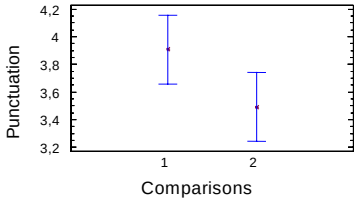


Figure 5 – Results of text comparison with 3 variables

Figure 6 shows the results for the same experiment but in this case 5 different variables were analyzed: Type/Token Ratio, percentage of Hapax Legomena, percentage of function words, sentence length and word length. As it can be noticed, dispersions do not overlap and there is a clear tendency in the comparisons of 1 (texts by the same author) to be superior to the comparisons of 2 (texts by different authors). The p-value for this analysis is <0.00 (against the 0.06 from the analysis shown at Figure 5). So that we can deduce that we are before two well differentiated populations. It is expected that when combining all the style characteristics that provide some information about the author (about 40) better results can be reached.

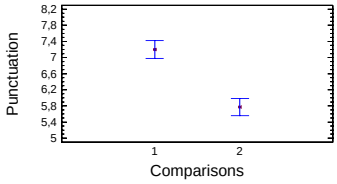


Figure 6 – Results of text comparison with 5 variables

5. Conclusions

In this article, a proposal to measure an author's style faithfulness has been presented. Halliday's conception about the structure of language and Sanders' ideas on stylistics offer a suitable theoretical framework for the objectives of this study. Our goals were to describe and to measure both such a general concept as 'style' and the intra and inter-authorial variation in order to make a transfer to a more abstract level from the analysis of the language structures of two texts to the structures of thought so that they help to determine the authorship of a text.

The project methodology, the first experiments carried out (with function words, lexical frequency, adverbial typology, etc.) and the first results are shown in this paper. Observing the work done, we can confirm that authors really tend to make the same choices from the variety of options that the language system provide them. The results suggest that even though there is still much work to do, we are on the right path.

6. Aknowledgements

This work would not have been possible without the support received from my family, friends and colleagues. Particularly, I would like to thank Rogelio Nazar for his helpful guidance in Perl programming, Jaume Llopis for his support in the statistical analyses and Diego Burgos for his help in the translation of this paper into English.

7. References

Dugast, D. (1980). *La statistique lexicale. Travaux de linguistique quantitative*, 9. Editions Slatkine, Genève.

Halliday, M.A.K. (1978) Language as a Social Semiotic. The Social Interpretation of Language and Meaning. Open University Set Book, London.

Holmes D. (1996). A Stylometric Analysis of Conversational Speech of Aphasic Patients. University of the West of England, Bristol, UK. Literary and Linguistic Computing 11(3):133-140.

Sanders, W. (1977) *Linguistische Stilistik. Grundzüge der Stylanalyse sprachlische Kommunikation.* Kleine Vandenhoeck-Reihe, Göttingen.

Stamatatos *et al*. (2000). Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics, 26(4), December 2000, pp. 471-495.

Ullmann, S. (1968). *Lenguaje y estilo*. Colección cultura e historia, editorial Aguilar 1977, Madrid.

Identifying Criteria to Automatically Distinguish between Scientific and Popular Science Registers

Lorraine Goeuriot Estelle Dubreil Béatrice Daille Emmanuel Morin
Université de Nantes, LINA - FRE CNRS 2729
2 chemin de la Houssiniere, BP 92208, 44322 Nantes Cedex 3, France
{lorraine.goeuriot,estelle.dubreil,beatrice.daille,emmanuel.morin}@lina.univ-nantes.fr

ABSTRACT
Stylistic analysis includes the automatic identification of the register of a document such as legal, literary...
In this paper we propose a typology of criteria to distinguish between science and popular science, based on observations of Web documents, forming a part of a comparable corpus. This typology holds two dimensions: the external dimension, characterizing the web site creation context, and the internal one, characterizing the communicative tasks of the website and its documents (graphical and textual characteristics). Then the criteria of the typology will be implemented to automatically classify Web documents between scientific and popular-science register. That is a first step to the automate of a comparable corpus of Web documents.

Keywords
Comparable corpora, scientific register, popular-science register, web genres, web categorization

1. INTRODUCTION
The use of corpora for the automatic compilation of monolingual, bilingual, or multilingual dictionaries is an important trend in Natural Language Processing (NLP); see for example research in the compilation of bilingual dictionaries from parallel texts [13]. However, the use of parallel corpora raises two problems:

- Because a parallel corpus is pairs of original and translated texts, the vocabulary appearing in the translated text is strongly influenced by the source text, especially in technical domains;
- Translated texts are artefacts that neutralize the social and cultural gap between the source and the target language. This is higly adverse in many cases where languages are typologically and culturally different.
- Such corpora are difficult to obtain for pairs of languages not involving English.

New methods try to make use of comparable corpora. Bowker and Pearson [5, p.93] worked on the use of corpora, and defined comparable corpora as follows:

> "Comparable corpora consist of sets of texts in different languages that are not translations of each other."

They use the term *comparable* to

> "indicate that the texts in the different languages [...] have some characteristics or features in common."

Topic, subject, period, type of text are examples of features. Most of the works that use or compile comparable corpora are based on the same definition [7], [14]. In some papers, the notion of *comparability* is developed through that of *degree of comparability* which depends on the common features of the text [6].

The web provides ample amounts of documents which can be used to compile comparable corpora. Monolingual web documents are linked to their social and cultural environment, and should present distinctive features, at least for different languages, subject areas and registers.

The purpose of our study is to automate the compilation of comparable corpora through the categorization of web pages. Web page categorization is presently an active research domain. The amount of data on the Web is exponentially increasing, and a lot of work has been done to integrate categorization with web search. Recent works [11], [10] try to extract information from the structure of web documents (*HTML* code, images, videos, meta data, etc.). Automatic categorization is based on different kinds of categories, such as web genre[1].

In this paper, we present our research on automatic categorization of French corpus documents between scientific and

[1]E.g. information pages, research pages and personal home pages in [11]

popular science registers. In section 2, we describe the compilation of a French corpus from documents on the Web. In section 3, we present and discuss our categorization criteria. In section 4, we state our conclusions and make suggestions for further study.

2. CORPUS COMPILATION

In this section, we focus on the compilation of the French monolingual part of the comparable corpus. The documents collected for the corpus share the common subject, *"Diabète et alimentation"* ("Diabetes and nutrition"), more specifically *"Diabète et obésité"* ("diabetes and obesity") which both interests medical professional and general public. We want our corpus to be representative of the full range of linguistic possibilities in French language. That is why we excluded French language documents not from France (Belgian, Swiss, Quebec French...).

In this paper, we distinguish the terms "website" and "web page" (or "web document"): a web page is an element of the website hierarchy. The documents we collect are web pages of specific websites.

2.1 Documents research

Documents for the French corpus are extracted from the web. We use three techniques to find them :

1. National web search engines[2], using keywords;
2. Local search engines (in the internal documents of a site), using keywords;
3. Surfing and link collections.

The first two techniques require keywords. The combination of the used keywords should adequatly represent the subject of our corpus. We have three main keywords: *diabète, alimentation* and *obésité* (diabetes, nutrition and obesity). In order to collect more documents, we needed to widen our set of keywords. Two supplementary methods were used to accomplish this: an external and an internal approach. First, we used a thesaurus (synonym finder)[3], which gave us a set of synonyms for each keyword. We noticed that this method produced only some equivalent terms and can not provide us other semantically linked terms. To resolve this problem, we complemented our set with terms occurring in the documents. These terms are not necessary occuring with a high frequency, but are used in the same context as our keywwords. We name these terms *equivalent terms*. Figure 1 presents *obesity* synonyms and equivalent terms.

We underlined words found in the document. It shows that thesaurus do not make the difference between employed words and others. A manual selection is still necessary.

2.2 Document formats

Websites contain documents in various formats (of which *HTML* is just the most widespread). We classified these formats in two categories:

[2]Such as Google France : `http://www.google.fr`
[3]`http://elsap1.unicaen.fr/cgi-bin/cherches.cgi` (French)

Synonyms	Equivalent terms
adipose	excès de poids
adiposité	surpoids
bouffissure	diabésité
engraissement	surplus de poids
graisse	perte/prise de poids
grosseur	kilos (en trop)
lipomatose	surcharge pondérale
polysarcie	excès pondéral
corpulence	poids excessif
embonpoint	rondeurs
rotondité	excès de masse grasse

Figure 1: New keywords corresponding to *obésité* (obesity)

- Documents interpretable by a web browser: HTML, PHP, SHTML, ASP, etc.;
- Documents not interpretable by a web browser: PDF, PS, MS Word, Powerpoint, etc.

It must be possible to convert our documents to simple text, and this text should be representative of the French lexicon, syntax, etc., as well as of the subject of the corpus. (The absolute majority of collected documents satisfied these requirements.)

2.3 Web genres

The texts we have found correspond to Beauvisage's [2] classification of websites:

- informal, private: personal home pages;
- public, commercial: home pages for the general public[4];
- interactive pages: pages with feed-back: searchable indexes, customer dialogue;
- journalistic materials: press: news, editorials, reviews, e-zines;
- reports: scientific, legal and public materials; formal text;
- other texts;
- FAQs;
- link collections;
- other listings and tables;
- discussions;
- error messages.

[4]In this genre we inventory sub-genres such as commercial, institutional, association, etc.

As we were compiling a corpus, some of these categories were not interesting for us. Interactive pages, for example, are not textual documents, as well as link collections, tables and error messages.

Figure 2 presents our corpus documents repartion into the holded web genres.

Genre	Scientific documents	Popular science documents
Informal, private	1	43
Public, commercial	9	80
Journalistic material	20	65
Reports	30	9
Other texts	7	8
FAQs	0	0
discussions	0	30

Figure 2: Documents repartition into the web genres

The most frequent kinds of documents belong to the categories *informal/private*, *public*, and *journalistic*. Reports are plenty in institutional websites (academic, government, associations, hospitals). We found many discussions for diabetics, but we did not keep everything (we only need a sample of this particular discourse). The "other texts" category includes: course notes (lectures), reviews of books, pages offering advice on nutrition, etc.

The most useful websites are portals: huge websites, incorporating search engines, large amounts of documents, collections of links, discussions, lexicons (with specialized vocabulary), etc.

We stopped collecting as it became hard to find new documents: we explored the twenty first results returned by web search engines, and stopped when they contained no new documents, whatever the combination of keywords.

We set the desirable size of the corpus at 400 000 words (200 000 in each register). We actually reached 300 000 words for popular science and 140 000 for science, which represent about 20 Mo, and 250 web pages.

3. STYLISTIC ANALYSIS

As we have said in the previous section, the trilingual corpus is compiled in order to elicit register classification criteria (proper to French documents) that should be used to automatically categorise web documents.

First of all, let us define what is considered as popular-science and scientific documents. According to Mortureux [1], popular-science discourse is a discourse involved in specialized knowledge diffusion ("discours intervenant dans la diffusion de connaissaces spécialisées"). It is hard to find a general definition of scientific discourse. We found three major different ways to define it. First, scientific discourse is composed of specialized terms (terms which belong to a specialized scientific vocabulary). Second, scientific documents are defined as having a restricted subject, oriented toward people having an implicit knowledge of this subject [8]. Third, scientific documents are defined in terms of register.

3.1 Genre, register and style

In France, the term "genre" is generally used to categorize different kinds of literary texts (novel, comedy, etc.). In view of the recent developments of discourse analysis, "genre" came to refer to more general phenomena. [3] [4] Genres are text categories distinguished by matured speaker, for example genres for English are: Intimate personal interaction, Informational interaction, Scientific exposition, Learned exposition, Imaginative narrative, General narrative exposition, Situated reportage, Involved persuasion.

Since 1995, Biber uses the term "register" for this broader conception of genre. However, we are going to abide by the previous term because we keep a limited conception of the typology. However, we use the term "register" to distinguish between the different kinds of discourse (in the sense of Rastier [9]), for example legal vs. literary vs. political vs. scientific. This notion of register is quite vague, but makes sense in our stylistic analysis.

In this paper, we define *stylistics* as the inventory and the analysis of the variable characteristics proper to each register (science or popular science): "the presence or absence of some of a large range of structural and lexical features" [12]. We expect that our stylistic analysis will product a new typology in terms of sub-genre.

3.2 Classification criteria

Our criteria come from two techniques: first we studied works about categorization and classification [15] and then update them while collecting our documents (observations). While we were compiling our corpus, each document was manually classified to its proper register. We could not base our work on definitions above. The following criteria helped us.

We distinguished two dimensions: the first one, the external dimension, relates to the physical characteristics of the websites of documents. The second one, the internal dimension, relates to the websites and mostly web documents features (appearance, text).

3.2.1 The external dimension

The external dimension corresponds to a set of criteria pertaining to the website creation context. These criteria concern the website that contains a given document, and include:

- the website location (URL);
- document formats in the website;
- the architecture of the website ;
- the size of the website;
- the origin of the website.

3.2.2 The internal dimension

The internal dimension corresponds to a set of criteria pertaining to the communicative tasks of the document and of its website. We divided this category into 3 sub-categories: graphical characteristics, structural characteristics and semantico discursive characteristics.

Graphical characteristics

Graphical characteristics are the apparent components of a website and the appearance of its documents:

- frames;
- colors;
- images, animations;
- advertising.

Structural characteristics

These are the characteristics of the textual structure of a document:

- the title (existence, font, color, position, length, punctuation);
- presence of keywords, an abstract, a plan, a bibliography;
- presence of an introduction and/or a conclusion;
- the appearance of the main text (length, typography, punctuation, numeric characters, links, sections and subsections, references, citations);
- name(s) of author(s);
- examples.

Semantico-discursive characteristics

These are the stylistic and linguistic characteristics of the document. This category includes:

- pragmatic criteria: narrative devices, logical connectors, paraphrases;
- characteristics of sentences: sentence types, length, punctuation, noun phrases;
- lexical criteria: specialized vocabulary, collocations, morphology, terminological density, proper nouns, initials, unknown words (specialized terms or spelling mistakes), numeric characters, symbols, word length.

3.3 Observations

In the course of our first experiments, we surveyed almost fifty documents (half of them scientific, the other half popular) according to the above criteria. Some of them seemed as an effective means to discriminate scientific from popular-science documents.

External criteria

Popular: Most of popular-science documents occur in personal websites, although medical websites also provide popular-science information for diabetics.

Scientific: Most of scientific documents occur in institutional websites (academic, government, etc.) or in websites for medical professionals. The majority of *PDF* documents are scientific.

Internal dimension

Popular: Popular-science documents are highly colored, we can find images everywhere, especially in margin frames. Often there is advertising, also located in margin frames.

Titles are short (2 or 3 full words), colored and sometimes have interrogative form. We have never found any keywords, bibliography, or abstract. Introductions and conclusions are rare, and often tally with the characteristics of the journalistic style. The main text has variable size, and soft colors. It can be well stretched out, as well as compact; short, as well as long. It is often broken into paragraphs, and it often contains lists of item and sometimes, links. It is set in the default font. It often contains *HTML tables*, and rarely the name of the author (at the end of the text).

Vocabulary is quite basic and euphemistic (especially where it concerns obesity). There are many paraphrased or reformulated sentences, and questions are often used. One particular case is discussions which employ pronouns like "I" and "you". Such texts are adressed to diabetics.

Scientific: Scientific documents are less colored. The background color is often white, and one highlighting color is used1 (blue in the majority of cases). They contain images, especially in margin frames (mostly the top one) and rarely in the text. Very frequently, there is a logo located in the top frame (link to the home page). Advertising is rare.

The main text always has a title which, most of the time, is aligned on the left. The title is quite small and colored. Its length is variable (from 1 to 15 full words). Sometimes the text is preceded by a plan, less frequently by an abstract. Depending on the genre of the text, an introduction and a conclusion may be present. The text is set in the default font and is quite compact and long. It is always broken into sections and paragraphs. Sentences are lengthy. Texts contain few itemized lists and few links, but often are provided with references. *HTML tables* are often present and contain numeric characters. A bibliography can sometimes be found at the end of the text, and often there are the name(s) of the author(s) (sometimes with their academic or occupational title).

The vocabulary is restricted, contains scientific (medical) terms and acronyms. We have found characteristical features of scientific discourse: symptom descriptions, words like "patients" (patients), "malades" (sick persons), etc. The text is in delocutive ways form. This means that such texts talk about diabetics, but never adress them directly. The tone is very objective, diabetes and obesity are only considered as pathologies. Interrogative forms are very rare.

3.4 Application of criteria

In order for our set of criteria to be more effective, it has to be filtered to preserve only operational and discriminatory criteria.
A criterion is said to be operational if it can be implemented in a program. It is called discriminatory if its variation characterizes a register.
Unfortunately, the criteria corresponding to our requirements are few. None of our criteria (§3.2) is completely characterizes a register. That is why we will have to examine combinations of different criteria to find the right register.

One combination, called *score* will evaluate the degree of membership of the document in the scientific or popular science register. However, to evaluate the score, we need to select some criteria according to their operationality and their discriminatory effect.
The score s is the sum of criteria weighted according to their discriminatory effect:

$$s = \sum_{i=1}^{n} w_i \cdot c_i$$

where n is the number of criteria, w_i is the weight of the i^{th} criterion, and $c_i = 1$ if the i^{th} criterion is present and 0 otherwise.

A threshold should be determined: if the score of a given document is above the threshold, then it is scientific; else it is popular science.

Meanwhile, there is a problem with this notion of threshold: what about documents whose score is immediately around the threshold? Indeed, we noticed in our discussion of stylistic analysis (§3.2) that these registers cannot be well defined. The previous definitions show us how subjective they are. This arbitrariness of threshold points out that there is a continuum between scientific and popular-science documents. Actually, we can not venture on the task of automatic classification without manually evaluating ambiguous documents.

4. CONCLUSION

In this paper, we have presented our work on register categorization of web pages and have proposed a new typology of web document registers. We have compiled a monolingual comparable corpus in the medical domain centering around the topic *diabetes, nutrition and obesity*. This corpus was manually built in order to elicit the criteria distinguishing scientific from popular-science register. In this task, we use the text of documents as well as their structure. The criteria have two dimensions, internal and external. It is our plan to integrate these criteria in the implementation of an automatic categorization between the two registers, with the final goal to automate the compilation of comparable corpora.

5. FUTURE WORK

The notion of *score* presented in section 3.4 has to be analysed to alleviate the continuum effect between the scientific and the popular-science registers. In perspective, the automatic recognizer will be integrated into a program for comparable corpora compilation. Our goal is to extend these tools to other languages, in the first place, to Japanese and Russian.

6. ACKNOWLEDGMENTS

This research program has been funded by CNRS as part of the TCAN program for 2004. We thank Boris Smilga for his help in preparing the English text of the article.

7. ADDITIONAL AUTHORS

8. REFERENCES

[1] Comment définir la propriété d'un mot? Berne, Peter Lang, 1994.

[2] T. Beauvisage. Morphosyntaxe et genres textuels. TAL, 2001.

[3] D. Biber. A typology of english texts. Linguistics, 1989.

[4] D. Biber. Representativeness in corpus design. In A. Zampolli, N. Calzolari, and M. Palmer, editors, Current Issues in Computational Linguistics: Essays in Honour of Don Walker, pages 377–407. Giardini Editori e Stampatori and Kluwer Academic Publishers, Pisa and Dordrecht, 1994.

[5] L. Bowker and J. Pearson. Working with Specialized Language: A Practical Guide to Using Corpora. London/New York: Routeledge, 2002.

[6] H. Déjean and E. Gaussier. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. Lexicometrica, 2002.

[7] P. Fung. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora, Mar. 08 1998.

[8] M. D. Heaulme, M. Membrado, S. Ameli, and F. Vexler. Ambiguïté et paraphrase dans le langage médical et leur traitement par translog. In l'Ambiguïté et la paraphrase.

[9] D. Malrieu and F. Rastier. Genres et variations morphosyntaxiques. Texto !, 2002.

[10] J. M. Pierre. Practical issues for automated categorization of web sites, June 01 2000.

[11] A. Prakash, A. Kranthi, and K. Ravi. Web page categorization based on document structure, Dec. 23 2001.

[12] J. Sinclair. Preliminary recommendations on corpus typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards), 1996.

[13] J. Veronis, editor. Parallel Text Processing. Kluwer Academic Publishers, 2000.

[14] F. Zanettin. Bilingual comparable corpora and the training of translators. Meta, Vol. 43, numéro 4, 1998.

[15] P. Zweigenbaum and N. Grabar. Learning derived words from medical corpora. In 9th Conference on Artificial Intelligence in Medicine Europe.

Style versus Expression in Literary Narratives

Özlem Uzuner and Boris Katz
Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, 32-252
Cambridge, MA 02139
{ozlem,boris}@csail.mit.edu

ABSTRACT

Style and expression are related: they both refer to linguistic elements people use while conveying content. However, style refers to the linguistic choices of authors that persist over their works, *independently* of content while expression refers to the way people convey *particular* content. Style helps us identify the works of a particular author for authorship attribution; expression helps us identify a unique work for copyright infringement detection.

The differences between expression and style are more than qualitative. In this paper, to expose the differences of expression and style in more concrete terms, we present computational definitions for each. These definitions show that style can be adequately captured in terms of syntactically uninformed features such as function words. However, these features are not sufficient for capturing the expression of content that is unique to a particular work of an author. In order to recognize individual works from their expression of content, we need syntactic information.

1. INTRODUCTION

Style refers to the linguistic choices of authors that can identify their writings even when these writings vary in content; information about authors' styles have frequently been used in the text classification literature for authorship attribution. Expression refers to the way people convey particular content; identifying expression is especially useful for copyright infringement detection. In this paper, we present a comparative study of expression and style; we evaluate the strengths and weaknesses of various sets of features for identifying each.

In particular, we study two corpora consisting of novels in order to capture the computational difference between expression and style. Our first corpus includes books that have been translated into English from foreign originals. We refer to the original works as the *titles* and the translations of these titles as *books*; translations of a title help us create a computational definition of expression. Our second corpus contains multiple books by each of eight authors and helps us create a computational definition of style.

The statistical tests and classification experiments presented in this paper provide a way of quantifying the qualitative and functional difference between expression and style. Our experiments show that syntactic features, which we call syntactic elements of expression, are more successful at recognizing expression than style, whereas function words are more successful at recognizing style than expression.

2. RELATED WORK

Linguistic similarity between works has been studied for identifying the style of an author in terms of a variety of features, including distribution of word lengths [14, 24] and sentence lengths [16, 17, 18, 24], distribution of function words [13, 14], and measures of richness of vocabulary [6, 19]. Overall, both linguistically uninformed features, e.g., sequences of letters [8, 10], and linguistically more informed features, e.g., syntactic classes (part of speech) of words [5, 9] and their ngrams [4], have been successfully used for capturing an author's style.

Expression and style are both based on linguistic elements of authors' writings. Which linguistic features are more useful for identifying expression and which are more useful for style depends on the group of authors and works that are studied. But in general, different groups of features would be used to define an author's overall style and to define his unique expression in a work. For example, if an author always uses long sentences, his style can partly be described in terms of the length of his sentences; however, this information is not enough for capturing expression as it does not indicate which work is copied. On the other hand, the author may use predominantly left-embedded sentences in one work and predominantly right-embedded sentences in another. This information can be used to capture the different expressions of his works, but would not help define his style.

To capture expression, we need to identify the linguistic choices of authors that are unique to a work, and that differentiate it from the expressions of other authors who write about similar content as well as the expression of other content by the same author. To capture style, we need to identify the linguistic choices of authors that occur independently of the content, in all works of authors, within the same genre.

3. TOWARDS DEFINING EXPRESSION: STYLE FEATURES AND EXPRESSION

In this paper, we study the linguistic choices of authors in order to create computational definitions of expression and style. Style has been previously studied in stylometry and authorship attribution; however, expression is a new concept which we study for copyright infringement detection [22].

Various feature sets have been used in the text classification literature for stylometry and authorship attribution. These features mostly relate to the way people write; they capture the elements of an author's writing that reflect his style. Expression also depends on the way people write. Therefore, in order to select a set of features that capture expression, we first studied those existing features frequently used in the authorship attribution literature [5, 21]. Our goal was to evaluate these features for their contribution to differentiating between different expressions of the same content, to identify the features that hold promise for capturing expression, and later to explore these promising features in more detail to create a computational definition of expression.

Initially, we analyzed a set of surface, syntactic, and semantic features obtained from authorship attribution literature [5], evaluated the promise of each feature in this set for capturing expression using classification experiments and significance tests, ranked these features based on their contribution to identification of expression, and identified the features that needed to be studied in more detail to:

- capture expression, and
- compare expression and style.

The features we studied included:

- Baseline surface features:
  - Number of words in the document;
  - Type–token ratio, i.e., the ratio of the total number of unique words in the document to the number of words in the document;
  - Average and standard deviation of the lengths of words (in characters) in the document;
  - Average and standard deviation of the lengths of sentences (in words) in the document; and
  - Number of sentences in the document.
- Baseline syntactic features:
  - Sentence type:
    - Frequency of declarative sentences, i.e., constructs that follow the subject–verb–object pattern;
    - Frequency of interrogatives, i.e., constructs that exhibit subject–auxiliary inversion, sometimes accompanied by wh-phrases, e.g., what, which, who, why, etc., and appropriate punctuation, as well as wh-questions that do not exhibit subject–auxiliary inversion;
    - Frequency of imperatives, i.e., constructs that start with an imperative verb and do not have an explicit subject;
    - Frequency of fragmental sentences;
  - Voice:
    - Frequency of active voice;
    - Frequency of be-passives, i.e., passive constructs that use "be", e.g., "I was robbed";
    - Frequency of get-passives, i.e., passive constructs that use "get", e.g., "I got robbed";
  - Genitive use:
    - Frequency of 's-genitives, i.e., possessive "'s" observed in the "*noun*'s *noun*" construct;
    - Frequency of of-genitive, i.e., possessive "of" observed in the "*noun* of *noun*" construct; and
    - Frequency of noun phrases that do not include genitives.
- Baseline semantic features:
  - Frequency of overt negations, i.e., explicit negations such as "not", "no", "nowhere", "no one", "none", and several others; and
  - Frequency of uncertainty markers, i.e., words like "can", "could", "maybe", "may", "kinda", "probably", "possibly", etc.

3.1 Significance Testing for Feature Ranking

We studied the contribution of each of the baseline surface, syntactic, and semantic features to recognition of expression by focusing on the expressive differences that are solely due to the way people write content. For this purpose, we ran classification experiments and statistical tests on the distributions of mean cross-validation accuracies obtained from these classification experiments in order to rank the baseline features based on their contribution to identification of expression given similar content. Our data set for this experiment consisted of pairs of translations of the same title [2] which provided us with examples of the same content that differed in expression. These pairs served as our surrogate for infringement data. The particular translations used in this experiment included two translations of *20000 Thousand Leagues Under the Sea* (Verne), three translations of *Madame Bovary* (Flaubert), and two translations of *The Kreutzer Sonata* (Tolstoy).

In particular, we used the baseline surface, syntactic, and semantic features to obtain cross-validation accuracies on pairwise classification experiments [15] which included one experiment for differentiating between the two translations of *The Kreutzer Sonata*, one for differentiating between the two translations of *20000 Leagues under the Sea*, and three for differentiating between the three translations of *Madame Bovary*, i.e., compare translation 1 against translation 2, translation 1 against translation 3, translation 2 against translation 3. Next, we ran statistical tests on the distribution of mean cross-validation accuracies of these pairwise classification experiments: we obtained the distribution of average cross-validation accuracies in the presence of all n features. Next, we reran the same experiments with all possible subsets of $n-1$ features. We calculated the significance of the differences of the resulting distributions of average cross-validation accuracies (using $n-1$ features) from the original distribution (using n features).

To obtain a ranking of the features, we used these significance values and eliminated (one at a time) the features whose absence least affected the distribution of mean cross-validation accuracies, i.e., the features with the highest p-value. We repeated this process until only one feature was left.

This method ranked syntactically more informed features, such as the frequency of use of "get-passives", the frequency of use of "'s-genitives", and the frequency of use of "be-passives" in the top five. These and the remaining top ten most useful features thus identified are shown in Table 1.

Rank	Feature
1	Standard deviation of sentence lengths
2	Frequency of use of "get-passives"
3	Frequency of use of "'s-genitives"
4	Standard deviation of word lengths
5	Frequency of use of "be-passives"
6	Frequency of active voice sentences
7	Frequency of use of declaratives
8	Frequency of overt negations
9	Type–token ratio
10	Number of sentences in the document

Table 1: Ten most useful features for distinguishing between translators who translated the same content.

In general, the high rank of the syntactically more informed features, such as passives and genitives, when differentiating between the translations of the same original is expected: given the content, translations can be differentiated based on the way they are written, and this can be captured by analyzing syntax.

4. SYNTACTIC ELEMENTS OF EXPRESSION

Syntax plays a significant role in the way authors express content. For example, consider the following semantically equivalent excerpts from three different translations of *Madame Bovary* by Gustave Flaubert.

> Excerpt 1: "Where should he practice? At Tostes. In that town there was only one elderly doctor, whose death Madame Bovary had long been waiting for, and the old man had not yet breathed his last when Charles moved in across the road as his successor." (Translated by Unknown1.)
>
> Excerpt 2: "Where should he go to practice? To Tostes, where there was only one old doctor. For a long time Madame Bovary had been on the look-out for this death, and the old fellow had barely been packed off when Charles was installed, opposite his place, as his successor." (Translated by Aveling.)
>
> Excerpt 3: "And now where was he to practice? At Tostes, because at Tostes there was only one doctor, and he a very old man. For a long time past Madame Bovary had been waiting for him to die, and now, before the old fellow had packed up his traps for the next world, Charles came and set up opposite, as his accredited successor." (Translated by Unknown2.)

These excerpts differ in their use of vocabulary and syntax in several ways. For instance, they ask the question "where should he practice" in three different ways, using three different verb phrase structures exemplified by "practice", "go to practice", and "to practice"; they also explain the presence of the old doctor using three different sentential structures: "there was only one elderly doctor", "there was only one old doctor" and "there was only one doctor, and he a very old man". We captured some of these differences by studying syntactic features that relate to how people convey content, i.e, syntactic elements of expression. These features included:

- Distributions of various phrase types in sentence-initial and -final positions which can capture expressive differences at a very high level. This level of analysis identified, for example, the syntactic difference in sentences "Martha can finally put some money in the bank." and "Martha can put some money in the bank, finally.".
- Distributions of semantic classes of "non-embedding" verbs in documents [7, 11, 20, 22] which we identified using Levin's taxonomy of semantic verb classes [11]. We coupled this with information about the argument structure of the observed verbs in terms of the phrase-level constituents, such as noun phrases arguments and prepositional phrase arguments.
- Distributions of syntactic classes of "embedding" verbs in documents obtained using the taxonomy of embedding verb classes by Alexander and Kunz [1]. These syntactic classes are given in terms of phrasal and clausal elements, such as verb phrase heads (Vh), participial phrases (Particip.), indicative clauses (IS), subjunctives (Subj.), and small clauses (SC). For our studies, we used 29 such verb embedding classes and identified the distributions of these embedding classes in different works [20, 22].
- Linguistic complexity of sentences measured in terms of:
  - The mean and the standard deviation of the depths of the top-level left and right branches in sentences in terms of phrase depth.
  - The mean and the standard deviation of the number of prepositional phrases in sentences, as well as the mean and the standard deviation of the depths of the deepest prepositional phrases in sentences.
  - The percentage of left-heavy, right-heavy, and equal-weight sentences, e.g., sentences where the top-level right branch of the syntax tree is deeper than the top-level left branch are considered right-heavy.
  - The mean and the standard deviation of the number of embedded clauses in the top-level left and right branches in sentences.
  - The percentage of left-embedded, right-embedded, and equally-embedded sentences, e.g., sentences where the top-level right branch of the syntax tree embeds more clauses than the top-level left branch are considered right-embedded.
  - The mean and standard deviation of the depths of sentence-initial subordinating clauses in sentences.

All of these features are extracted from part-of-speech tagged text [3], using context-free grammars. More details about these features and examples of each can be found in [20, 22, 23].

4.1 Analysis of Text

Given sentence-initial and -final phrase structures, semantic classes of verbs and their argument structures, syntactic classes of verbs and their embeddings, and linguistic complexity features, i.e., syntactic elements of expression, studying the excerpts presented in Section 4 focuses our attention on the fact that:

- All of excerpts 1, 2, and 3 contain question constructs.
- The majority of sentences in all of these excerpts start with prepositional phrases, e.g., *"To Tostes"*, *"At Tostes"*, and *"For a long time"*.
- All three excerpts contain one sentence (an interrogative sentence) that ends with a verb phrase, i.e., *"practice"*; however, the majority of the sentences in all of the excerpts end with noun phrases, i.e., *"Tostes"*, *"his successor"*, *"only one old doctor"*, *"very old man"*, etc.
- Excerpt 1 includes three sentences, one of which is a fragment and does not include any subject–predicate pairs, i.e., *"To Tostes"*. Of the other two sentences, one contains only one pair, i.e., *he–practice*, and the other contains four pairs, i.e., *there–was...* where the subject is the existential "there" and the predicate is the verb phrase starting with "was", *Madame Bovary–had...been...*, *the old man–had...breathed...*, and *Charles–moved...*. All of these clauses have deeper right branches than left branches. Excerpts 2 and 3 can be analyzed similarly.
- Excerpts 1 and 2 include relative clauses marked with wh-words, e.g., *"...whose death Madame Bovary had been waiting for..."* and *"...where there was only one doctor."*
- Excerpt 2 uses passive voice, e.g., *"...been packed off..."*
- Most of the verbs used in these excerpts are non-embedding, e.g., *"practice"*, *"be"*, *" breathe"*, etc.
- All excerpts use copular *be*, though at different rates.
- Some of the non-embedding verbs in excerpt 1 are intransitive, i.e., verb phrase structure is denoted by "V", e.g., *"practice"*. Others are observed in structures denoted by "V+NP" which indicates a verb followed by a direct object noun phrase, e.g., *was only one elderly doctor"*; "V+prep" which indicates a verb followed by a preposition and no noun phrase, and typically occurs as a result of movement of the noun phrase subsumed by the prepositional, e.g., *"waiting for"*; and "V+PP" which indicates a verb followed by a prepositional phrase, e.g., *"moved in across the road..."*. Phrase structures of the other excerpts can be analyzed similarly.
- The differences in sentence structures of excerpts 1 and 3 result in different observations associated with the use of the verb *"wait"* in these two excerpts. In excerpt 1, presence of the relative clause and movement of the object noun phrase of the verb results in the structure "V+prep"; whereas in excerpt 3, due to lack of movement, the verb appears in the structure "V+PP".

5. EVALUATION

We used the syntactic elements of expression to computationally describe expression. We hypothesize that these elements capture expression of content and provide information on how people convey particular content. To test this hypothesis, we evaluated the syntactic elements of expression on identifying expression of content in individual books (even when some books are derived from the same original title). Given the focus of these features on how people convey content, we also tested them on recognizing the style of authors.

5.1 Baseline Features

To evaluate the syntactic elements of expression, we used as baselines, features that capture content and features that capture the way works are written. These baselines included:

- Tfidf-weighted Keywords: We excluded from this set proper nouns; proper nouns can identify books without having to capture style or expression and are therefore omitted from this experiment.
- Function Words: We used a set of 506 function words that included the function words used in the studies of Mosteller and Wallace [13], as well as 143 function words that are more frequently used in modern English. The list of function words can be found in [20].
- Distributions of Word Lengths: We used distributions of word lengths, although literature presents conflicting evidence on the usefulness of this measure for authorship attribution [12, 24].
- Distribution of Sentence Lengths: We used sentence length distributions, means, and standard deviations [6] as baseline features.
- Baseline Linguistic Features: We compared the syntactic elements of expression also with the original set of baseline surface, baseline syntactic, and baseline semantic features (presented in detail in Section 3).

5.2 Classification Experiments

To evaluate the strength of different sets of features on identifying expression and identifying style, we studied these features on two separate experiments: recognizing books even when some of them are derived from the same title (different translations) and recognizing authors. For these experiments, we used boosted decision trees [25].

For both experiments, we created a balanced data set of relevant classes, using 60% of the chapters from each class for training and the remaining 40% for testing. Parameter tuning on the training set showed that the performance of classifiers (regardless of feature set) stabilized at around 200 rounds of boosting. In addition, limiting our features to those that had non-zero information gain on the training set eliminated noisy features [26].

5.2.1 Recognizing Expression (Recognizing Books)

For evaluating different sets of features on recognizing expression, we used a corpus of parallel translations. This corpus contained 45 titles and 49 books derived from these titles. In this context, *title* refers to an original work. Some titles are translated by different translators on different occasions; each of these translations provide us with a *book* that is derived from that title. Our corpus included multiple books for 3 of the titles (3 books derived from the title *Madame Bovary*, 2 books from *20000 Leagues*, and 2 books from *The Kreutzer Sonata*). Given that these books were translated independently from each other, they each contain their own expression of content. An accurate description of expression needs to capture this difference adequately.

The remaining titles in this corpus included literary works from Jane Austen (1775-1817), Fyodor Dostoyevski (1821-1881), Charles Dickens (1812-1870), Arthur Doyle (1859-1887), George Eliot (1819-1880), Gustav Flaubert (1821-1880), Thomas Hardy (1840-1928), Ivan Turgenev (1818-1883), Victor Hugo (1802-1885), Washington Irving (1789-1859), Jack London (1876-1916), William Thackeray (1811-1863), Leo Tolstoy (1828-1910), Mark Twain (1835-1910), and Jules Verne (1828-1905).

To test different feature sets for recognizing books, we ran classification experiments on a collection 40–50 chapters from each book in this corpus. We found that syntactic elements of expression accurately recognized books 76% of the time and that these features significantly outperformed all baseline features (see Table 2). Further analysis of the results showed that syntactic elements of expression accurately recognized each of the paraphrased books 89% of the time (see right column on Table 2).

Feature Set	**Accuracy (complete corpus)**	**Accuracy (paraphrases only)**
Syntactic elements of expression	76%	89%
Tfidf-weighted keywords	66%	88%
Function words	61%	81%
Baseline linguistic	42%	53%
Dist. of word length	29%	72%
Dist. of sentence length	13%	14%

Table 2: Classification results on the test set for expression recognition even when some books contain similar content.

The fact that syntactic elements of expression can differentiate between translations of the same title indicates that translators add their own expression to works, even when their books are derived from the same title; the expressive elements chosen by each translator help differentiate between books derived from the same title.

5.2.2 Recognizing Authors

Style and expression, though different, both relate to the way people convey content. Then, an interesting question to answer is: Can the same set of syntactic features help recognize both expression and style?

Stylometry literature uses corpora consisting of literary works that are written by native speakers of English, that are in the same genre, and that are written around the same time periods [9, 12, 13, 14, 24]. By controlling time period and genre, these corpora help expose the linguistic differences that are due to authors. In order to evaluate syntactic expression features on authorship attribution, i.e., identifying the works of an author by studying his style, we used a similarly controlled corpus. The books in this corpus included:

- Jane Austen (1775-1817): *Northanger Abbey*, *Emma*, *Sense and Sensibility*, *Mansfield Park*, *Lady Susan*, *Persuasion*, *Pride and Prejudice*.
- Charles Dickens (1812-1870): *A Tale of Two Cities*, *David Copperfield*, *Old Curiosity Shop*, *Oliver Twist*, *Pickwick Papers*, *The Life and Adventures of Nicholas Nickleby*.
- George Eliot (1819-1880): *Adam Bede*, *Middlemarch*, *Daniel Deronda*, *The Mill on the Floss*.
- Thomas Hardy (1840-1928): *The Mayor of Casterbridge*, *A Laodicean: A Story of To-Day*, *The Hand of Ethelberta: A Comedy in Chapters*, *Far from the Madding Crowd*, *Jude the Obscure*, *Tess of the d'Urbervilles: A Pure Woman*.
- Washington Irving (1789-1859): *Life and Voyages of Christopher Columbus Vol. II*, *Chronicle of the Conquest of Granada*, *Knickerbockers History of New York*.
- Jack London (1876-1916): *The People of the Abyss*, *Adventure*, *The Little Lady of the Big House*, *The Sea Wolf*, *The Cruise of the Snark*, *Michael, Brother of Jerry*, *Burning Daylight*, *The Iron Heel*, *The Mutiny of the Elsinore*.
- William Makepeace Thackeray (1811-1863): *Catherine: A Story*, *The Memoirs of Barry Lyndon, Esq.*, *The Great Hoggarty Diamond*, *The Newcomes: Memoirs of a Most Respectable Family*, *The Tremendous Adventures of Major Gahagan*, *The History of Henry Esmond, esq: A Colonel in the Service of Her Majesty Queen Anne*, *The Virginians: A Tale of the Eighteenth Century*, *The History of Pendennis*, *The Book of Snobs*.
- Mark Twain (1835-1910): *The Mysterious Stranger*, *A Connecticut Yankee in King Arthur's Court*, *The Adventures of Huckleberry Finn*, *Following the Equator: A Journey Around the World*, *The Gilded Age: A Tale of Today*, *Those Extraordinary Twins*, *Christian Science*, *The Adventures of Tom Sawyer*.

Authors can be distinguished from other authors based on the way they write, independently of content. Therefore, for authorship attribution, we use as baselines only the features that capture the way authors write, i.e., distributions of function words, distributions of word lengths, distributions of sentence lengths, and the preliminary set of baseline linguistic features described in Section 3.

To test the ability of different sets of features to capture style, we trained models on a subset of the titles by the above listed eight authors and tested on a different subset of titles by the same authors. We repeated this experiment five times so that several different sets of titles were trained and tested on. At each iteration, we used 150 chapters from each of the authors for training and 40 chapters from each

of the authors for testing. Our results on the test set showed that function words outperform all other feature sets on authorship attribution (see Table 3). Top ten most predictive function words identified by information gain on this data set are: `the`, `not`, `of`, `she`, `very`, `be`, `her`, `'s`, `and`, and `it`.

Feature Set	Avg. Accuracy
Function words	87%
Syntactic elements of expression	62%
Distribution of word length	40%
Baseline linguistic	39%
Distribution of sentence length	34%

Table 3: Results for authorship attribution. Classifier is trained on 150 chapters from each author, and tested on 40 chapters from each author. The chapters in the training and test sets come from different titles.

These results indicate that syntactic expression features are not as effective as function words in capturing the style of authors. This finding is consistent with our intuition: we selected the syntactic elements of expression for their ability to differentiate between different books and titles, even when some titles are written by the same author. Recognizing the style of an author requires focus on the elements that are similar in the titles written by the same author, instead of focus on elements that differentiate these titles.

However, the syntactic elements of expression are not completely devoid of any style information; they successfully identify authors 62% of the time. Nevertheless, the results of this experiment highlight the computational difference between expression and style. These two concepts are different not only in their function, but also in their computational composition.

5.3 Conclusion

In this paper, we presented a comparative study of expression and style and we have identified a computational definition of each. Through experiments, we have shown that syntax plays a role in identifying expression; however, high-level information in the form of distribution of function words is sufficient to capture style.

Expression and style are both related to the way people write; however, the two concepts differ functionally and in their level of dependence on content: expression is dependent on content whereas style is independent of content. The experiments presented here enable us to computationally describe this qualitative difference between expression and style. We show that information about syntactic constructs can capture the differences in expression of content. These features can identify individual books, even when they share content, and even when they are written by the same person. They can also recognize independently copyrighted derivatives of the same title by highlighting the creative expression of each author even when two authors write about the same content. However, these features are not as successful in capturing style of authors—style can be better captured by features that are used similarly in different works of an author and that would not be able to differentiate between the author's works. Function words provide one such feature set; they recognize an author's style more accurately than any of the other feature sets.

6. REFERENCES

[1] D. Alexander and W. J. Kunz. Some classes of verbs in English. In *Linguistics Research Project*. Indiana University, June 1964.

[2] R. Barzilay and K. McKeown. Extracting paraphrases from a parallel corpus. In *ACL/EACL*, 2001.

[3] E. Brill. A simple rule-based part of speech tagger. In *3rd Conference on Applied Natural Language Processing*, 1992.

[4] M. Diab, J. Schuster, and P. Bock. A preliminary statistical investigation into the impact of an n-gram analysis approach based on word syntactic categories toward text author classification. In *6th International Conference on Artificial Intelligence Applications*, 1998.

[5] A. Glover and G. Hirst. Detecting stylistic inconsistencies in collaborative writing. In M. Sharples and T. van der Geest, editors, *The new writing environment: Writers at work in a world of technology*. Springer-Verlag, 1996.

[6] D. I. Holmes. Authorship attribution. *Computers and the Humanities*, 28, 1994.

[7] B. Katz and B. Levin. Exploiting lexical regularities in designing natural language systems. In *12th International Conference on Computational Linguistics, COLING '88*, 1988.

[8] D. Khmelev and F. Tweedie. Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 16(4), 2001.

[9] M. Koppel, N. Akiva, and I. Dagan. A corpus-independent feature set for style-based text categorization. In *IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[10] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khemelev. Using literal and grammatical statistics for authorship attribution. In *Problemy Peredachi Informatsii*, volume 37(2), April-June 2000.

[11] B. Levin. *English Verb Classes and Alternations. A Preliminary Investigation*. University of Chicago Press, 1993.

[12] T. C. Mendenhall. Characteristic curves of composition. *Science*, 11, 1887.

[13] F. Mosteller and D. L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302), 1963.

[14] R. D. Peng and H. Hengartner. Quantitative analysis of literary styles. *The American Statistician*, 56(3), 2002.

[15] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 2002.

[16] H. S. Sichel. On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society (A)*, 137, 1974.

[17] M. W. A. Smith. Recent experience and new developments of methods for the determination of authorship. *Association for Litrary and Linguistic Computing Bulletin*, 11, 1983.

[18] D. R. Tallentire. *An Appraisal of Methods and Models in Computational Stylistics, with Particular Reference to Author Attribution*. PhD thesis, University of Cambridge, 1972.

[19] R. Thisted and B. Efron. Did Shakespeare write a newly-discovered poem? *Biometrika*, 74, 1987.

[20] Ö. Uzuner. *Identifying Expression Fingerprints Using Linguistic Information*. PhD thesis, Massachusetts Institute of Technology, 2005.

[21] Ö. Uzuner, R. Davis, and B. Katz. Using empirical methods for evaluating expression and content similarity. In *37th Hawaiian International Conference on System Sciences (HICSS-37). IEEE Computer Society*, 2004.

[22] Ö. Uzuner and B. Katz. Capturing expression using linguistic information. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, 2005.

[23] Ö. Uzuner and B. Katz. A comparative study of language models for book and author recognition. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, 2005.

[24] C. B. Williams. Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. *Biometrika*, 62(1), 1975.

[25] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco, 2000.

[26] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML-97, 14th International Conference on Machine Learning*, 1997.

Term re-occurrence measures for analyzing style

Avik Sarkar[1], Anne De Roeck[1], Paul H Garthwaite[2]
[1] Department of Computing, [2] Department of Statistics
The Open University
Milton Keynes, MK7 6AA, UK
{a.sarkar, a.deroeck, p.h.garthwaite}@open.ac.uk

ABSTRACT

In this paper, we propose to investigate style through modeling burstiness in the occurrence patterns of terms in different collections. We set out a fine grained model that looks at gaps between the successive occurrence of the term using a mixture of exponential distributions. A Bayesian framework allows flexibility in fitting the model. The parameter estimates are then studied to understand the distributional properties of a term in various collections. We investigate the behaviour of a range of terms and conclude that the model brings out useful features that may be deployed in the analysis of style.

Categories and Subject Descriptors

H.1.0 [**Information Systems**]: Models and Principles; H.3.1 [**Information Systems**]: Information Storage and Retrieval-Content Analysis and Indexing

General Terms

Theory, Experimentation, Algorithms

Keywords

Term burstiness, Term re-occurrence, Bayesian analysis, mixture models, stylistic analysis, frequent terms

1. INTRODUCTION

Stylistic analysis is focused on two problem areas: authorship attribution (including applications in computational forensic linguistics) and genre detection or identification. Currently, research in this area mainly uses techniques based on term frequency counts: frequency data are collected for common terms, possibly together with other features in the document (such as sentence length) or collection (such as document length), and these data are then analyzed using a range of fairly standard statistical techniques [3] and some other approaches [1, 2].

The popularity of term frequency measures is in part explained by the ease with which word counts can be extracted and manipulated. Whilst cost-effective, frequency based measures only give us one kind of information, and lack granularity in some respects. For instance, function words or stop words, which tend to be very common, are often dropped from frequency counts, as are rare words, where the effectiveness of statistical approaches breaks down. Also, term frequency based approaches are grounded in the "bag-of-words" assumption, which stipulates that term occurrences (and, indeed, term re-occurrences) are independent of each other in text.

The raw frequency of a word in a document is not a useful measure unless it is interpreted relative to document length, simply because longer documents can be assumed to have more term occurrences. The effect of document length can be nullified by using the relative frequency, i.e. the frequency normalized with respect to the document length. However, Katz [9] has shown that relative frequency is insufficient to account for a term's distributional and re-occurrence patterns in a document. He demonstrates this by posing a question:

- If a certain word occurs on the first page of a 200-page book, and also on the first page of a 20-page paper, so is the chance of observing the word again in the remainder of the book about 10 times higher than observing it in the paper?

Katz claims that the answer to the above question is no. Yet, by counting all occurrences of a term in a single consolidated count, frequency based measures assume that once a term occurs in a document, its overall frequency in the entire document is the only useful measure that reflects the term's behaviour. In other words, they assume that additional positional information cannot lever any extra performance in the applications such measures inform. This assumption is appropriate if the stuff being counted is homogeneously distributed across the entire span of the collection. In that case, no consideration is needed of whether a term occurred in the beginning, middle or end of a document, or whether it occurred many times in close succession as opposed to a more even distribution throughout the text. Yet, for text, that assumption has been shown to be wrong [7]. Terms do not distribute homogeneously. Even very frequent terms, which are usually assumed to be mere "background noise" do not distribute in the same way throughout text [10, 5]. Frequency

based models lose fine grained information about distribution patterns of terms, the behaviour of phrases linked by function words, and co-occurrence relationships between different words. Intuitively, these would appear important for the investigation of style.

Analyzing style is meant to capture document structure. Terms are believed to occur in bursts [4, 9] and we would like to study this characteristic. Unlike other approaches, we will look at bursts not by looking at term occurrences, but by looking at the length of gaps between occurrences of terms. In a burst, the gaps between occurrences of a term will be comparatively short, whereas in between bursts, the gaps will be longer.

The organization of the paper is as follows. In section 2 we discuss the issue of burstiness in text and some work that demonstrates the failure of the "bag of words" assumption. We motivate our approach. In section 3 we describe the mixture model for studying gaps. Section 4 describes the Bayesian estimation theory, methodology and ways to interpret the parameters. In section 5 we describe the experimental framework and the datasets we have used. In section 6 we analyze the chosen terms in four distinct categories based on the parameters from our model. We provide conclusions and suggest directions for future work in section 7.

2. BURSTINESS

Burstiness is a phenomenon usually associated with content words. Once they have occurred in a text, the likelihood that they re-occur soon afterwards is much higher than the standard frequency based probability estimate would predict [4]. This is known as *within-document burstiness*, or the close proximity of all or some individual instances of a word within a document exhibiting multiple occurrences [9].

Usually, term burstiness is investigated by counting terms and after a term has occurred, adjusting its probability to account for a "burst". Church [4], for instance, uses Poisson Mixtures, and in a later approach, an adaptive language model [16] based on conditional probabilities. A measure of burstiness was proposed as a binary value that is based on the magnitude of average-term frequency of the term in the corpus [12]. This measure takes the value 1 (bursty term) if the average-term frequency value is large and 0 otherwise. The measure is too naive and incomplete to account for term burstiness.

Katz [9] adopts the same basic starting point of counting term occurrence, but uses K-mixtures, and proposes a model for within-document burstiness with three parameters:

- the probability that a term occurs in a document at all (document frequency)
- the probability that it will occur a second time in a document given that it has occurred once
- the probability that it will occur another time, given that it has already occurred k times (where k > 1).

There are several drawbacks to this model. First of all, it cannot handle non-occurrence of a term in a document, and so is unsuitable as a basis for looking at rare terms. Second, the model associates burstiness with content words only, and is unsuitable for predicting the behaviour of function words, which have also been shown to display burstiness. Finally, the model cannot account for the rate of re-occurrence of the term or the length of gaps.

Our model overcomes these drawbacks. Unlike other approaches, we will look at bursts not by looking at term occurrences, but by looking at the length of gaps between occurrences of terms. This has the advantage that it adds information about the distribution of words, whilst the main frequency based information also remains available. The models also allow frequency counts to be derived from them.

2.1 Homogeneity Assumption

The popular "bag of words" assumption for text states that a term's occurrence is uniform and homogeneous throughout. A measure of homogeneity or self-similarity of a corpus can be calculated, by dividing the corpus into two frequency lists based on the term frequency and then calculating the χ^2 statistic between them [10]. Various schemes for dividing the corpus were used [5] to detect homogeneity of terms at document level, within-document level and by choosing text chunks of various sizes. This work revealed that homogeneity increases by nullifying the within document term distribution pattern and homogeneity decreases when chunks of larger size are chosen, as they incorporated more document structure. Other work based on the same methodology [6] reveals that even very frequent function words do not distribute homogeneously over a corpus or document. These papers [5, 6] provide evidence of the fact that the "bag of words" assumption is invalid. Thus it sets the stage for a model that defies the independence assumption and considers term distribution patterns.

3. THE MODEL

We build an independent single model for every particular term of interest for each of the datasets [14]. Let us suppose we have chosen a certain collection of documents and we are interested in the term "x" in that collection.

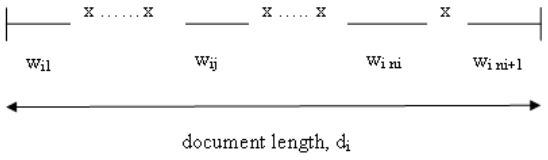


Figure 1: The document structure and the gaps between terms

Figure 1 shows the i^{th} document in the collection. Suppose for this document, the document length is d_i and the term "x" occurs n_i times in that document. w_{i1} denotes the position of first occurrence of the document and $w_{i2}, \ldots, w_{in_i}$ denotes the successive gaps between occurrences of term "x" in the document.

These gaps w_{ij} are modeled using a mixture of exponential distributions.

$$\phi(w_{ij}) = p\lambda_1 e^{-\lambda_1 w_{ij}} + (1-p)\lambda_2 e^{-\lambda_2 w_{ij}} \quad (1)$$

for $j \in \{2, \ldots, n_i\}$. Without loss of generality we let λ_1 be the larger of the two λs and let p and $(1-p)$ denote, respectively, their probabilities of membership.

When λ_1 and λ_2 differ substantially, the first exponential distribution (with the larger mean) mainly determines the rate with which the particular term will occur if it has not occurred before or it has not occurred recently. And the second exponential component (with the smaller mean) determines mainly the rate of re-occurrence in a document or text chunk given that it has already occurred recently. This component captures the bursty nature of the term in the text (or document) i.e. the *within-document burstiness*. The roles of the two exponentials become less distinct as λ_1 and λ_2 become closer together.

The first occurrence of the term in a document is not the result of a re-occurrence from the past. For this reason, while modeling the first occurrence w_{i1}, the second exponential component that accounts for burstiness is not required. Thus,

$$\phi_1(w_{i1}) = \lambda_1 e^{-\lambda_1 w_{i1}}$$

The term x occurs n_i times in the document, but according to our model we assume that the term will occur a further time, i.e. an $(n_i+1)^{th}$ time, if the document had continued. However, since the document ended at length d_i, the event of observing the $(n_i+1)^{th}$ occurrence of the term is censored. The length of censoring is the number of positions from the n_i^{th} occurrence of the term to the end of the document. The information about the model parameters that is given by the censored occurrence is,

$$\begin{aligned} Pr(w_{in_i+1} > cen_i) &= \int_{cen_i}^{\infty} \phi(x)dx \\ &= pe^{-\lambda_1 cen_i} + (1-p)e^{-\lambda_2 cen_i} \end{aligned}$$

where,

$$cen_i = d_i - \sum_{j=1}^{n_i} w_{ij}$$

The big advantage of censoring the event of observing a term is that we can handle the non-occurrence of a term in a document as most terms are unlikely to occur in a particular document. When a term does not occur we censor the event of observing that particular term in that document at the document length.

4. BAYESIAN ESTIMATION

Two philosophically different approaches to statistical inference are classical or *frequentist* statistics and Bayesian statistics. The essential difference between them is that the Bayesian approach allows any unknown quantity to have a probability distribution while the frequentist approach only allows random variables to have a probability distribution. For instance, an unproven conjecture in mathematics is that any positive even number can be written as the sum of two prime numbers (12=7+5; 26=3+23; etc). A Bayesian might state that 0.9 is the probability that this conjecture is true. Frequentist statistics says that there is no random uncertainty, so the probability that the conjecture is true is either 0 or 1, and can be nothing in between. The distinction means that the parameters of a model, such as λ_1 and λ_2, can have probability distributions with the Bayesian approach but not with the frequentist approach.

In the Bayesian approach, a *prior* distribution is used to convey the information about model parameters that was available before data were gathered. This is combined with the information supplied by the data, which is contained in the *likelihood*, to yield a *posterior distribution*. Formally,

$$\text{posterior} \propto \text{prior} \times \text{likelihood},$$

where $\propto$ means 'is proportional to'. Prior distributions are a strength of Bayesian statistics in that they enable background knowledge to be incorporated into a statistical analysis. However, they are also a weakness because a prior distribution must always be specified, even if no useful background information is available or if one does not wish to use background knowledge, perhaps to ensure the analysis is transparently impartial, or because it can be difficult and time-consuming to specify a prior distribution that provides a good representation of the available prior knowledge. Consequently, in practice a prior distribution is almost always chosen in a mechanical way that yields a distribution designed to be non-informative.

Bayesian methods have sprung to prominence over the last fifteen years. This is because of the development of good computational techniques, notably **Markov Chain Monte Carlo (MCMC)** methods, that have solved many of the numerical problems formally associated with the practical application of the Bayesian approach. With these new techniques, Bayesian methods can now analyse complex problems that frequentist methods cannot handle. This has led to the general acceptance of Bayesian methods.

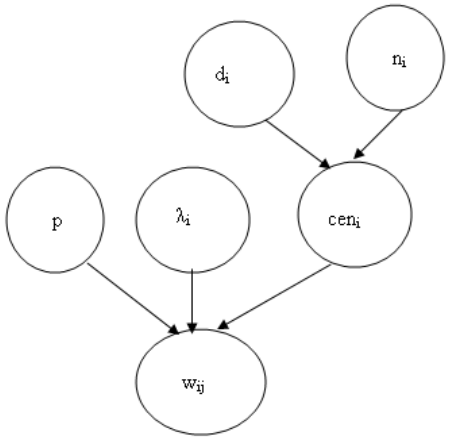


Figure 2: Bayesian dependencies between the parameters

For the model defined in section (3), let $\vec{\Theta} = \{p, \lambda_1, \lambda_2\}$ denote its parameters and let $\vec{W} = \{w_{i1}, \ldots, w_{in_i}, , w_{in_i+1}\}$ denote the data. We define the following:

- $f(\vec{\Theta})$ is the **prior distribution** of $\vec{\Theta}$. So as to obtain a non-informative prior distribution (figure 2) we suppose that $f(\vec{\Theta})$ specifies
  $p \sim Uniform(0, 1)$, and
  $\lambda_1 \sim Uniform(0, 1)$
  Also, to tell the model that λ_2 is the larger of the two λs, we put $\lambda_2 = \lambda_1 + \gamma$, where $\gamma > 0$, and
  $\gamma \sim Uniform(0, 1)$

- $f(\vec{W}|\vec{\Theta})$ is the **likelihood function**. It is our model for the data $\vec{W}$ conditional on the parameters $\vec{\Theta}$. (As well as the observed data, the likelihood also conveys the information given by the censored values)

- $f(\vec{\Theta}|\vec{W})$ is the **posterior distribution** of $\vec{\Theta}$, given $\vec{W}$. It describes our beliefs about the parameters given the information we have.

From Bayes Theorem,

$$f(\vec{\Theta}|\vec{W}) = \frac{f(\vec{W}|\vec{\Theta})f(\vec{\Theta})}{f(\vec{W})} \tag{2}$$

where $f(\vec{W})$ is simply a normalizing constant, independent of $\vec{\Theta}$. Thus equation 2 has the form,

$$f(\vec{\Theta}|\vec{W}) \propto f(\vec{W}|\vec{\Theta})f(\vec{\Theta}). \tag{3}$$

In many cases, as in ours, it is impossible to find a closed form expression for the posterior distribution, $f(\vec{\Theta}|\vec{W})$, because $f(\vec{W})$ in equation (2) cannot readily be determined. However, we may use a simulation process based on random numbers, **Markov Chain Monte Carlo (MCMC)** [8], to generate a very large sample of observations from the distribution $f(\vec{\Theta}, \vec{W})$, even though we cannot write down an expression for this distribution. These observations are a sample of values of $\vec{\Theta}$ and they can be used to make inferences and estimates for $\vec{\Theta}$.

The MCMC method used here is **Gibbs Sampling** [8], which is a popular MCMC method and the simplest. It provides an elegant way for sampling from the joint distribution of a vector of parameters. Initial random values are assigned to the parameters and then samples are repeatedly generated from the conditional distribution of each parameter in turn. (Each conditional distribution is a univariate distribution and the conditions set all parameters to their most recently sampled values, apart from the parameter whose distribution is now being sampled.) The process is based on the Markov chain assumption, which states that the next generated value only depends on the present value and does not depend on the values previous to it. Based on mild regularity conditions, the chain of generated values will gradually *forget* its initial starting point and will eventually converge to a unique *stationary distribution*. The values generated from the start to the point where the chain settles down are discarded and are called the *burn-in* values. Thereafter, the generated values are from the stationary distribution, which is the posterior distribution, $f(\vec{\Theta}, \vec{W})$. Inferences and estimates are based on these generated values and will be subject to random variation, but the effects are insignificant provided a very large number of samples are generated.

We are modeling the length of gaps by a *mixture distribution* (*cf* equation (1)). Fitting this type of distribution is tricky and requires special techniques. Here, data augmentation is used to make it feasible. For details about this, see [13] which describes in detail the fitting of mixture models in MCMC methods.

4.1 Parameter Estimation

Parameter estimation was carried out using Gibbs Sampling on the WinBUGS software [15]. Values from the first 1000 iteration were discarded as burn-in. It had been observed that in most cases the chain reached the stationary distribution well within 1000 iterations. A further 5000 iterations were run to obtain the parameter estimates.

4.2 Interpretation of Parameters

The parameters of the model can be interpreted in the following manner:

- $\widetilde{\lambda_1} = 1/\lambda_1$ is the mean of an exponential distribution with parameter λ_1. $\widetilde{\lambda_1}$ measures the rate at which this term is expected in a running text corpus. $\widetilde{\lambda_1}$ determines the rarity of a term in a corpus, as it is the average gap at which the term occurs if it has not occurred recently. Thus, a large value of $\widetilde{\lambda_1}$ tells us that the term is very rare in the corpus and vice-versa.

- Similarly, $\widetilde{\lambda_2}$ measures the *within-document burstiness*, i.e. the rate of occurrence of a term given that it has occurred recently. It measures the term re-occurrence rate in a burst within a document. Small values of $\widetilde{\lambda_2}$ indicate the bursty nature of the term.

- $\widetilde{p}$ and $1 - \widetilde{p}$ denote, respectively, the probabilities of the term occurring with rate $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ in the entire corpus. Hence $\widetilde{p}$ denotes the proportion of times the term does not occur in a burst, and $1 - \widetilde{p}$ denotes the proportion of times the term occurs in a burst.

Table 1 presents some heuristics for drawing inference based on the values of the parameter estimates.

	$\widetilde{\lambda_1}$ small	$\widetilde{\lambda_1}$ large
$\widetilde{\lambda_2}$ small	frequently occurring and common function word	topical content word occurring in bursts
$\widetilde{\lambda_2}$ large	comparatively frequent but well-spaced function word	infrequent and scattered function word

Table 1: Heuristics for inference, based on the parameter estimates.

Based on the parameters and their interpretation, the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristic seems useful for inference.

5. EXPERIMENTAL FRAMEWORK

We are interested in investigating whether term burstiness patterns can contribute usefully to the analysis of style - i.e. whether burstiness models can uncover useful information about the behaviour of terms that is not typically available from methods that are based on frequency counts alone. We are interested in investigating the behaviour of different kinds of terms, because it seems reasonable to expect that the contribution made by function words, for instance, may be of a different nature than that of content or rare terms. Also, some words may be closely associated with a particular style and may therefore display a behaviour in documents written in that style, that is different to their "standard" behaviour.

We set up a series of experiments. First of all, we selected five different datasets, and we equate each with a different genre. We are careful not to make the assumption that any collection automatically amounts to a style or genre. These are initial experiments and we picked standard, high quality collections from the TIPSTER dataset, which can be reasonably argued to represent different genres. Table 2 gives a short description of each.

Dataset	Contents of the documents
AP	Copyrighted AP Newswire stories from 1989.
DOE	Short abstracts from the Department of Energy.
FR	Issues of the Federal Register (1989), reporting source actions by government agencies.
PAT	U.S. Patent Documents for the years 1983-1991.
SJM	Copyrighted stories from the San Jose Mercury News (1991).

Table 2: The contents of dataset

Table 3 sets out some basic profiles of these collections. They show that the datasets are quite different from each other with respect to size and average document length. The type-to-token ratio (ratio of new words to old) per million words provides a rough estimate of the breadth of the corpus.

Dataset	Corpus Length	Average doc. length	Type to token ratio
AP	114,438,101	471.1	106.845
DOE	26,882,774	119.0	94.778
FR	62,805,175	1,370.7	144.866
PAT	32,151,785	4,790.9	134.017
SJM	39,546,073	438.1	102.149

Table 3: Basic statistics for each of the datasets

We identified four different types of terms for modeling. We chose very frequent and less frequent function words, some terms that are used in connection with reported speech and reporting styles (which are relevant to two of the datasets: AP and SJM), and some terms that might behave like content words in some of the collections. For each of these, we collected frequency based information across the different datasets, as shown in table 6. *Relative document frequency* tells us about the proportion of documents in the collection that contain the particular term. *Rate of incidence* measures the relative frequency of a particular term in the entire dataset, and it provides a measure of the term's distributional density across the entire corpus (rate of incidence = (total number occurrences of the term in the corpus) / (corpus length)). In this case the rate of incidence is expressed as the incidence of the term per 100,000 words in the collection.

We then ran our model for the full set of terms using a random selection (due to hardware limitations) of $1-10\%$ of the documents from each collection, and building the models using [15] on a desktop computer. We compared the burstiness patterns that emerged across different datasets. We paid special attention to differences between terms display similar behaviour according to the frequency based measures, to identify where our model adds information.

6. ANALYSIS OF CHOSEN TERMS

In this section, we shall choose one group of terms at a time and discuss the findings based on our model as compared to those based on relative document frequency and rate of incidence.

6.1 Very frequent function words

We selected the very frequent function words *the*, *of* and *are* (table 4), because they are ubiquitous. They are often subjected to stop word removal because they are thought to behave like background noise in any collection. Certainly, their frequency based profiles show that *the* and *of* occur pretty much in all documents in each collections, and pretty much at indistinguishable rates of incidence. *Are* is less ubiquitous perhaps occurs less frequently in shorter documents.

Term	Dataset	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
are	AP	0.45	47.39	45.48	1.04
	DOE	0.32	31.85	30.24	1.05
	FR	0.07	101.30	33.70	3.01
	PAT	0.27	423.73	84.32	5.03
	SJM	0.01	473.26	45.13	10.49
of	AP	0.53	38.65	36.63	1.06
	DOE	0.62	21.10	19.72	1.07
	FR	0.01	200.28	24.05	8.33
	PAT	0.02	86.06	21.54	3.99
	SJM	0.04	204.37	39.45	5.18
the	AP	0.59	16.58	16.11	1.03
	DOE	0.29	20.49	12.72	1.61
	FR	0.01	194.89	13.47	14.47
	PAT	0.02	68.07	10.36	6.57
	SJM	0.02	168.52	17.80	9.47

Table 4: Parameter estimates of very frequently occurring function words

The terms *the* and *of* have low values of $\widetilde{\lambda_1}$ indicating frequent usage of these terms in most datasets. In FR and SJM $\widetilde{\lambda_1}$ values are higher possibly due to the fact that some documents contain notices or instructions, which are not plain English hence *the* and *of* does not occur in them. Small values of $\widetilde{\lambda_2}$ indicates frequent re-occurrence. The $\widetilde{\lambda_2}$ values for the different datasets are in a close proximity. In FR and SJM, however, the behaviour of *of* and, surprisingly, *the* is clearly much burstier than it is in the other datasets, with long gaps separating close bursts.

The term *are* has very close values of $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ for the AP and DOE datasets, indicating the fact that this term occurs evenly across these two datasets. The behaviour of *are* in the other datasets is quite different. It has high values of $\widetilde{\lambda_1}$ for FR, PAT and SJM. This is combined with low $\widetilde{\lambda_2}$ value for the SJM dataset, leading to a relative bursty behaviour. The model for *are* again shows a distinctive behaviour in SJM, particularly as compared to AP, even though the associated frequency based profiles are indistinguishable.

6.2 Less frequent function words

Not all function words are as frequent as the examples in the previous section. Though they are often removed as stop words, less frequent function words tend to be associated with certain types of syntactic structure, and hence may be indicative of style. Some less frequent function words we study are *could*, *should*, *as*, *expect* and *in* (table 5).

Syntactically speaking, *could* and *should* are both modals and have comparable usage in English. Table 6 shows that they have different relative document frequency values, but that their rate of incidence across the different collections is almost equivalent. Hence, even using linguistic knowledge these terms cannot be differentiated on that basis. Our model, on the other hand, shows a consistently bursty behaviour in the FR and PAT collections, indicating a different usage pattern in government reports and in patent documents. Both these sets use comparatively formalized styles and document structures that are not uniform throughout the document.

The term *as* is quite interesting. It occurs in quite a high proportion of documents in all the datasets, with a uniform rate of occurrence. This is borne out by the $\widetilde{\lambda_1}$ values which show a uniform distance between bursts. However, in FR and PAT within-burst distance is larger, and it behaves like a relatively scattered function word, whereas in AP, DOE and SJM the very low values of $\widetilde{\lambda_2}$ depict a very bursty behaviour.

The frequency based profile of *except* is diverse. In our model, it has large values of $\widetilde{\lambda_1}$ for all the datasets, and also has quite large values of $\widetilde{\lambda_2}$. Hence based on table 1 and on the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics this term appears to behave as a scattered function word. The exception is the PAT dataset, where the term occurs much more burstily.

The term *in* displays very similar behaviour across the board, using both types of measures, and in all collections. Though it is a preposition, like *of*, their behaviours are markedly distinct.

Term	Dataset	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
could	AP	0.52	1631.85	539.37	3.03
	DOE	0.61	3095.02	1078.98	2.87
	FR	0.74	3810.98	293.17	13.00
	PAT	0.74	9174.31	273.90	33.50
	SJM	0.53	2741.23	450.86	6.08
should	AP	0.45	2890.17	1020.30	2.83
	DOE	0.62	4677.27	1715.56	2.73
	FR	0.48	1423.08	101.50	14.02
	PAT	0.73	5065.86	268.96	18.83
	SJM	0.58	5063.29	627.35	8.07
as	AP	0.93	241.55	7.60	31.76
	DOE	0.93	215.52	6.85	31.47
	FR	0.45	287.85	72.46	3.97
	PAT	0.27	300.75	68.54	4.39
	SJM	0.90	256.61	6.30	40.72
except	AP	0.82	19755.04	3650.97	5.41
	DOE	0.60	17908.31	3593.24	4.98
	FR	0.49	7668.71	1056.97	7.26
	PAT	0.83	13622.12	192.31	70.84
	SJM	0.67	29120.56	6309.15	4.62
in	AP	0.13	94.79	42.55	2.23
	DOE	0.17	91.74	36.81	2.49
	FR	0.02	359.07	50.68	7.08
	PAT	0.08	137.17	41.70	3.29
	SJM	0.10	141.48	48.17	2.94

Table 5: Parameter estimates of some less frequent function words

6.3 Style indicative terms

Some terms may be associated with particular styles or genres such as verbs indicating reported speech, or a specific way of attributing sources or information. We chose to investigate the behaviour of three such terms: *called*, *report* and *said* (table 7)

Table 6 shows that *report* occurs in a smaller proportion of documents in DOE and PAT as compared to the other collections. At the same time, DOE exhibits the smallest value of $\widetilde{\lambda_1}$ and PAT has the largest (table 7). This may seem incompatible but may be explained by the fact that DOE consists of short abstracts whereas PAT has large patent articles in the form of reports. The term also has hugely differing values of $\widetilde{\lambda_2}$, though the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics indicates similar overall behaviour when comparing between-burst and within-burst gaps. Our model here helps determine that *report* may be an important term in the stylistic analysis of these collections, something that simple frequency based measures do not reveal.

The term *called* has close values of document frequency for both the AP and PAT collections, and the values are large enough to be comparable to a function word. However, table 3 shows the average document length for PAT to be much larger than that of AP, with a much higher occurrence rate. Our model presents the term as more bursty in PAT than in AP. Using our heuristics, *called* does not behave like a function word in PAT. Also, the rate of incidence is of the same order of magnitude in the FR and PAT collections, supported by close values of the $\widetilde{\lambda_1}$ parameter. In the FR

Dataset Term	AP	DOE	FR	PAT	SJM
are	0.595	0.556	0.639	0.927	0.534
	331.5	861.4	463.0	587.6	384.2
as	0.728	0.382	0.689	0.999	0.620
	459.7	543.4	639.2	769.3	468.3
associated	0.482	0.041	0.106	0.323	0.210
	110.5	39.9	23.4	30.5	55.5
called	0.214	0.013	0.031	0.208	0.151
	58.5	11.8	4.3	9.5	44.3
could	0.316	0.045	0.136	0.322	0.272
	103.0	45.1	41.7	19.6	101.2
current	0.074	0.059	0.190	0.238	0.057
	19.2	70.6	47.9	65.7	15.9
data	0.028	0.154	0.207	0.241	0.029
	8.8	202.1	89.9	159.3	11.0
energy	0.041	0.165	0.151	0.173	0.024
	16.2	258.6	48.6	27.3	9.5
except	0.030	0.008	0.154	0.235	0.032
	6.6	7.1	37.1	12.1	8.0
in	0.980	0.863	0.898	0.951	0.944
	2189.7	2361.9	1840.4	2039.3	1842.3
of	0.985	0.978	0.975	1.000	0.946
	2656.3	5022.3	4081.9	4267.3	2328.7
report	0.152	0.062	0.182	0.036	0.136
	63.0	65.4	53.7	1.9	47.3
require	0.031	0.012	0.165	0.218	0.027
	8.1	10.3	49.1	7.3	7.1
required	0.042	0.039	0.319	0.504	0.032
	10.1	36.6	126.1	31.9	8.5
requirements	0.015	0.027	0.338	0.140	0.009
	3.9	28.4	161.3	5.2	2.6
requires	0.021	0.014	0.166	0.238	0.019
	4.9	12.4	35.0	8.6	4.8
said	0.899	0.003	0.070	0.812	0.602
	1265.7	4.7	10.4	867.5	635.2
should	0.185	0.037	0.464	0.494	0.169
	57.4	37.6	153.2	30.5	59.9
the	0.998	0.986	0.957	1.000	0.977
	6108.6	7703.9	6814.6	8388.1	5271.1

Table 6: Table showing values of relative document frequency(proportion of documents where the term occurs) and rate of incidence((total occurrences of the term in the corpus)x(10^5) / (corpus length)) for the chosen terms across all the datasets. The top value in each cell is the relative document frequency and the lower value is the rate of incidence (x10^5)

collection, however, this term is of a much more bursty nature, having comparatively smaller $\widetilde{\lambda_2}$ values. Hence based on our $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics, *called* behaves like a bursty content term for FR.

Probably the most interesting term in table 7 for analyzing style is *said*. It directly indicates a document or a collection referring to a conversation. This term has high values of relative document frequency and rate of incidence for the AP, PAT and SJM datasets. This is perhaps unsurprising for the AP and SJM as they are about news. This is supported by our parameter estimates for these datasets. For PAT however, *said* has a rather bursty nature due to large value of $\widetilde{\lambda_1}$ combined with small $\widetilde{\lambda_2}$ value and it behaves like a rare and scattered function word.

Term	Dataset	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
called	AP	0.48	3780.72	997.01	3.79
	DOE	0.72	12861.74	1826.82	7.04
	FR	0.82	38804.81	68.78	564.22
	PAT	0.79	32637.08	656.60	49.71
	SJM	0.71	8237.23	489.96	16.81
report	AP	0.85	4472.27	94.61	47.27
	DOE	0.97	2474.63	5.56	444.69
	FR	0.68	4315.93	71.74	60.16
	PAT	0.94	259875.26	303.49	856.29
	SJM	0.85	8264.46	112.20	73.66
said	AP	0.04	687.76	68.97	9.97
	DOE	0.67	61349.69	12224.94	5.02
	FR	0.84	26385.22	392.62	67.20
	PAT	0.06	2080.30	13.43	154.94
	SJM	0.16	2460.63	92.34	26.65

Table 7: Parameter estimates of terms related to the style of reporting

6.4 Content terms

We also looked at a selection of content term, or terms that might refer to a topic in the collections. The terms we study in this section (table 8) are *associated*, *current*, *data* and *energy*.

The term *current* is interesting, because it is ambiguous between an adjective (in"present" time period) and a noun (in "electricity"). Table 6 shows uniform rate of incidence and relative document frequency values for each of the collections. Our model on the other hand records bursty behaviour with low $\widetilde{\lambda_2}$ values in the DOE and PAT collections. Our heuristics classify this behaviour as that of a content word for the DOE and PAT collections and as a function word for the other collections. This is consistent with the nature of these collections: DOE and PAT contain technical documents. Whilst this requires further investigation, it would appear that our model may be useful in disambiguation with an approach along these lines.

If we were to calculate the inverse document frequency for *associated*, the term would have maximum weight for DOE (lowest relative document frequency) and least weight for AP (highest relative document frequency). The rate of incidence is almost similar for all the datasets. In our model, the $\widetilde{\lambda_1}$ values show pretty similar distances between bursts in FR, DOE and AP, but in DOE, the term is very scattered across the whole collection, and hence has the characteristics of a rare function word. For AP and PAT, though the rate of occurrence is quite high, the re-occurrence rate is quite small, which leads to large values of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio. Here, the term has the characteristics of a bursty content word.

The behaviour patterns of the term *data* help us identify a drawback of frequency based measures. The values for document frequency and the rate of incidence for this term are quite low for AP and SJM when compared to the other collections. A pure frequentist approach would have reason to treat this term as an informative content word in AP and SJM, and as a function word in the other collections. Doing so would ignore the issue of burstiness. Our model shows

Term	Dataset	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
associated	AP	0.68	8019.25	36.44	220.05
	DOE	0.40	7968.13	2461.24	3.24
	FR	0.83	8928.57	522.47	17.09
	PAT	0.50	13104.44	330.91	39.60
	SJM	0.93	2453.99	6.69	366.87
current	AP	0.32	17445.92	4027.39	4.33
	DOE	0.92	3642.99	69.78	52.20
	FR	0.69	4299.23	366.17	11.74
	PAT	0.36	7189.07	60.68	118.48
	SJM	0.86	14039.03	884.96	15.86
data	AP	0.76	24673.08	243.49	101.33
	DOE	0.82	1591.85	67.11	23.72
	FR	0.58	2833.66	64.77	43.75
	PAT	0.23	5336.18	48.50	110.03
	SJM	0.90	46468.40	188.89	246.00
energy	AP	0.96	10711.23	78.74	136.03
	DOE	0.77	1548.71	43.18	35.87
	FR	0.67	14863.26	107.46	138.32
	PAT	0.49	11195.70	86.88	128.86
	SJM	0.71	28457.60	1258.34	22.62

Table 8: Parameter estimates of terms with some dependence on topic and genre

Term	Dataset	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
require	AP	0.32	26624.0	10554.0	2.5
	DOE	0.51	19406.1	6211.1	3.1
	FR	0.56	4975.1	554.0	8.9
	PAT	0.83	36818.8	676.1	54.4
	SJM	0.52	24703.5	9363.3	2.6
required	AP	0.57	22172.9	4780.1	4.6
	DOE	0.60	6377.5	1094.5	5.8
	FR	0.72	1325.9	109.3	12.1
	PAT	0.55	9505.7	678.4	14.0
	SJM	0.65	24813.9	4844.9	5.1
requirements	AP	0.60	27181.3	9950.2	2.7
	DOE	0.81	7745.9	724.1	10.7
	FR	0.58	1229.8	93.5	13.1
	PAT	0.83	40849.6	884.1	46.2
	SJM	0.82	46232.1	3408.3	13.5
requires	AP	0.51	28785.2	9624.6	2.9
	DOE	0.67	82034.4	1578.2	51.9
	FR	0.79	4784.6	337.6	14.2
	PAT	0.80	33590.8	699.7	48.0
	SJM	0.53	37383.1	10232.2	3.6

Table 9: Parameter estimates of terms originating from the common root word *require*

small values of $\widetilde{\lambda_2}$ for DOE, FR and PAT as compared to those of AP and SJM, and our $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics will class this term as a content word for DOE, FR and PAT; and as a scattered rarely occurring non-informative term in AP and SJM. Though they would need some means of confirmation, these findings are plausible given the content of the datasets.

The term *energy* is quite an interesting term in each of the collections. It is a content word in general but, like all content words, could behave as a non-informative function word in an appropriate specialized domain - in this case about energy, such as the DOE collection. This term has a high rate of occurrence, $\widetilde{\lambda_1}$, in all the datasets except DOE, and bursty nature as indicated by the $\widetilde{\lambda_2}$ value for most of the collections. Because of this, the term will be considered as an informative content term in the AP, FR and PAT datasets. The two lowest values of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristic are from DOE, where the within-burst and between-burst gaps are smallest, and SJM where the within-burst and between-burst gaps are largest. Our heuristic predicts that in both collections, the term behaves like a function word, but in DOE it has the characteristics of a frequent function word whereas in SJM it has those of a rarely occurring and scattered one. We believe the patterns associated with this term demonstrate the strength of our model in differentiating the behaviours of a pervasive content word behaving like a function word in a collection.

7. CONCLUSIONS AND FURTHER WORK

In this paper we have described a model to show burstiness patterns in term re-occurrence, by measuring gaps between successive occurrences of a term either as part of a burst, or between bursts. We have used the model to investigate the behaviour of a range of terms, and have contrasted the information it provides with that supplied by pure frequency based approaches.

We believe we can conclude that the model provides additional, fine grained information about the behaviour of terms in different collections. As such, we believe it to be a promising addition to the range of techniques that can be used in style analysis.

On the other hand, our study has been limited in scope, and whilst we have drawn some plausible and promising preliminary conclusions, we need to conduct further evaluation and verification. In particular, we have conducted all experiments on full text, without engaging either in stemming or affix stripping (as would be applied in many search and retrieval applications [11]), and in the absence of syntactic information as might be supplied by a part of speech tagger, so homographs are treated together. Also, our findings will be affected by morphological variations such as plurals or tense formation. Morphosyntactic features do seem to affect the results the model throws up, as we have tried to show in table 9.

Table 9 shows clear differences between strings that, at least under some interpretation, are closely related, such as the group *require*, *requires* and *required* marking person and tense differences on the one hand, and also *requirements* as a related noun. This points to the need to investigate the impact of syntactic features on evidence of term burstiness.

Our future work will be directed towards extending the range of terms whose behaviour we investigate, drawing better inferences from the parameters offered by the model, and looking at the contribution syntactic and part of speech information might play in detecting burstiness profiles.

8. REFERENCES

[1] S. Aaronson. Stylometric clustering: a comparison of data driven and syntactic features, 1999.

[2] S. Argamon, M. Koppel, J. Fine, and A. R. Shimony. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346, 2003.

[3] C. Chaski. Empirical evaluation of language-based author identification techniques. *Forensic Linguistics: International Journal of Speech, Language and Law*, 8(1):1–64, 2001.

[4] K. Church. Empirical estimates of adaptation: The chance of two Noriega's is closer to p/2 than p^2. In *COLING*, pages 173–179, 2000.

[5] A. De Roeck, A. Sarkar, and P. H. Garthwaite. Defeating the homogeneity assumption. In G. Purnelle, C. Fairo, and A. Dister, editors, *Proceedings of 7th International Conference on the Statistical Analysis of Textual Data (JADT)*, pages 282–294, De Louvain, Belgium, 2004. UCL Presses Universitaries.

[6] A. De Roeck, A. Sarkar, and P. H. Garthwaite. Frequent term distribution measures for dataset profiling. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, editors, *Proceedings of the 4th International conference of Language Resources and Evaluation (LREC)*, pages 1647–1650, Paris, France, 2004. European Language Resources Association (ELRA).

[7] A. Franz. Independence assumptions considered harmful. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 182–189, 1997.

[8] W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics Series. Chapman and Hall, London, UK, 1996.

[9] S. M. Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–60, 1996.

[10] A. Kilgarriff. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings of ACL-SIGDAT Workshop on very large corpora*, Hong Kong, 1997.

[11] R. Krovetz. Viewing Morphology as an Inference Process,. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–203, 1993.

[12] K. L. Kwok. A new method of weighting query terms for ad-hoc retrieval. In *ACM SIGIR*, pages 187–195, 1996.

[13] C. P. Robert. Mixtures of distributions: inference and estimation. In W. Gilks, S. Richardson, and D. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 441–464, 1996.

[14] A. Sarkar, P. H. Garthwaite, and A. De Roeck. A Bayesian mixture model for term re-occurrence and burstiness. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 48–55, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[15] D. Spiegelhalter, A. Thomas, N. G. Best, and D. Lunn. WinBUGS: Windows version of Bayesian inference Using Gibbs Sampling, version 1.4, 2003.

[16] K. Umemura and K. Church. Empirical term weighting and expansion frequency. In *Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 117–123, 2000.

Combining Text and Linguistic Document Representations for Authorship Attribution

Andreas Kaster, Stefan Siersdorfer, Gerhard Weikum
{kaster,stesi,weikum}@mpi-sb.mpg.de

Max-Planck-Institute for Computer Science, Germany

ABSTRACT

In this paper, we provide several alternatives to the classical Bag-Of-Words model for automatic authorship attribution. To this end, we consider linguistic and writing style information such as grammatical structures to construct different document representations. Furthermore we describe two techniques to combine the obtained representations: combination vectors and ensemble based meta classification. Our experiments show the viability of our approach.

General Terms

text classification, authorship attribution

Keywords

linguistic information, writing style information, combination techniques, stylometry

1. INTRODUCTION

1.1 Motivation

Automatic document classification is useful for a wide range of applications such as organizing Web, intranet or portal pages into topic directories, filtering news feeds or mail, focused crawling on the Web or in intranets and many more [11].

Most of the text classification approaches deal with topic-oriented classification (e.g., classifying documents into classes like "Sports", "Politics" or "Computer Science"). Here the Bag-Of-Words model, taking just the occurences of words into account (often using additional techniques like stemming, stopword elimination, and different weighting schemes), has been shown to be very effective for this task [20, 45].

But these techniques have limitations for other classification tasks such as authorship recognition. In this context, application scenarios include tasks like plagiarism detection [22], author identification forensics [12], author tracking in discussion forums [31], or solving problems of disputed authorship for historic documents such as the Federalist problem [18, 29].

Although in the case of authorship attribution, there occurs also a certain amount of topic and word correlation (in books written by Doyle we will typically find the names "Holmes" and "Watson", in books written by Christie we have, e.g., "Poirot" and "Marple"), alternative features (e.g., features, that do not contain any information about a document's content at all) may become important.

In this paper we study, in addition to some known approaches (like function words, filtering based on Part-Of-Speech tagging), several new approaches for feature construction. These include writing style features using syntax trees, considering constituents, and statistical measures on tree depths. As a result we obtain different and, to a certain degree, orthogonal document representations.

We combine these representations using two different techniques: combination vectors and meta classification.

1.2 Contribution

The paper makes the following contributions:

- In addition to using some known techniques we describe several novel approaches for the construction of document features for authorship attribution.
- We describe two different ways to combine and weight distinct feature spaces.
- We provide an experimental study of the pros and cons of a variety of methods.

1.3 Outline

The rest of the paper is organized as follows. In Section 2 we briefly review the technical basics of automatic classification and some linguistic basics. In Section 3 we describe different feature representations of documents. Combination methods for different document representations are described in Section 4. Section 5 presents experiments on a dataset based on Project Gutenberg [1]. Finally, in Section 6 we discuss related work in comparison to our own research.

2. TECHNICAL BASICS

2.1 Machine Learning

Classifying text documents into thematic categories usually follows a supervised learning paradigm and is based on training documents that need to be provided for each topic. Both training documents and test documents, which are later given to the classifier, are represented as multi-dimensional feature vectors. In the prevalent bag-of-words model the features are derived from word occurrence frequencies, e.g. based on tf*idf feature weights [6, 26]. Often feature selection algorithms are applied to reduce the dimensionality of the feature space and eliminate "noisy", non-characteristic features, based on information-theoretic measures for feature ordering (e.g., relative entropy or information gain).

The resulting compact feature vectors are used to derive a classification model for each topic, using probabilistic (e.g., Naive Bayes) or discriminative models (e.g., SVM). Linear support vector machines (SVMs) construct a hyperplane $\vec{w} \cdot \vec{x} + b = 0$ that separates the set of positive training examples from a set of negative examples with maximum margin. This training requires solving a quadratic optimization problem whose empirical performance is somewhere between quadratic and cubic in the number of training documents and linear in the number of features [9]. For a new, previously unseen, document $\vec{d}$ the SVM merely needs to test whether the document lies on the "positive" side or the "negative" side of the separating hyperplane. The decision simply requires computing a scalar product of the vectors $\vec{w}$ and $\vec{d}$. SVMs have been shown to perform very well for text classification (see, e.g., [14, 19]).

2.2 Linguistic Basics

For the understanding of techniques described below, we introduce some basic concepts. Further details will be provided later when needed. Consider a set Σ of tags for linguisitic corpus annotation (e.g. the Penn-Treebank-Tagset [27]). Let s be a sentence and $T_s := (V, E, \sigma)$ an ordered tree with a set of nodes V, a set of edges E and a labeling function $\sigma : V \rightarrow \Sigma$, that assigs a label $l \in \Sigma$ to each node of the tree. We call T_s the syntax tree of sentence s. T_s is the tree representation of a probabilistic contextfree grammar (PCFG). A PCFG is a contextfree grammar enriched by transition probabilities for each rewriting rule ([26]). For example, consider Figure 1. There, the sentence *Next, he examined the framework of the door we had broken in, assuring himself that the bolt had really been shot.* is represented as a syntax tree. The leaves of the tree represent the words themselves, i.e. terminal symbols, where the higher nodes represent the PCFG Tags, i.e., non terminal symbols. Non-terminals can be subdevided into other non-terminals or terminals, e.g. NP (a noun phrase) into DT (determiner, an article) and NN (a noun in singular case) and NN into "framework". Intuitively, a syntax tree represents the structure of a sentence, and, in some way, the writing style of an author, which we use for feature construction as described in Section 3.4.

3. CONSTRUCTION OF FEATURES

3.1 Word-Based Features

Using word based features is the most popular and, despite of its simplicity, very effective feature construction method. We briefly describe several variants from the literature, that we will consider as a baseline for other methods.

3.1.1 Bag-Of-Words

In the Bag-Of-Words approach the ordering of the words is not considered. Optionally a stopword list can be used to eliminate very common terms like articles, prepositions, etc. Often additional techniques like stemming [35] are applied to the words. There are different options to construct feature weights: taking the absolute or relative frequency of term occurrences as compontents, constructing a binary feature vector by just considering the pure occurrence of a term, computing the tf*idf values of the terms, etc. [30, 37]. Because it is the state-of-the-art method for feature construction in automatic document classification we will consider Bag-Of-Words as baseline for our experiments.

3.1.2 Function Words

In case of authorship attribution it can make sense to use "content-free" features, i.e., terms, that do not contain information about the document's content such as prepositions, pronouns, determiners etc. These terms are called *function words* (see e.g. Diederich et. al. [13]). In our implementation we regard as function words all words other than nouns, verbs and adjectives.

3.1.3 POS Annotation

In this approach, the part of speech (POS) group of the words (e.g. verb, noun, adjective) is taken into account [33]. This can be used to filter documents, e.g., by considering only nouns or verbs. POS is also used for simple disambiguation, e.g., by distinguishing the verb "book" from the noun "book".

3.1.4 Feature Selection

The idea of feature selection is to just take the most discriminating features into account. Intuitively a well discriminating term for two classes A and B occurs frequently in documents of class A and infrequently in documents of class B or vice versa. Examples of feature selection measures are Mutual Information, Information Gain, and Chi Square [46].

3.1.5 Semantic Disambiguation

Here a thesaurus, e.g. Wordnet [15], is used to disambiguate terms (treating synonyms like "automobile" and "car" as the same feature). In some approaches also more complex relationships between words are taken into account [36, 38].

3.2 Using Linguistic Constituents

The structure of natural language sentences shows that word occurrences follow a specific order, called word order. Words are grouped into syntactic units, *constituents*, that can be deeply nested. Such constituents can be detected by their being able to occur in various positions and showing uniform syntactic possibilities for expansion (see [26]). Consider again the sentence *Next, he examined the framework of the door we had broken in, assuring himself that the bolt had really been shot.* and its syntax tree representation in Figure 1. In particular, consider the part *he examined the framework*. This part is a constituent of the sentence with sub-constituents, e.g. *"the framework"*. The sub-constituents can change their positions inside the bigger constituent. Just considering that specific part, *he examined the framework* has the same meaning as *the framework he examined*. We can use this information about the word relationships by extracting constituents for feature construction. To this end, we first subdivide the document into sentences, and then construct a syntax tree as shown in Figure 1 for each sentence (note that the grey-boxed parts belong to another technique, the writing style, described in 3.4). In our framework, we use the Connexor Machinese Phrase Tagger [41] to subdivide a document into sentences and Lexparser [2] to build the syntax trees. We define a minimal and maximal length, *min* and *max*, of the constituents that we want to use for feature construction.

The simplest way to construct features would be to just concatenate the features inside a constituent using an appropriate separation character (e.g. "$").
From our example sentence, this would result features such as $he\$examined\$the\$framework$. But such very specific features may occur very rarely in the document corpus. To obtain more "common" features a combination of some of the following options is applied:

- Performing stemming and stopword elimination on the words contained in a constituent.
- Abstracting from the ordering of words by putting the words into lexicographic order.
- Instead of a feature $x_1\$x_2\$\ldots\$x_n$ consider pairs $x_i\$x_j$ or triples contained in the constituents (bi- and trigrams).
- Perform a feature selection on the constituent-features themselves. This can be done completely analogously to the feature selection for simple words.

In Section 5, we provide experiments on using whole constituents with stemming and stopword-elimination as well as using bigrams (also stemmed and stopword-cleaned).

3.3 Functional Dependencies

Functional dependencies represent relational information in sentences. Consider again a part of the sentence used in Figure 1, *he examined the framework of the door*. Here, *he* is the subject (agent) and *framework* and *door* are the objects of the predicate *examined* (action). We used the Connexor Machinese Syntax [41] to determine such dependencies. Our features have the form

$$x_1\$x_2\$\ldots\$x_n \tag{1}$$

where x_1 is the subject of an action, x_2 is the predicate and x_3 through x_n are the objects. To obtain a canonical form, words are reduced to their base forms, using Connexor Machinese Phrase Tagger [41], and objects are sorted in lexicographic order. In our example case, we get the feature $he\$examine\$door\$framework$.

3.4 Writing Style: Using Syntax Trees

Different authors may construct sentences in their writings in a completely different way. The idea is to consider a syntax tree representation of their sentences as features. In the extreme case we could encode the whole tree into a string; but this would result in very sparse feature spaces. Instead we should restrict ourselves to nodes up to a certain maximum tree depth. In our experiments we observed that considering just the children of the root nodes of sentences and sub-clauses (labeled with S) provides us already with interesting features. So our example tree in Figure 1 could be encoded into the features $ADVP\$, \$NP\$VP$, VP, and two times $NP\$VP$ (emphasized by the grey boxes). Note that this method does not use any word information at all. Table 1 shows the top-5 features for the authors A. C. Doyle and R. Burton according to their mutual information values (we considered only books available from the Gutenberg Project [1]; see Section 5 for details). We do not apply any kind of filtering mechanism to the structure features such as removing punctuation marks. Experiments showed that those marks provide interesting information about the sentence structure. Note, that "," in the feature $ADVP\$, \$NP\$VP$ represents an annotation tag for a word phrase in the syntax tree, not the comma itself.

A.C. Doyle		R. Burton	
Feature	**MI**	**Feature**	**MI**
S$,$CCS.	0.23	S$:$S	0.26
PPNPVP$.	0.16	SCCS	0.23
SBAR$,$NPVP.	0.14	XXNP$VP	0.21
SBAR$,$XNPVP$.	0.13	S$:$S$.	0.20
PP$,$NPVP.	0.11	S$:$CC$S	0.18

Table 1: TOP 5 MI Features by Writing Style

3.5 Syntax Tree Depth

Other research discovered the benefit of sentence length as feature either by computing the average sentence length [12, 13], or by using histograms over the sentence length [42]. Another simple but, to our knowledge, novel approach to distinguish different writing styles is to consider the depth of the syntax trees in the documents. We consider two approaches.

3.5.1 Statistical Moments

Statistical Moments are one way to characterize a distribution. The *k-th moment* of a random variable X is defined as $E(X^k)$. The expression $E([X-E(X)]^k)$ is called the *k-th central moment* of X.

For a given document d, containing n sentences (and so n trees), we can approximate the k-th moment and the k-th central moment as follows:

$$E(X^k) = \frac{1}{n}\sum_{j=1}^{n} x_j^k \tag{2}$$

and

$$E([X-E(X)]^k) = \frac{1}{n}\sum_{j=1}^{n} [X-E(X)]^k \tag{3}$$

where x_i is equal to the syntax tree depth of the i-th sentence of document d. Note, that $E(X)$ is known as the expectation value and $E([X-E(X)]^2)$ the variance of the random variable X.[1]

The values for different k vary in their order of magnitude. To avoid an overestimation of higher moments we introduce a normalization by taking the k-th root of the k-th moment. For the construction of the feature vectors we consider the first three moments and the second and third central moments [2]. Thus we can represent our document d as the following vector:

$$\Big(E(X), \sqrt{E(X^2)}, \sqrt[3]{E(X^3)}, \sqrt{E([X-E(X)]^2)}, \sqrt[3]{E([X-E(X)]^3)}\Big) \tag{4}$$

3.5.2 Histogram Approach

The most common form of a histogram is obtained by splitting the range of the data into equal-sized bins (called

[1]The fact that the central moment is not perfectly unbiased is not an issue for large n.
[2]The first central moment, $E([X-E(X)])$, is equal to 0.

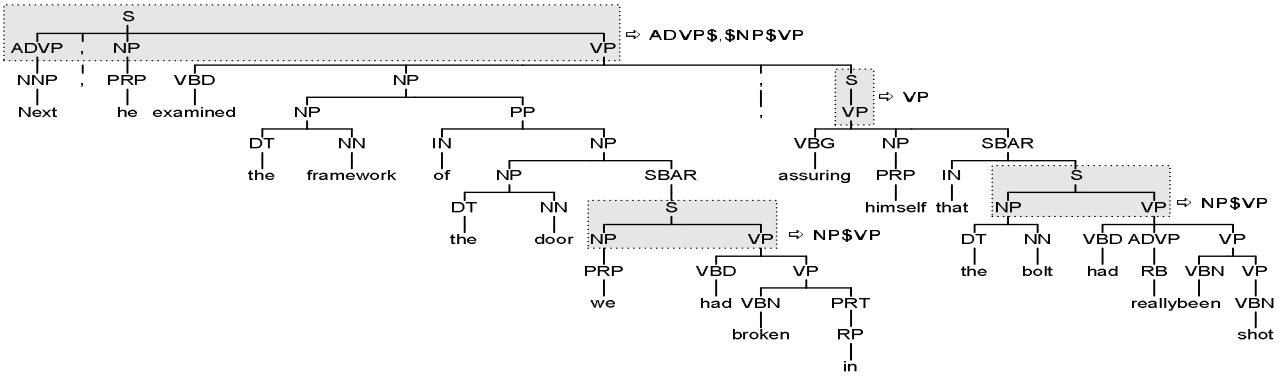


Figure 1: PCFG-Tree and Writing Style Features

classes). Then for each bin, the number of points from the data set that fall into the bin are counted [3].

In our scenario the data consists of the syntax tree depths of a document d. The value assigned to a bin is the number of trees within a certain range of depth (for example all trees of depth 10 to 12). Let $b(i)$ be the value of the i-th bin. As components of the feature vector for document d we consider these values normalized by the overall number n of trees in d and obtain the following vector:

$$\left(\frac{b(1)}{n}, \ldots, \frac{b(m)}{n}\right) \tag{5}$$

We used 5 as concrete bin-size in our implementation. Figure 2 shows a comparison between the tree depth distributions for the authors A. C. Doyle and R. Burton (again books from Gutenberg Project [1], see cpt. 5) in the form of histograms.

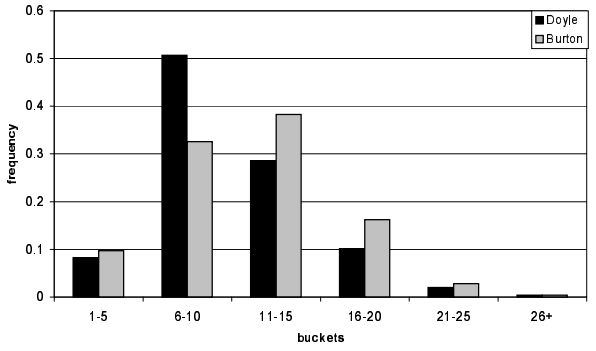


Figure 2: Tree Depth Histogram for two Authors

4. COMBINING FEATURES

In the previous section we have described several different document representations, providing us with different kinds of information about content and style. In this section we describe two approaches to put these pieces of information together.

4.1 Combination Vectors

The idea of combination vectors is to merge the vectors obtained by different document representations into a single vector. This can be done by the concatenation of feature spaces. More precisely we are given k vector representations

$$\vec{v_1}(d), \ldots, \vec{v_k}(d) \tag{6}$$

for document d with

$$\vec{v_i}(d) = (v_{i1}(d), \ldots, v_{im_i}(d)) \tag{7}$$

where m_i is the size of the feature space for the i-th representation.

These vectors can be combined into a combination vector as follows:

$$\left(\frac{v_{11}(d)}{n_1}, \ldots, \frac{v_{1m_1}(d)}{n_1}, \ldots\ldots, \frac{v_{k1}(d)}{n_k}, \ldots, \frac{v_{km_k}(d)}{n_k}\right) \tag{8}$$

Here the values n_i are normalization constants. The rationale for this normalization is that strong variations between the order of magnitude of the components of the feature vectors might occur (this holds, e.g., for Bag-of-Words vs. moments of syntax tree depth distributions). We choose the normalization constants such that the average component value is the same for all subspaces corresponding to the original feature spaces. Formally, for a document set D, we choose the constants n_i such that the following requirement is satisfied:

$$\frac{1}{n_i}\frac{1}{m_i}\sum_{d\in D}\sum_{l=1}^{m_i} v_{il}(d) = \frac{1}{n_j}\frac{1}{m_j}\sum_{d\in D}\sum_{l=1}^{m_j} v_{jl}(d)$$
$$\text{for all } i,j \in \{1,\ldots,k\} \tag{9}$$

We can assign one of the n_i an arbitrary value (say $n_1 = 1$); then the other normalization constants can be computed by elementary transformations of equations 9.

4.2 Meta Classification

For meta classification we are given a set $V = \{v_1, \ldots, v_k\}$ of k binary classifiers, obtained by supervised learning based on the features for distinct document representations, with Results $R(v_i, d)$ in $\{+1, -1, 0\}$ for a document d, namely, $+1$ if d is accepted for the given topic by v_i, -1 if d is rejected, and 0 if v_i abstains. We can combine these results into a meta result: $Meta(d) = Meta(R(v_1, d), \ldots, R(v_k, d))$

in $\{+1, -1, 0\}$ where 0 means abstention. A family of such meta methods is the linear classifier combination with thresholding [39]. Given thresholds t_1 and t_2, with $t_1 > t_2$, and weights $w(v_i)$ for the k underlying classifiers we compute $Meta(d)$ as follows:

$$Meta(d) = \begin{cases} +1 & \text{if } \sum_{i=1}^{n} R(v_i, d) \cdot w(v_i) > t_1 \\ -1 & \text{if } \sum_{i=1}^{n} R(v_i, d) \cdot w(v_i) < t_2 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

This meta classifier family has some important special cases, depending on the choice of the weights and thresholds:
1) voting [8]: Meta returns the result of the majority of the classifiers.
2) unanimous decision: if all classifiers give us the same result (either +1 or -1), Meta returns this result, 0 otherwise.
3) weighted averaging [43]: Meta weighs the classifiers by using some predetermined quality estimator, e.g., a leave-one-out or k-fold-crossvalidation estimator for each v_i.

The restrictive and tunable behavior is achieved by the choice of the thresholds: we dismiss the documents where the linear result combination lies between t_1 and t_2. For real world data there is often a tradeoff between the fraction of dismissed documents (the *loss*) and the fraction of correctly classified documents (the *accuracy*). The idea of restrictive classification is to classify a subset of the test documents, but to do so with a higher reliability.

If a fixed set U of unlabeled documents (that does not change dynamically) is given, we can classify the documents with a user-acceptable loss of L as follows:

1. for all documents in U compute their classification confidence $\sum_{i=1}^{n} R(v_i, d) \cdot w(v_i)$
2. sort the documents into decreasing order according to their confidence values
3. classify the $(1-L)|U|$ documents with the highest confidence values according to their sign and dismiss the rest

In our experiments we assigned equal weights to each classifier, and instead of $R(v_i, d)$, we considered a "confidence" value $conf(v_i, d)$ for the classification of document d by the classifier. For SVM we considered the SVM scores, i.e., the distance of the test points from the hyperplane. A more enhanced method to map SVM outputs to probabilities is described, e.g., in [34].

5. EXPERIMENTS

5.1 Setup

For the validation of the presented techniques, we considered a literature data set obtained from the Gutenberg Project [1], a volunteer effort to digitize, archive, and distribute cultural works. We selected 10 English and American authors with a sufficient number of books (listed in Table 2). For each author we divided each book into parts with 20 paragraphs and stored each part as a document in the database. From these documents, we randomly choose 600 per class for our experiments. We divided these documents, that we obtained for each author a training set (100 documents) and an evaluation set (500 documents).

For our experiments we considered binary classification on all 45 possible pairs of authors (e.g. "Burton" vs. "Dickens"). For every pair we chose $T \in \{20, 40, 60, 80, 100\}$ documents from the authors' training sets as positive and the same number of documents as negative samples. The classification was performed on the union of both evaluation sets.

Then, we computed the micro-averaged *error*, i.e. the ratio of incorrectly classified documents to all test documents. For restrictive meta classification we considered in addition the *loss*, the fraction of documents dismissed by the restrictive classifier. Additionally, we computed the 95 percent confidence interval for the error.

We compared the following methods for feature construction:

1. word based features
   (a) Bag-of-Words using porter stemming and stopword elimination - see Section 3.1.1 (**BoW**)
   (b) Function words - see Section 3.1.2 (**FW**)
   (c) Part of Speech extraction of nouns and verbs; annotation with Connexor Machinese Phrase Tagger, using base forms of words constructed by Connexor - see Section 3.1.3 (**N&V**)
   (d) n-grams within constituents; using the Stanford Lexparser, considering constituents of each sentence represented as PCFG-tree - see Section 3.2 (**Constit.**)
2. structure based features
   (a) functional dependencies using Connexor Machinese Syntax for dependency tagging - see Section 3.3 (**FunctDep**)
   (b) writing style using the Stanford Lexparser - see Section 3.4 (**Style**)
   (c) histograms for syntax tree depth distribution - see Section 3.5.2 (**Hist.**)
3. combination vectors using Bag-of-Words, writing style, and tree depth histograms - see Section 4.1 - (**Combi**)

As classification method we chose standard linear SVM with parameter $C = 1000.0$. We used the popular *SVMlight* implementation [19].

5.2 Results

In our first experiment we compared the classification results of the different feature construction methods and their combination (see Table 3, Figure 3 for a chart representation of Bag-of-Words - the best base classifier - vs. combination methods). As meta method we used a simple unanimous decision (**Unanimous Decision**) classifier with base classifiers based on Bag-of-Words, writing style, and tree depth histograms.

In a second experiment we took the confidence values for classification into account and induced different loss values for the meta classification as described in Section 4.2 (Table 4 and Figure 4).

The main observations are:

- The stylistic features work significantly better than random; nevertheless Bag-Of-Words provides us with better results. Obviously, in the Gutenberg corpus there is a high correlation between authors and topics as well as distinct word pools.

Author	# Books	# Test Documents
Richard Burton	49	7425
Charles Dickens	55	7869
Arthur Conan Doyle	40	3473
Henry Rider Haggard	55	6882
George Alfred Henty	60	7169
Jack London	38	3566
Edgar Allan Poe	7	636
William Shakespeare	89	4025
Robert Louis Stevenson	45	6451
Mark Twain	129	9087

Table 2: Authors used from Gutenberg Corpus

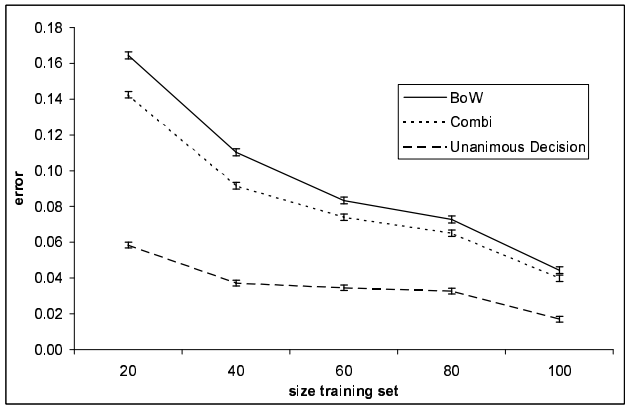


Figure 3: Comparison: Bag-of-Words and Combination Techniques on the Gutenberg Corpus

- By combining Bag-Of-Words with the alternative features, we obtained significant improvements. For combination vectors we have especially improvements for a low number of training documents. With restrictive meta methods we accept a certain loss, but obtain a much lower error on the remaining documents.

6. RELATED WORK

There is considerable prior work about alternatives to the Bag-Of-Words approach for document classification [28]. These include: using Part-Of-Speech (POS) tags ("verbs", "nouns", "adjectives", etc.) [33] either for disambiguation or for feature selection, using a thesaurus like Wordnet [15] for feature construction [38, 36], and feature selection based on statistical measures like Mutal Information or Information Gain [46]. N-grams of characters are popular for distinguishing different languages [10, 7]; also word based n-grams and phrases were examined for the text classification task [40, 24].

The problem of authorship attribution is different from the classical topic based classification task. Here, stylometric features may become important [17]. Baayen et. al. [4] show the occurrence of some kind of "stylistic fingerprint" for authors by considering a text corpus produced by student writers of different age and education level. They use the most frequent function words and apply principal component analysis (PCA) as well as linear discriminant analysis

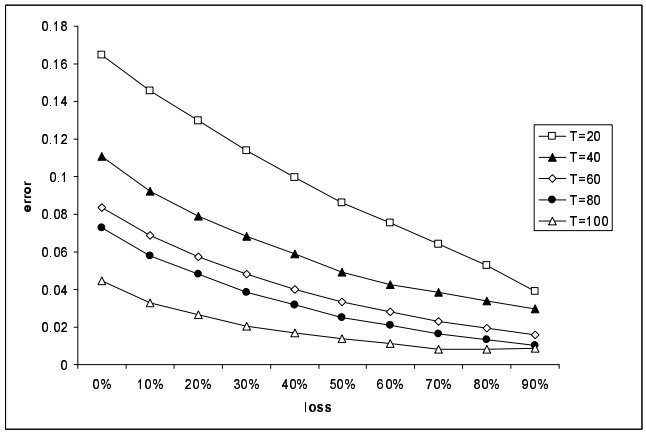


Figure 4: Comparison: Classification Results for Restrictive Meta Classification on the Gutenberg Corpus

Loss	T = 20	T = 40	T = 60	T = 80	T = 100
0 %	0.165	0.111	0.083	0.073	0.044
10 %	0.146	0.092	0.069	0.058	0.033
20 %	0.130	0.079	0.058	0.048	0.027
30 %	0.114	0.068	0.048	0.039	0.021
40 %	0.099	0.059	0.040	0.032	0.017
50 %	0.086	0.049	0.033	0.025	0.014
60 %	0.075	0.043	0.028	0.021	0.011
70 %	0.064	0.038	0.023	0.016	0.008
80 %	0.053	0.034	0.019	0.013	0.008
90 %	0.039	0.030	0.016	0.010	0.009

Table 4: Error for Different User-Provided Loss Values using a Meta Classifier with BoW, Writing Style, and Functional Dependencies on the Gutenberg Corpus

(LDA).

Diederich et al. [13] present a study on authorship attribution with Support Vector Machines. Their feature set consists of "full word forms" (in fact Bag-Of-Words) and so called tagwords, a combination of function words and grammatical information. Here, simple Bag-Of-Words outperforms their combination techniques with more enhanced linguistic features, in contrast to our combination vectors and meta methods.

In [5], Baayen et al. present a methodological study on the usefulness of stylometry-based features. They investigate features related to the writing style technique described above, taking grammatical rewriting rules derived from syntax trees into account.

The identification of unique users among a set of online pseudonyms using features such as simple words, misspellings, punctuation etc., is described in [31]. De Vel's work [12] deals with the exploration of style based features for identification of email authors. They use features such as style markers (average sentence or word length, total number of function words, vocabulary richness, etc.) and structural attributes (availability of signatures, number of attachments, etc.).

There are also several alternative learning paradigms for authorship attribution, e.g., Khmelev and Tweedie [21] con-

T	BoW error	FW error	N&V error	Style error	FunctDep error	Hist. error	Constit. error	Bigrams error
20	0.164 ±0.0034	0.354 ±0.0044	0.225 ±0.0039	0.186 ±0.0036	0.171 ±0.0035	0.299 ±0.0042	0.356 ±0.0044	0.458 ±0.0046
40	0.110 ±0.0029	0.240 ±0.0039	0.159 ±0.0034	0.138 ±0.0032	0.123 ±0.0030	0.278 ±0.0041	0.279 ±0.0041	0.390 ±0.0045
60	0.083 ±0.0026	0.177 ±0.0035	0.096 ±0.0027	0.123 ±0.0030	0.123 ±0.0030	0.273 ±0.0041	0.245 ±0.0040	0.323 ±0.0043
80	0.073 ±0.0024	0.147 ±0.0033	0.084 ±0.0026	0.114 ±0.0029	0.116 ±0.0030	0.275 ±0.0041	0.221 ±0.0038	0.285 ±0.0042
100	0.044 ±0.0019	0.115 ±0.0030	0.065 ±0.0023	0.103 ±0.0028	0.089 ±0.0026	0.272 ±0.0041	0.204 ±0.0037	0.230 ±0.0039

T	Combination error	Meta Unanimous Decision error	Meta Unanimous Decision loss
20	0.142 ±0.0032	0.059 ±0.0016	0.482
40	0.092 ±0.0027	0.037 ±0.0014	0.399
60	0.074 ±0.0024	0.035 ±0.0013	0.362
80	0.065 ±0.0023	0.033 ±0.0013	0.349
100	0.040 ±0.0018	0.017 ±0.0010	0.345

Table 3: Error for Classification based on Different Features and their Combination on the Gutenberg Corpus

sidering learning models for authorship attribution tasks using Markov chains of characters, or Oakes [32] using a kind of swarm intelligence simulation technique called Ant Colony Optimization.

Combination vectors are used for authorship attribution (e.g. [42, 23, 13]), but neither explicit component weighting nor normalization are considered. The machine learning literature has studied a variety of meta methods such as bagging, stacking, or boosting [8, 44, 25, 16], and also combinations of heterogeneous learners (e.g., [47]). But, to our knowledge, meta classification was not applied in the context of authorship recognition.

7. CONCLUSION AND FUTURE WORK

In this paper we described classification with different document representations. In addition to well known features like document terms in the Bag-Of-Words model, POS tagging, etc., we considered alternative stylistic features like the depth or the structure of syntax trees. We combined the feature representations using two techniques: 1) combination vectors, where we constructed a single vector from the different feature vectors with automatically normalizing the combination vector's components so that the average component value is the same for all subspaces, 2) meta methods combining the classification results based on the different representations into a meta result. Our experiments on the author recognition task show that our new features are suitable for discriminating different styles and, used within combination techniques, lead to significant improvements of the classifier performance.

Our ongoing and future work includes a number of relatively obvious directions like 1) the improvement of existing and the construction of new alternative features, 2) application of different feature spaces and their combination for clustering, 3) the use of enhanced features for expert queries in specialized search engines.

8. REFERENCES

[1] Gutenberg project. *http://www.gutenberg.org/*.

[2] Lexparser. *http://www-nlp.stanford.edu/downloads/lex-parser.shtml*.

[3] National institute of standards and technology.

[4] H. Baayen, H. van Halteren, A. Neijt, and F. Tweedie. An experiment in authorship attribution. *JADT*, 2002.

[5] H. Baayen, H. van Halteren, and F. Tweedie. Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 11(3):121–131, 1996.

[6] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[7] K. R. Beesley. Language identifier: A computer program for automatic natural-language identification on on-line text. In *29th Annual Conference of the American Translators Association*, 1988.

[8] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[9] C. Burges. A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 1998.

[10] W. B. Cavner and J. M. Trenkle. Text categorization and information retrieval using wordnet senses. In *Third Annual Symposium on Document Analysis and Information Retrieval*, 1994.

[11] S. Chakrabarti. *Mining the Web*. Morgan Kaufmann, 2003.

[12] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30(4):55–64, 2001.

[13] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1-2):109–123.

[14] S. Dumais and H. Chen. Hierarchical classification of Web content. *SIGIR*, 2000.

[15] C. Fellbaum. *WordNet: An Electronic Lexical*

Database. MIT Press, 1998.

[16] Y. Freund. An adaptive version of the boost by majority algorithm. *Workshop on Computational Learning Theory*, 1999.

[17] D. Holmes. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(9):111–117, 1998.

[18] D. Holmes and R. Forsyth. The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, 10(2):111–127, 1995.

[19] T. Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. *ECML*, 1998.

[20] T. Joachims. A statistical learning model of text classification for Support Vector Machines. *SIGIR*, 2001.

[21] D. Khmelev and F. Tweedie. Using Markov Chains for Identification of Writers. *Literary and Linguistic Computing*, 16(3):299–308, 2001.

[22] D. V. Khmelev and W. J. Teahan. A repetition based measure for verification of text collections and for text categorization. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 104–110, New York, NY, USA, 2003. ACM Press.

[23] M. Koppel, S. Argamon, and A. Shimoni. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[24] D. Lewis. *Representation and learning in information retrieval*. PhD thesis, Department of Computer and Information Science, University of Massachusetts, 1992.

[25] N. Littlestone and M. Warmuth. The weighted majority algorithm. *FOCS*, 1989.

[26] C. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[27] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[28] A. Moschitti and R. Basili. Complex linguistic features for text classification: A comprehensive study. In *ECIR*, 2004.

[29] F. Mosteller and D. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

[30] N. Nanas, V. Uren, and A. de Roeck. Learning with positive and unlabeled examples using weighted logistic regression. In *15th International Workshop on Database and Expert Systems Applications(DEXA'04), Zaragoza, Spain*, 2004.

[31] J. Novak, P. Raghavan, and A. Tomkins. Anti-aliasing on the web. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 30–39. ACM Press, 2004.

[32] M. Oakes. Ant colony optimisation for stylometry: The federalist papers. In *Proceedings of the 5th International Conference on Recent Advances in Soft Computing*, pages 86–91. Nottingham Trent University, 2004.

[33] B. Pang and L. Lee. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP*, 2002.

[34] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers, MIT Press*, 1999.

[35] M. Porter. An algorithm for suffix stripping. *Automated Library and Information Systems*, 14(3).

[36] P. Rosso, E. Ferretti, D. Jimenez, and V. Vidal. Text categorization and information retrieval using wordnet senses. In *The Second Global Wordnet Conference GWC*, 2004.

[37] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 1988, p. 513-523.*

[38] S. Scott and S. Matwin. Text classification using wordnet hypernyms. In *Workshop on Usage of Wordnet in Natural Language Processing Systems*, 1998.

[39] S. Siersdorfer, S. Sizov, and G. Weikum. Goal-oriented methods and meta methods for document classification and their parameter tuning. In *ACM Conference on Information and Knowledge Management (CIKM 04), Washington*, 2004.

[40] C. M. Tan, Y. F. Wang, and C. D. Lee. The use of bigrams to enhance text categorization. *Information Processing and Management., vol. 30, No. 4, pp. 529-546, 2002.*

[41] P. Tapanainen and T. Jõrvinen. A non-projective dependency parser. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 1997.

[42] H. van Halteren. Writing Style Recognition and Sentence Extraction. *Workshop on Text Summarization, DUC*, 2002.

[43] H. Wang, W. Fan, P. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. *SIGKDD*, 2003.

[44] D. Wolpert. Stacked generalization. *Neural Networks, Vol. 5, pp. 241-259*, 1992.

[45] Y. Yang. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1(1/2), 1999.

[46] Y. Yang and O. Pedersen. A comparative study on feature selection in text categorization. *ICML*, 1997.

[47] H. Yu, K. Chang, and J. Han. Heterogeneous learner for Web page classification. *ICDM*, 2002.

Computational Stylistics in Forensic Author Identification

Carole E. Chaski
Institute for Linguistic Evidence, Inc
25100 Trinity Drive
Georgetown, Delaware 19947-6585
USA
1-302-856-9488
cchaski@linguisticevidence.org,
cchaski@aol.com

ABSTRACT
This paper describes criteria for forensic authorship attribution which meet both legal and scientific requirements, and demonstrates that computational stylistics can readily meet these criteria. A forensic author identification method based on syntactic analysis and discriminant function analysis has previously attained a cross-validated accuracy rate of 95%, but the method was not subjected to a "real-life" simulation of actual forensic casework. This paper presents the results of 618 tests of questioned documents drawn from author-pairs. Results show that seven of the ten authors are accurately identified as the questioned document's author at least 80% of the time, with three of the ten authors are accurately identified as the author less reliably.

Categories and Subject Descriptors
J.5 Linguistics, I.5.3 Clustering

General Terms
Reliability, Experimentation, Legal Aspects.

Keywords
Authorship attribution, author identification, Questioned document examination, Computational stylistics.

1. INTRODUCTION
The style of a document can become crucial evidence in the forensic setting when a document's authorship is questioned. If the document is e-mail or word-processed, the style factor becomes the most important means of determining authorship because traditional techniques such as handwriting examination or ink analysis naturally do not apply [3, 7. 8]. In addition to the pervasive spread of electronic documents, legal developments in the United States and Europe have also demonstrated the need for a scientifically-grounded and empirically reliable method for determining authorship [4, 10, 11]. Thus, the time is ripe for computational stylistics in forensic author identification.

In this article, I describe four criteria for forensic author identification, all of which can be met by the computational stylistics paradigm. Second, I briefly summarize some recent results of experimentation related to author identification in the forensic setting. Third, I present the results of an experiment simulating the most common author identification problem, two potential authors of one questioned document [10].

2. FOUR CRITERIA FOR FORENSIC AUTHOR IDENTIFICATION
Authorship attribution in the forensic setting must meet certain criteria in order to be admitted as scientific evidence or entertained seriously as investigative support. These criteria are linguistic defensibility, forensic feasibility, statistical testability, and reliability.

First, the method must be linguistically defensible. Basic assumptions about language structure, language use, and psycholinguistic processing should undergird the method. The linguistic variables which are ultimately selected should be related in a straightforward way to linguistic theory and psycholinguistics; the linguistic variables should be justifiable. For example, function words have been used in many lexical approaches to authorship attribution, perhaps most famously by Mosteller and Wallace [11]. Function words can be justified as a potential discriminator for two reasons: first, function words are a lexical closed class, and second, function words are often indicators of syntactic structure. Psycholinguistically, function words are known as a distinct class for semantic processing and the syntactic structures which function words shadow are known to be real. A method based on function words is linguistically defensible because there is a fairly obvious way for a linguist to relate this class of discriminators to what we already know about language structure and psycholinguistic processing.

Second, the method must be forensically feasible. Specifically, a forensically feasible method must be sensitive to the actual limitations of real data and the basis of expert opinion. Foremost, the method must be designed to work within the typical forensic situation of brevity and scarcity of texts. The importance of this criterion can not be ignored because forensic feasibility will impact both the selection of linguistic variables as well as the selection of statistical procedures. Many of the lexical approaches which have been developed within literary studies have rightfully exploited the lexical richness and high word counts of such literary data, but these same approaches are not forensically feasible because the typical forensic data is too short or too lexically restricted. Further, statistical procedures which require hundreds of cases to fit a large number of variables are not always forensically feasible because in the typical forensic situation there are not hundreds of texts to be analyzed. Due to the scarcity of texts, either the texts can be separated into smaller units to provide additional cases or the linguistic variables can be collapsed. But in either text-decomposition or variable-reduction, again

linguistic defensibility must be maintained. For example, it was once suggested that split-half reliability testing be performed at the word level: every other word of a document was extracted and that extracted portion was tested against the remainder of the original document [10]. While this kind of text-decomposition is understandable as a way of dealing with the scarcity of texts, this particular technique is linguistically indefensible because, by relying on a basic assumption that language is just a "bag of words" rather than a structured system, the approach totally ignores the fact that there is a linearized and syntactic structure in text which is psychologically real to the author of the document.

Another impact of the forensic feasibility criterion concerns the basis of expert opinion. In the forensic setting, the expert witness stakes his or her reputation on the accuracy of the data analysis. Therefore, any "black box" methods which are automatized to the extent that the analyst cannot supervise, error-correct or otherwise intervene in the basic data analysis may not be acceptable to forensic practitioners or linguists who do not wish to serve as mere technician-servants of the machine. On the other hand, automatization of many types of linguistic analysis provides a welcome way to avoid examiner bias and fatigue. The best approach, therefore, appears to be an interactive, user-assisted automatic computerized analysis, since the machine can provide objective, rule-based analysis and the human can correct any analytical errors the machine might make.

Third, the method must be statistically testable. Specifically, this criterion requires that the linguistic variables—even if they are categorical—can be operationally defined and reproduced by other linguists. This criterion does not reject categorical linguistic variables which may have their basis in qualitative analysis, but it does reject subjective reactions to style such as "sounds like a Clint Eastwood movie" or "not what a blue-collar worker would write." These quotations are not facetious, but actual comments from experts whose reports I have personally read.

Fourth, the method must be reliable, based on statistical testing. The level of reliability can be obtained through empirical testing. Naturally, the most accurate method is most welcome in the forensic setting, but even a method with an empirically-based, statistically-derived overall accuracy rate of only 80% or 90% is better than any method whose reliability is unproven, untested, anecdotal or simply hypothesized and then stated as accomplished fact.

If an authorship attribution method meets these scientific criteria, it will surely meet success within the legal arena under the Daubert-Joiner-Kumho criteria as well as the Frye standard. Linguistic defensibility speaks to general acceptance among peers; linguists are certainly far more likely to accept any method which is based on standard techniques of linguistic theory as well as conceptions of language congruent with linguistic theory and psycholinguistic experimentation than one based on prescriptive grammar or literary sensibility. Forensic feasibility speaks to the appropriate application of the method to typical forensic data and the credibility of the testimony. Finally, both statistical testing and reliability speak to the error rate, and again, the credibility and weight of the testimony. It is obvious to anyone familiar with computational stylistics that this research paradigm can produce a method which meets these four criteria.

3. RECENT RESULTS IN THE COMPUTATIONAL STYLISTICS PARADIGM

Within the last five years, some very exciting work has been conducted by deVel and his colleagues on e-mail authorship [8], Argamon and his colleagues on multi-author classification [1], Koppel and his colleagues on variable selection [9].

Using machine learning algorithms, deVel et al [8] recorded performance rates from 60% to 100% with style variables related to formatting, lexical metrics, function words and punctuation totals. Argamon et al [1] present very promising work for multi-author classification of newsgroup posts, using function words, net abbreviations, lexical metrics and formatting. This work may be especially important in security situations where sorting of massive amounts of documents is required, rather than the usual forensic situation in which there is a narrow pool of potential authors who can be tested individually against the questioned document [13]. Koppel and Schler [9] tested part-of-speech tags and intuition-based style variables (known as idiosyncracies in traditional handwriting examination), also using machine learning algorithms, with performance results ranging from 37% to 72%. This is an especially interesting result because this is the first and only time that many such style variables have ever been subjected to empirical error-rate testing, although some were also tested by Chaski [4] with similar results using a much simpler statistical test. Meanwhile, the non-computational, non-statistical proponents of the intuition-based style variables have not produced any error rate experiments and their testimony is being restricted or excluded [13, 5].

Also within the computational stylistics paradigm, Baayen et al [1], Stamatatos et al [14, 15] and Tambouratzes et al [16] have been using discriminant function analysis with lexical, syntactic and punctuation variables. These studies show a remarkably consistent performance of cross-validated discriminant function analysis with such stylometric features, since each of these report performance rates in the 87% to 89% range. Chaski [6,7] reports an overall performance rate of 95% using syntactic markedness, syntactically-classified punctuation and average word length in a cross-validated discriminant function analysis. This result surely suggests that these variables are worth pursuing in further research. However, Chaski's [6, 7] result is based on an experimental design which does not mirror an actual forensic case analysis, in which two suspects' documents are compared to one questioned document. This "real-life" experimental design is the one reported below in Section 4.

4. EXPERIMENTAL SUBJECTS, DATA AND DESIGN

Based on sociolinguistically-relevant demographics and the amount of text, ten authors were drawn from Chaski's Writing Sample Database, a collection of writings on particular topics designed to elicit several registers such as narrative, business letter, love letter and personal essay [3, 4]. Sociolinguistically-relevant demographics include sex, race, education and age. These demographic features can be used to define dialects. Controlling for these features tests the ability to differentiate authors at an individual rather than group level. Although this dataset was not as tightly constrained as the dataset in Chaski [4], because it includes both men and women and a wider age range, this dataset has been controlled for race and education. The five women and five men are all white adults who have completed high school up to three years of college at open-admission colleges. The authors range in age from 18 to 48. The authors all have extensive or lifetime experience in the American English, Delmarva dialect of the mid-Atlantic region of the United States. The authors are "naïve writers" (in terms of Baayen, et al [1]) with similar background and training. The authors volunteered to write, wrote at their leisure, and were compensated for their writings through grant funding from the National Institute of Justice, US Department of Justice.

Another control for the dataset is the topic. Controlling the topic tests the ability to differentiate authors even though they are writing about the same topic. The authors all wrote on similar topics, listed in Table 1, but they wrote over a range of topics and registers.

Table 1: Topics in the Writing Sample Database

Task ID	Topic
1.	Describe a traumatic or terrifying event in your life and how you overcame it.
2.	Describe someone or some people who have influenced you.
3.	What are your career goals and why?
4.	What makes you really angry?
5.	A letter of apology to your best friend
6.	A letter to your sweetheart expressing your feelings
7.	A letter to your insurance company
8.	A letter of complaint about a product or service
9.	A threatening letter to someone you know who has hurt you
10.	A threatening letter to a public official (president, governor, senator, councilman or celebrity)

Further, the author selection took into consideration the quantity of writing which the authors had produced. Authors who met the sociolinguistic demographics but produced only three documents were not included in this dataset lest the lack of data produce misleading results. In order to have enough data for the statistical procedure to work, but in order to make this experiment as forensically feasible as possible, the number of documents for each author was determined by however many were needed to hit targets of approximately 100 sentences and/or 2,000 words. One author needed only 4 documents to hit both targets, while two authors needed ten documents. Three authors needed 6 documents to hit the sentences target but only one of these three authors exceeded the words target. The exact details are shown in Table 2: Authors and Texts.

Table 2: Authors and Texts

Race, Gender	Topics by Task ID	Author ID Number	Number of Texts	Number of Sentences	Number of Words	Average in Words (Min, Max)
WF	1 - 4, 7, 8	16	6	107	2,706	430 (344, 557)
WF	1 - 5	23	5	134	2,175	435 (367, 500)
WF	1 - 10	80	10	118	1,959	195 (90, 323)
WF	1 - 10	96	10	108	1,928	192 (99, 258)
WF	1 - 3, 10	98	4	103	2,176	543 (450, 608)
WF Total			**35**	**570**	**10,944**	
WM	1 - 8	90	8	106	1,690	211 (168, 331)
WM	1 - 6	91	6	108	1,798	299 (196, 331)
WM	1 - 7	97	6	114	1,487	248 (219, 341)
WM	1 - 7	99	7	105	2,079	297 (151, 433)
WM	1 - 7	168	7	108	1,958	278 (248, 320)
WM Total			**34**	**541**	**9,012**	
Grand Total			69	1,111	19,956	

Each text was processed using ALIAS, a program developed by Chaski [3, 4] for the purpose of databasing texts, lemmatizing, computing lexical frequency ranking, calculating lexical, sentential and text lengths, punctuation-edge counting, Part-Of-Speech-tagging n-graph and n-gram sorting, and markedness subcategorizing. ALIAS is thus able to provide a large number of linguistic variables. In this study, however, only three types of variables are used: punctuation classified by the syntactic edge it marks, syntactic structures categorized by markedness, and average word length (including both content and function words in the computation).

While a thorough description of the variables is provided in Chaski [6, 7], for the reader's sake, these three variable types are described again herein.

Chaski (2001) showed that syntactically-classified punctuation had a slighter better performance than simple punctuation marks for discriminating authors while preserving intra-author classification. Authors may share the same array of marks, but the placement of the marks appears to be what matters. This approach to using punctuation as an authorial identifier is very different from the approach advocated by questioned document examination (Hilton, 1993), forensic stylistics (McMenamin 2003), as well as the computational stylometric studies discussed earlier.

After each text is automatically split into sentences, the user interacts with ALIAS to categorize punctuation within each sentence by the syntactic edge which it marks. These syntactic edges are the clause, the phrase and the morpheme (word-internal). For example, the end-of-clause (EOC) marks may be commas, semi-colons, hyphens; the particular marks are not counted separately, but any and every EOC mark is counted. There are three variables for syntactically-classified punctuation. ALIAS then exports these counts to a spreadsheet.

Language is structured by binary distinctions, asymmetrically, so that one member of the binary opposition is less frequent, more restricted and in other ways more marked than the other. Markedness is the basic asymmetry in language which pervades the binary substructure of linguistic signs. The unmarked contrast is the most common and often the most easily parsed, while the marked contrast is typically less frequent and sometimes more difficult to parse because it can pose several different parsing attachments. For example, the head-position of the noun, in universal terms, can be either initial or final (the binary contrast). This head-position parameter distinguishes English and Spanish, since in simple noun phrases, the English noun will be in final position, after any modifiers, while the Spanish noun will be in initial position, before any modifiers. The unmarked noun phrase in English is head-final while the unmarked noun phrase in Spanish is head-initial. But in both English and Spanish, the marked variants are possible. In English, noun phrases which are not head final include head-medial structures such as NP[DP AP N PP] and NP[DP AP N CP] as well as the more rare head-initial structure NP[N AP] (such as 'forest primeval'). Phrases around each head (noun, verb, adjective, preposition, modifier) are parsed.

After each word is tagged for its part-of-speech, ALIAS searches and sorts syntactic head patterns and flags the pattern exemplars as either marked or unmarked. These flags can be checked by the user. ALIAS then counts the marked and unmarked exemplars for each syntactic head, collapse them into two variables (marked XP and unmarked XP) and outputs these counts to a spreadsheet for statistical analysis.

Finally, one lexical variable was included. Following the lead of Tambouratzis et al. (2004) and many other stylometric studies, average word length for each document was computed. All words, both function and content words, were included in this computation.

In sum, as listed below, there are three syntactically-classified punctuation variables, two syntactic markedness variables and one lexical variable, for a total of six variables per document.

A method based on these same variables without any collapsing of the heads to marked or unmarked XP is also available for use at the sentence level (rather than document level), and results have also been higher than 90% accuracy, but the document level analysis is the version which was used in the experiment reported here.

The design of this experiment followed the usual case analysis: one questioned document either belongs to author A or author B. There are known authenticated documents of author A and author B. Discriminant function analysis can be used to differentiate between the known writings of A and B, and then apply this function to classify the questioned document.

Each author was paired with each other author. For each author pair, each document was treated as a questioned document. That is, (1) for author pair 16 and 23, document 16-1 was removed, (2) a discriminant function analysis, with cross-validation, was run on the remaining documents by author 16 and the documents by author 23, (3) the function classified document 16-1 as either belonging to author 16 or author 23. This was repeated for all of the documents by 16 and by 23. Thus, given the number of documents and number of pairs, 618 classification tests were run.

SPSS was used to run each discriminant function analysis. The SPSS options were set to stepwise entry of the variables, using the Mahalanobis distance, and keeping F to enter and F to remove at the default SPSS settings. Given these settings, there are times when no variables qualify. In order to produce as standard an analysis as possible, these settings were not manipulated. Previous work by Chaski [6] showed that using the stepwise entry, Mahalonobis distance and default values for F resulted in the highest accuracy rates with the syntactic markedness and syntactically-classified punctuation variables.

5. RESULTS

Table 3 shows the results of these tests. The number of documents for each author is listed in parentheses in the cell under the Author ID. If another parenthesized number occurs in this column, it denotes the number of documents which were able to be tested with the discriminant function analysis (those remaining in case some documents sets resulted in no variables qualifying). For each author, the final bolded hits and misses percents are averages of the hits and misses percents for the documents paired with each other author.

For these ten authors, the average hits percents range from a low of 64% to a high of 89%. Three authors of the ten had average hit rates of less than 70%. Author 98, whose

questioned documents yielded the lowest discriminability at 64%, still had hit rates of 100% when paired with authors 91, 97 and 168. Author 99, whose questioned documents also yielded low discriminability at 66%, still had hit rates of 100% when paired with authors 16 and 23, and hit rates of 86% when paired with authors 98 and 168. Author 91, whose questioned documents yielded low discriminability at 69%, still had hits rates of 100% when paired with authors 16, 23 and 98. Author 91 also had the most instances of no variables qualifying for the discriminant function analysis. On the other hand, seven of the ten authors obtained average hit rates from 80% to 89%.

The detailed results are shown in Table 3.

Table 3: Results of Questioned Document Simulation Tests For Each Author

Author 16	paired with	Hits	Misses	Hits %	Misses %
(6 docs)	23	6	0	100%	0%
	80	5	1	83%	17%
	96	6	0	100%	0%
	98	4	2	67%	33%
	90	5	1	83%	17%
	91	5	1	83%	17%
	97	5	1	83%	17%
	99	5	1	83%	17%
	168	6	0	100%	0%
				87%	**13%**
Author 23	paired with	Hits	Misses	Hits %	Misses %
(5 docs)	16	4	1	80%	20%
	80	4	1	80%	20%
	96	5	0	100%	0%
	98	3	2	60%	40%
	90	4	1	80%	20%
	91	4	1	80%	20%
	97	3	2	60%	40%
	99	4	1	80%	20%
	168	5	0	100%	0%
				80%	**20%**
Author 80	paired with	Hits	Misses	Hits %	Misses %
(10 docs)	16	10	0	100%	0%
	23	10	0	100%	0%
	96	8	2	80%	20%
	98	10	0	100%	0%
	90	8	2	80%	20%
(3 docs)	91	1	2	33%	67%
	97	8	2	80%	20%
(8 docs)	99	5	3	63%	38%
	168	8	2	80%	20%
				80%	**20%**
Author 96	paired with	Hits	Misses	Hits %	Misses %
(10 docs)	16	10	0	100%	0%
	23	10	0	100%	0%
	80	6	4	60%	40%
	98	10	0	100%	0%
	90	5	5	50%	50%
	91	9	1	90%	10%
	97	8	2	80%	20%
	99	10	0	100%	0%
	168	10	0	100%	0%
				87%	**13%**
Author 98	paired with	Hits	Misses	Hits %	Misses %
(4 docs)	16	2	2	50%	50%
	23	2	2	50%	50%
	80	2	2	50%	50%
	96	2	2	50%	50%
	90	2	2	50%	50%
	91	4	0	100%	0%
	97	4	0	100%	0%
	99	1	3	25%	75%
	168	4	0	100%	0%
				64%	**36%**
Author 90	paired with	Hits	Misses	Hits %	Misses %
(8 docs)	16	8	0	100%	0%
	23	8	0	100%	0%
	80	7	1	88%	13%
	96	3	5	38%	63%

	98	8	0	100%	0%
	91	6	2	75%	25%
	97	8	0	100%	0%
	99	7	1	88%	13%
	168	7	1	88%	13%
				86%	**14%**
Author 91	paired with	Hits	Misses	Hits %	Misses %
(6 docs)	16	6	0	100%	0%
	23	6	0	100%	0%
(4 docs)	80	0	4	0%	100%
	96	3	3	50%	50%
	98	6	0	100%	0%
	90	3	3	50%	50%
(4 docs)	97	3	1	75%	25%
(4 docs)	99	3	1	75%	25%
	168	4	2	67%	33%
				69%	**31%**
Author 97	paired with	Hits	Misses	Hits %	Misses %
(6 docs)	16	6	0	100%	0%
	23	6	0	100%	0%
	80	4	2	67%	33%
	96	3	3	50%	50%
	98	6	0	100%	0%
	90	4	2	67%	33%
(5 docs)	91	5	0	100%	0%
	99	6	0	100%	0%
	168	5	1	83%	17%
				85%	**15%**
Author 99	paired with	Hits	Misses	Hits %	Misses %
(7 docs)	16	7	0	100%	0%
	23	7	0	100%	0%
(6 docs)	80	2	4	33%	67%
	96	2	5	29%	71%
	98	6	1	86%	14%
	90	4	3	57%	43%
(3 docs)	91	1	2	33%	67%
	97	5	2	71%	29%
	168	6	1	86%	14%
				66%	**34%**
Author 168	paired with	Hits	Misses	Hits %	Misses %
(7 docs)	16	7	0	100%	0%
	23	7	0	100%	0%
	80	5	2	71%	29%
	96	7	0	100%	0%
	98	7	0	100%	0%
	90	7	0	100%	0%
	91	5	2	71%	29%
	97	6	1	86%	14%
	99	5	2	71%	29%
				89%	**11%**

6. CONCLUSIONS

Seven of the ten authors, when subjected to a "real-life" simulation of forensic author identification, were identified as the correct author of the questioned document for at least 80% of their documents. Three of the ten authors were identified better than chance, but less than 70%. These results are important avenues for helping us determine the limits of such a method. Any of the best forensic methods (DNA, toxicology, reconstructive engineering) have established well-known limits for their applications. In future research, these results will be examined in detail to determine what patterns of data distribution, variable values and so forth may explicate the method's limitations. For instance, it may be that the minimal amount of documents (rather than words or sentences) is the best guide for data collection, and that the minimal amount of documents is a specific number.

The variables used in this experiment are not only linguistically-defensible, they can also be justified through many years' worth of psycholinguistic experimentation. Other linguistic features suggested in the literature such as sentence length or spelling errors have not been shown to be particularly useful at discriminating authors in the forensic setting [4, 9]. In future work, it is hoped that linguistic variables can be selected based on both theoretical, cognitive and empirical grounds.

These results also suggest that the computational stylistics paradigm in general and in particular the use of discriminant function analysis with syntactically-motivated variables are promising lines of research for developing a linguistically-

defensible, forensically feasible, empirically tested and statistically reliable method for forensic author identification.

7. ACKNOWLEDGMENTS

I wish to acknowledge my gratitude for financial support, Grant 98-LB-VX-0065-S1, from the National Institute for Justice, US Department of Justice. Opinions expressed in this article are the author's and are not necessarily the official opinion of the US Department of Justice. I also wish to thank Harry Chmelynski, Gregory Zarow and Larry Solan.

8. REFERENCES

[1] Argamon, S., Saric, M., Stein, S. S. (2003). *Style Mining of Electornic Messages for Multiple Authorship Discirmination: First Results.* KDD-2003, Washington, DC, ACM.

[2] Baayen, H., van Halteran, Hans, Neijt, Anneke, Tweedie, Fiona (2002). "An Experiment in Authorship Attribution." *Journees internationales d'Analyse statistique des Donnees Textuelles* **6**.

[3] Chaski, C. E. (1997). "Who Wrote It? Steps Toward A Science of Authorship Identification." *National Institute of Justice Journal*: 15-22.

[4] Chaski, C. E. (2001). "Empirical Evaluations of Language-Based Author Identification Techniques." *Forensic Linguistics* **8**(1): 1-65.

[5] Chaski, C. E. (2005)."Forensic Linguistics." In Cyril Wecht and John Rago, editors. *Foundations of Forensic Science and Law: Investigative Applications in Criminal, Civil and Family Justice.* Boca Raton: CRC Press.

[6] Chaski, C. E. (2005). "Discriminant Analysis Results for Authorship Attribution in the Forensic Setting." *Forensic Linguistics* **(to appear)**.

[7] Chaski, C. E. (2005). "Who's At the Keyboard? Authorship Attribution in Digital Evidence Investigations." *International Journal of Digital Evidence.* Spring. Available at www.ijde.org.

[8] deVel, O., A. Anderson, M. Corney, G. Mohay (2001). "Multi-topic E-Mail Authorship Attribution Forensics." *ACM Conference on Computer Security-Workshop on Data Mining for Security Applications.* Philadelphia, PA.

[9] Koppel, M. a. S., Jonathan (2003). "Exploiting Stylistic Idiosyncrasies for Authorship Attribution." Available at www.cs.biu.ac.il/aaahtmlfiles/indexpeoplefiles/fmembers.html.

[10] Miron, M. S. (1983). Content Identification of Communication Origin. Advances in Forensic *Psychology and Psychiatry*. R. W. Reiber. Norwood, New Jersey, Ablex.

[11] Mosteller, F., Wallace, D L (1984). *Applied Baesian and Classical Inference: The Case of the Federalist Papers.* New York, Springer-Verlag.

[12] Risinger, D. M., and Saks, M.J. (1996). "Science and nonscience in the courts: Daubert meets handwriting identification expertise." *Iowa Law Review* **82**(1): 21-74.

[13] Solan, L. M., Tiersma, Peter M (2005). *Speaking of Crime: The Language of Criminal Justice.* Chicago, University of Chicago Press.

[14] Stamatatos, E., Fakotakis, N, Kokkinakis, G (2000). "Automatic Text Categorization in Terms of Genre and Author." *Computational Linguistics* **26**(4): 471-495.

[15] Stamatatos, E., Fakotakis, N., Kokkinakis, G. (2001). "Computer-Based Authorship Attribution Without Lexical Measures." *Computers and the Humanities* **35**: 193-214.

[16] Tambouratzis, G., Markantonatou, Stella, Hairetakis, Nikolaos, Vassiliou, Marina, Carayannis, George, Tambouratzis, Dimitrios (2004). "Discriminating the Registers and Styles in the Modern Greek Language -- Part 2: Extending the feature Vector to Optimize Author Discrimination." *Literary & Linguistic Computing* **19**(2): 221-242.

Experiments with Mood Classification in Blog Posts

Gilad Mishne
Informatics Institute, University of Amsterdam
Kruislaan 403, 1098SJ Amsterdam, The Netherlands
gilad@science.uva.nl

ABSTRACT

We present preliminary work on classifying blog text according to the mood reported by its author during the writing. Our data consists of a large collection of blog posts – online diary entries – which include an indication of the writer's mood. We obtain modest, but consistent improvements over a baseline; our results show that further increasing the amount of available training data will lead to an additional increase in accuracy. Additionally, we show that the classification accuracy, although low, is not substantially worse than human performance on the same task. Our main finding is that mood classification is a challenging task using current text analysis methods.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing

Keywords

Subjective content, blogs, moods

1. INTRODUCTION

With the increase in the web's accessibility to the masses in the last years, the profile of web content is changing. More and more web pages are authored by non-professionals; part of this "publishing revolution" is the phenomenon of *blogs* (short for web-logs) – personal, highly opinionated journals publicly available on the internet. The *blogspace* – the collective term used for the collection of all blogs – consists of millions of users who maintain an online diary, containing frequently-updated views and personal remarks about a range of issues.

The growth in the amount of blogs is accompanied by increasing interest from the research community. Ongoing research in this domain includes a large amount of work on social network analysis, but also content-related work, e.g., [5, 7]. Some of this work is intended to develop technologies for organising textual information not just in terms of topical content, but also in terms of metadata, subjective meaning, or stylometric aspects. Information needs change with the type of available information; the increase in the amount of blogs and similar data drives users to access textual information in new ways – for example, analyzing consumers' attitudes for marketing purposes [16].

In this paper, we address the task of classifying blog posts by mood. That is, given a blog post, we want to predict the most likely state of mind with which the post was written: whether the author was depressed, cheerful, bored, and so on. As in the vast majority of text classification tasks, we take a machine learning approach, identifying a set of features to be used for the learning process.

Mood classification is useful for various applications, such as assisting behavioral scientists and improving doctor-patient interaction [8]. In the particular case of blog posts (and other large amounts of subjective data), it can also enable new textual access approaches, e.g., filtering search results by mood, identifying communities, clustering, and so on. It is a particularly interesting task because it also offers a number of scientific challenges: first, the large variety in blog authors creates a myriad of different styles and definitions of moods; locating features that are consistent across authors is a complex task. Additionally, the short length of typical blog entries poses a challenge to classification methods relying on statistics from a large body of text (these are often used for text classification). Finally, the large amounts of data require a scalable, robust method.

The main research questions addressed in this paper are:

- In what way does mood classification in blogs differ from mood classification in other domains? Many types of features seem to be good indicators of mood – which ones are effective in blogs? How complex is the task to begin with?
- How much data is required for reliable training, and how many features are required for each instance of the data? It has been observed that, for NLP tasks, continuously increasing the training set size improves results consistently [1]; does this hold in our domain?

Put differently, the paper is largely exploratory in nature, taking a large collection of blog posts, broad sets of features, and varying the amount of training data exploited by the

Style2005, Stylistic Analysis Of Text For Information Access

machine learner, evaluating the effect on the classification accuracy.

The remainder of the paper is organized as follows. In Section 2 we survey existing work in affect analysis and related fields. In Section 3 we describe the collection of blog posts we use for our experiments. Section 4 follows with details regarding the features we used for the classification process, dividing them into sets of related features. Our experiments and results are reported in Section 5, and we conclude in Section 6.

2. RELATED WORK

Most work on text classification is focused on identifying the topic of the text, rather than detecting stylistic features [28]. However, stylometric research – in particular, research regarding emotion and mood analysis in text – is becoming more common recently, in part due to the availability of new sources of subjective information on the web. Read [23] describes a system for identifying affect in short fiction stories, using the statistical association level between words in the text and a set of keywords; the experiments show limited success rates, but indicate that the method is better than a naive baseline. We incorporate similar statistical association measures in our experiments as one of the feature sets used (see Section 4). Rubin *et al.* investigated discriminating terms for emotion detection in short texts [24]; the corpus used in this case is a small-scale collection of online reviews. Holzman and Pottenger report high accuracy in classifying emotions in online chat conversations by using the phonemes extracted from a voice-reconstruction of the conversations [9]; however, the corpus they use is small and may by biased. Liu *et al.* present an effective system for affect classification based on large-scale "common-sense" knowledge bases [17].

Two important points differentiating our work from existing work on affect analysis are the domain type and its size. As far as we are aware there is no published work on computational analysis of affect in blogs; this is an interesting and challenging domain due to its rising importance and accessibility in recent years, and the properties that make it different from other domains (e.g., highly personal, subjective writing style and the use of non-content features such as emoticons – see Section 4).

Closely related areas to mood classification are the fields of authorship attribution [19, 15] and gender classification [14], both of which are well-studied. Since these tasks are focused on identifying attributes that do not change over time and across different contexts, useful features typically employed are non-content features (such as the usage of stopwords or pronouns). In contrast, moods are dynamic and can change – for the same author – in a relatively short span. This causes both the features used for mood classification to be more content-based features, and the documents used for classification to be different: while authorship attribution and gender detection work well on long documents such as journal articles and even books, mood classification should be focused on short, time-limited documents.

Finally, a large body of work exists in the field of Sentiment Analysis. This field addresses the problem of identifying the semantic polarity (positive vs. negative orientation) of words and longer texts, and has been addressed both using corpus statistics [31, 6], linguistic tools such as WordNet [11], and "common-sense" knowledge bases [17]. Typically, methods for sentiment analysis produce lists of words with polarity values assigned to each of them. These values can later be aggregated for determining the orientation of longer texts, and have been successfully employed for applications such as product review analysis and opinion mining [3, 30, 21, 20, 2, 4].

3. A BLOG CORPUS

We now describe the collection of blog entries we used for our experiments.

We obtained a corpus of 815494 blog posts from Livejournal,[1] a free weblog service with a large community (several millions of users; considered the largest online blogging community). The web interface used by Livejournal, allowing users to update their blog, includes – in addition to the input fields for the post text and date – an optional field indicating the "current mood." The user can either select a mood from a predefined list of 132 common moods such as "amused", "angry" and so on, or enter free-text. If a mood is chosen while adding a blog entry, the phrase "current mood: X" will appear at the bottom of the entry, where X is the mood chosen by the user.

One obvious drawback of the mood "annotation" in this corpus is that it is not provided in a consistent manner; the blog writers differ greatly from each other, and their definitions of moods differ accordingly. What may seem to one person as a frustrated state of mind might appear to another as a different emotional state – anger, depression, and so on. Of course, this is also an advantage in a way, since unlike other corpora, in this case we have direct access to the writer's opinion about her state of mind at the time of writing (rather than an external annotator).

The blog corpus was obtained as follows. First, for each one of the 132 common moods given by Livejournal as predefined moods, we used the Yahoo API [32] to get a list of 1000 web pages containing a Livejournal blog post with that mood. Since the Livejournal web pages contain multiple blog posts (up to 20), some of the web pages overlapped; in total, our list contained 122624 distinct pages, from 37009 different blogs. We proceeded to download the posts in these pages, getting in total the 815494 posts mentioned above – 22 posts per blog, on average. Of these posts, 624905 (77%) included an indication of the mood; we disregarded all other posts.

As expected, the distribution of different moods within the posts follows a power law. The number of unique moods in the corpus is 54487, but 46558 of them appear only once, and an additional 4279 appear only twice; such moods are inserted by users in the free-text field rather than chosen from the predefined list. Table 3 shows the distribution of the most popular moods in our corpus (percentages are calculated from the total number of posts with moods, rather than from the total number of posts altogether).

[1] http://www.livejournal.com

Mood	Occurrences		Mood	Occurrences		Mood	Occurrences	
amused	24857	(4.0%)	contemplative	10724	(1.7%)	anxious	7052	(1.1%)
tired	20299	(3.2%)	awake	10121	(1.6%)	exhausted	6943	(1.1%)
happy	16471	(2.6%)	calm	10052	(1.6%)	crazy	6433	(1.0%)
cheerful	12979	(2.1%)	bouncy	10040	(1.6%)	depressed	6386	(1.0%)
bored	12757	(2.0%)	chipper	9538	(1.5%)	curious	6330	(1.0%)
accomplished	12200	(1.9%)	annoyed	8277	(1.3%)	drained	6260	(1.0%)
sleepy	11565	(1.8%)	confused	8160	(1.3%)	sad	6128	(1.0%)
content	11180	(1.8%)	busy	7956	(1.3%)	aggravated	5967	(1.0%)
excited	11099	(1.8%)	sick	7848	(1.3%)	ecstatic	5965	(1.0%)

Table 1: Frequently occurring moods in our corpus

To ensure a minimal amount of training data for each mood we attempt to classify, we use only posts for which the mood is one of the top 40 occurring moods in the entire corpus. This leaves us with 345014 posts, the total size of which is 366MB (after cleanup and markup removal). The number of words in the corpus is 69149217 (average of 200 words per post), while the unique number of words is 596638.

An additional point important to note about our corpus is that while it contains a large amount of different authors, it does not constitute a representative sample of adult writers. In fact, many of the blog maintainers are not even adults: according to Livejournal, the median age of blog authors is about 18, so half of the writers are actually teenagers.

4. FEATURE SET

When designing a classification experiment, the most important decision – more important than the choice of the learning algorithm itself – is the selection of features to be used for training the learner. In the case of text classification, several feature sets such as word counts are commonly used; in the blog domain, additional sets of features seem beneficial. In this section we list the features we used in our experiments, grouped by "feature family".

First, we employ "classic" features in text analysis – features which are used in various types of classification tasks, both style-related and topic related.

Frequency Counts

Perhaps the most common set of features used for text classification tasks is information regarding the occurrence of words, or word n-grams, in the text. The absolute majority of text classification systems treat documents as simple "bag-of-words" and use the word counts as features [28]. Other measures commonly used as features in text classifiers are frequencies of Part-of-Speech (POS) tags in the text. In our experiments, we used both the word counts and the POS tag counts as features; an additional feature that we used was the frequencies of word lemmas. Both the POS tags and the lemmas were acquired with TreeTagger [27].

We have experimented both with single word/POS/lemma features and with higher-order n-grams; due to time and space constraints, in this paper we report only on unigram features.

Length-related

Four features are used to represent the length of a blog post: the total length in bytes, the number of words in the post, the average length of a sentence in bytes, and the average number of words in a sentence. A naive method was used for sentence splitting, taking standard punctuation marks as sentence delimiters.

Next, we make use of features that are related to the subjective nature of text in blogs – the fact that they tend to contain a larger amount of personal, opinionated text than other domains.

Semantic Orientation Features

Semantic orientation seems like a particularly useful feature for mood prediction: some moods are clearly "negative" (annoyed, frustrated) and some are clearly "positive" (cheerful, loved); it is anticipated that positive blogs posts will have, on average, a more positive orientation than negative ones.

In our experiments, we use both the total orientation of a blog post and the average word orientation in the blog as features. Since the estimation of word semantic orientation is highly dependent on the method used for calculating it, we use two different sources for the word-level orientation estimation.

The first source is a list of 21885 verbs and nouns, each assigned with either a positive, negative, or neutral orientation. The method used for creating this list is described by Kim and Hovy in [12]. In a nutshell, the method uses the WordNet distances of a word from a small set of manually-classified keywords. For calculating the total and average orientation of a post, we assign a value of +1 to every positive word and -1 to every negative one, summing (or averaging) the words.

The second source we use is a similar list of 1718 adjectives with their corresponding real-numbered polarity values, either positive or negative. This list was constructed using Turney and Littman's method described in [30]; their method is based on measuring the co-occurrence of a word with a small set of manually-classified keywords on the web.

Examples of words with their values in both lists are given in Table 2, illustrating the occasional disagreement between the different sources.

Word	Kim&Hovy	Turney&Littman
pricey	Positive	-4.99
repetitive	Positive	-1.63
teenage	Negative	-1.45
momentary	Negative	+0.01
fair	Positive	+0.02
earnest	Positive	+1.86
unparalleled	Negative	+3.67
fortunate	Positive	+5.72

Table 2: Semantic orientation values of words

Mood PMI-IR

The next set of features we use is based on Pointwise Mutual Information (PMI, [18]). PMI is a measure of the degree of association between two terms, and is defined as

$$\mathrm{PMI}(t_1, t_2) = \log \frac{p(t_1 \& t_2)}{p(t_1)p(t_2)}$$

PMI-IR [29] uses Information Retrieval to estimate the probabilities needed for calculating the PMI using search engine hitcounts from a very large corpus, namely the web. The measure thus becomes

$$PMI\text{-}IR(t_1, t_2) = \log \frac{\mathrm{hitcounts}(t_1 \& t_2)}{\mathrm{hitcounts}(t_1) \cdot \mathrm{hitcounts}(t_2)}$$

When estimating the total PMI of a text with a certain concept, it is common practice to sum the individual PMI values of all words in the text and the concept [30]. Since we are classifying text by mood, our "concepts" are all possible moods, and we would like to measure the association between words used in the blog entry and various moods. Thus, we pre-calculated the PMI-IR of the 2694 most frequently occurring words in the corpus with the top 40 occurring mood (for a total of $2694 \cdot 40 = 107760$ PMI-IR values). For the search engine hitcounts we used the Yahoo API; some example PMI-IR values are given in Table 3 (higher values depict greater association).

Word	Mood	PMI-IR
nap	great	-15.51
hugged	great	-25.61
mirror	great	-40.23
goodnight	sleepy	-22.88
moving	sleepy	-26.58
install	sleepy	-28.87
homework	content	-29.24
homework	annoyed	-26.04
homework	bored	-25.52

Table 3: Example PMI-IR values of ⟨word,mood⟩ pairs

After calculating the PMI-IR values between the frequent words and the frequent moods, we used 80 additional features for each blog post: for each mood of the top 40 moods, we included two features representing the association of the post to that mood: the total PMI and the average PMI. The numerical values of the features are simply the sum of the normalized PMI-IR values of words contained in the post (and included in the list of 2694 most frequent words for which PMI was pre-calculated), and the average of the PMI values. This approach is somewhat similar to the one used in [23].

Finally, we turn to features that are unique to online text such as blogs, as well as email and certain types of web pages.

Emphasized Words

Historically, written online text such as email was unformatted (that is, raw ASCII was used, without layout modifiers such as different font sizes, italic text and so on). This led to alternative methods of text emphasis, including using all-capitalized words ("I think that's a GREAT idea"), and using asterisks or underscores attached to a word on both sides ("This is *not* what I had in mind", "Did you bother _checking_ it before sending??").

While today most online text has extensive formatting options, usage of these emphasis methods is still popular, especially in cases where text is added through a standard text-box on a web page, containing no formatting options – the way many blog hosting services provide access to the blog maintainer.

We use as a feature the frequency of each emphasized word in a post, as well as the total number of stressed words per post. The intuition is that since these are words that the writer chose to emphasize, they may be important indicators of the written text.

Special Symbols

This set of features captures the usage of two types of special characters in the blog posts. The first type is punctuation characters such as ellipsis, exclamation marks, and so forth. The intuition behind modeling the frequencies of these symbols is that in some cases increased usage of them is beneficial for characterizing specific kinds of text [26]. Indeed, punctuation marks proved suitable in some text classification tasks, such as detecting email spam [25]. We use as features the frequencies of 15 common special symbols in each blog post; these include punctuation marks and some additional non-alphanumeric symbols such as asterisks and currency signs.

The second type of special symbols we use as feature are *emoticons* (emotional icons). Emoticons are sequences of printable characters which are intended to represent human emotions or attitudes; often, these are sideways textual representations of facial expressions. Examples of such emoticons are :) (representing a smile) and ;) (representing a wink) – both viewed sideways. Usage of emoticons originated in email messages and quickly spread to other forms of online content; it is currently very popular in many domains including blogs. Similarly to the first set of special symbols, we use the frequencies of 9 popular emoticons in the blog posts as features.

5. EXPERIMENTAL EVALUATION

In this section we describe the experiments we performed for classifying the mood of a blog post. We start with an overview of the classification environment, and follow with a description of the experiments performed and their results.

5.1 Classification

Setup

For our experiments, we use SVMlight, a support-vector machine package.[2] SVMs are popular in text classification tasks since they scale to the large amount of features often incurred in this domain [10]; also, they have been shown to significantly outperform other classifiers for this type of experiments [33].

Although SVMs can effectively handle large feature spaces, for efficiency reasons we chose to reduce the number of features related to word frequencies in the text. This is a common practice in text classification tasks, due to the large feature space; many text classifiers employ methods for Feature Selection – choosing only some of the features to actually be used in the learning process.

The intuition behind our feature selection method is that each mood has a set of words (and similarly, POS tags and lemmas) that are more characteristic of text associated with this mood than of other texts. We identify this set of features for each mood, then aggregate the separate sets to a combined feature set of all words which are characteristic of at least one mood. The identification of the characteristic set of features per mood is done using standard tests for comparing frequency distributions, where we compare the distribution of the words in texts associated with a mood with the distribution of the words in all other texts, and similarly for POS tags and word lemmas.

More formally, for each mood m we define two probability distributions, Θ_m and $\Theta_{\overline{m}}$, to be the distribution of all words in texts associated with m and in other texts, respectively.[3] We rank all the words in Θ_m, according to their log likelihood measure [22], as compared with $\Theta_{\overline{m}}$. We then set as the "set of characteristic features for mood m" the top N-ranked features. Once we have completed this process for all moods, we combine all the characteristic sets obtained to one feature set. In the experiments reported in this paper, we set N to 50.

Examples of characteristic word n-grams in our corpus for some moods are given in Table 4; characteristic POS and lemma features were calculated similarly.

Since the vocabulary of stressed words is substantially smaller than that of all words, we do not employ any mechanisms for reducing the amount of features, as we did with the frequencies of words.

Experiments

We performed two sets of experiments. The first set is intended to evaluate the effectiveness of identifying specific, individual moods in a blog post, and to examine the effect of changes in the training set size on classification accuracy. For each mood we created a training set with randomly drawn instances from the set of posts associated with that mood as positive examples, and an equal amount of negative examples, randomly drawn form all other moods. The test set we used contained, similarly, an equal amount of random positive and negative instances, distinct from those used for training.

For the second set of experiments, we manually partitioned the moods into two "mood sets" according to some abstraction, such as "positive moods" vs. "negative moods". We then repeated the training and testing phase as done for the individual mood classification, treating all moods in the same set as equivalent. The purpose of these experiments was to test whether combining closely-related moods improves performance, since many of the moods in our corpus are near-synonyms (e.g., "tired" and "sleepy").

In the experiments reported in this paper, we use the entire list of features given above, rather than select subsets of it and experiment with them separately. This was done due to space constraints; our ongoing work includes evaluating the performance gain contributed by each feature subset.

For classifying individual moods, our training set size was limited to a maximum of a few thousand positive and a few thousand negative examples, since many moods did not have large amounts of associated blog posts (see Table 3). For classifying the mood sets, we used a larger amount of training material.

Since both our training and test sets contain the same number of positive and negative examples, the baseline to all our experiments is 50% accuracy (achieved by classifying all examples as positive or all examples as negative).

Mood	Top words	Top bigrams	Top trigrams
hungry	hungry eat bread sauce	am hungry hungry and some food to eat	I am hungry is finally happened I am starving ask my mother
frustrated	n't frustrated frustrating do	am done can not problem is to fix	I am done am tired of stab stab stab I do not
loved	love me valentine her	I love love you love is valentines day	I love you my god oh i love him love you so

Table 4: Most discriminating word n-grams for some moods

5.2 Results

Table 5 lists the results of the classification of individual moods. The test sets contained 400 instances; for the training sets we used varying amounts, up to 6400 instances; the table lists the results when training with 1600 instances and with 6400 instances. The results of the classification of two mood partitions – active/passive and positive/negative – are shown in Table 6.

[2]http://svmlight.joachims.org/

[3]We describe the process for word features, but it is equivalent for POS tag and word lemma features.

Mood	Correct		Mood	Correct	
	1600	6400		1600	6400
confused	56.00%	65.75%	bored	51.52%	55.25%
curious	60.25%	63.25%	sleepy	44.25%	55.00%
depressed	58.25%	62.50%	crazy	54.00%	55.00%
happy	54.50%	60.75%	blank	56.00%	54.50%
amused	57.75%	60.75%	cheerful	52.50%	54.25%
sick	54.75%	60.25%	anxious	51.75%	54.25%
sad	53.00%	60.25%	aggravated	52.75%	54.25%
frustrated	57.00%	60.25%	content	50.75%	54.00%
excited	55.50%	59.75%	awake	51.50%	53.75%
ecstatic	54.00%	59.75%	busy	50.75%	53.50%
bouncy	51.00%	59.50%	cold	50.25%	53.25%
thoughtful	52.75%	59.00%	exhausted	52.50%	52.50%
annoyed	57.00%	59.00%	drained	47.50%	52.25%
loved	57.00%	57.75%	hungry	51.50%	50.75%
blah	53.75%	57.75%	good	48.50%	50.50%
hopeful	51.50%	57.50%	creative	47.75%	50.50%
cranky	55.00%	57.25%	okay	46.75%	49.00%
contemplative	53.25%	57.00%	calm	44.75%	49.00%
accomplished	54.75%	55.75%			
400 test instances; 1600 and 6400 training instances					

Table 5: Classification performance: individual moods

Size of training set	Active/Passive	Positive/Negative
800	50.51%	48.03%
1600	50.93%	53.00%
3200	51.50%	51.72%
6400	51.77%	54.92%
20000	53.53%	54.65%
40000	55.26%	57.33%
80000	57.08%	59.67%

Table 6: Classification performance: active vs. passive moods (size of test set: 65936) and positive vs. negative moods (size of test set: 55495)

5.3 Discussion

The classification performance on most moods is modest, with an average of 8% improvement over the 50% baseline (with 6400 training examples); a few moods exhibit substantially higher improvements, up to 15% improvement over the baseline, while a small number of moods are performing equivalently or worse than the baseline. Examining the better and worse performing moods, it seems that the better ones are slightly more concrete and focused than the worse ones, e.g., "depressed", "happy" and "sick" compared to "okay" and "calm". However, this is not consistent as some concrete moods show low accuracy ("hungry") whereas some of the non-focused moods perform averagely ("blah"): the reasons for the different behavior on different moods need to be explored further.

Somewhat surprising, the classification of the aggregated sets does not seem to be an easier task than classifying a single mood, despite the substantial increase in the amount of training examples.

In general, it seems that the classification task is indeed a complex one, and that methods and features used for other stylistic analysis – even when augmented with a range of additional features – do not provide sufficient results. Disappointed by these results, we decided to let humans perform the individual mood classification task, and see if this yields substantially higher performance. For each one of the 40 most frequent moods, we randomly selected 10 posts annotated with that mood, and 10 posts annotated with a random other mood. We then presented these 20 posts to a human assessor without their accompanying moods; the assessor was told that exactly 10 out of the 20 posts are of mood X (the mood was explicitely given), and was asked to select which ones they are. This process simulates the same test data used with the machine learning experiments. The accuracy of the human over these 800 posts was 63%, and the assessor commented that in many of the cases, it seemed to him that much less than 10 posts were in fact related to the given mood, therefore driving him to choose randomly.

Some possible reasons for the low accuracy – both of the human and the machine – on this task are as follows.

- First, the subjective nature of the "annotation" in our corpus is problematic due to the large amount of widely-varying authors in it. Unlike lab-controlled experiments, where annotators follow guidelines and try to be consistent, our annotated corpus is fairly unstable.

- Additionally, the nature of the blog posts is problematic for text classification purposes. The average size of an entry is, as stated earlier, is 200 words; this is usually not enough to gather meaningful statistics, creating very sparse training data. Some posts hardly contain text at all – just a few links or pictures – and others yet are not even in English.

- Finally, the mood categories themselves – as defined by the Livejournal interface – are highly subjective; for any given mood there are lots of different situations that may bring this mood about, and correspondingly there could be many different types of blog entries labelled with the same mood.

One clear observation is the increasing the size of the training set affects favorably the performance in the vast majority of the cases, particularly for single-mood classification, and to a lesser extent also for mood-set classification. We believe this indicates that our results can still improve by simply further increasing the training data size.

6. CONCLUSIONS

We presented preliminary experiments in classifying moods of blog text, using as our corpus a large collection of blog posts containing the authors' indication of their state of mind at the time of writing. We use a variety of features for the classification process, including content and non-content features, and some features which are unique to online text such as blogs. Our results are show a small, if consistent, improvement over a naive baseline; while the success rates are relatively low, human performance on this task is not substantially better.

Going back to our research questions, we witness that mood classification in blogs is a complex task—for humans as well as machines—and that the wealth of features available for the learning process does not ensure high performance. We do experience, however, a consistent improvement over a baseline for almost all given moods. Furthermore, our results indicate that increasing the amount of training data results in a clear improvement in effectiveness, and that our experiments did not reach a saturation point in the improvement – i.e., further improvement is expected with more training data.

In the future, we intend to thoroughly analyze which features are more beneficial for effective classification, and modify our feature set accordingly; preliminary investigations in this direction show that the mood PMIs are prominent features throughout all moods. In this context, we are examining the notion of "feature stability" [13] for identifying important features for style analysis in blogs. Additional directions we intend to explore are the relation between blog post length and the success in classifying it, and the reasons for the different performance of the classification process on different moods. Finally, to increase the level of "annotator agreement"—the consistency level regarding moods among bloggers—we intend to reduce the amount of different authors in the corpus, focusing on a relatively small amount of bloggers, with a large amount of posts each.

Acknowledgments

The author wishes to thank Maarten de Rijke for valuable comments and discussions, and Soo-Min Kim and Ed Hovy for providing their polarity-tagged lists. This work was supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001.

7. REFERENCES

[1] M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings ACL 2001*, pages 26–33, 2001.

[2] S. Das and M. Chen. Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. In *EFA 2001*, 2001.

[3] K. Dave, S. Lawrence, and D. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW2003: the 13th international conference on World Wide Web*, 2003.

[4] G. Grefenstette, Y. Qu, J. Shanahan, and D. Evans. Coupling niche browsers and affect analysis. In *RIAO'2004*, 2004.

[5] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW2004: the 13th international conference on World Wide Web*, pages 491–501, New York, NY, USA, 2004. ACM Press.

[6] V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings EACL 1997*, 1997.

[7] S. Herring, L. Scheidt, S. Bonus, and E. Wright. Bridging the gap: A genre analysis of weblogs. In *HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4*, page 40101.2, Washington, DC, USA, 2004. IEEE Computer Society.

[8] J. Hodgson, C. Shields, and S. Rousseau. Disengaging communication in later-life couples coping with breast cancer. *Families, Systems, & Health*, (21):145–163, 2003.

[9] L. Holzman and W. Pottenger. Classification of emotions in internet chat: An application of machine learning using speech phonemes. Technical Report LU-CSE-03-002, Lehigh University, 2003.

[10] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pages 137–142, London, UK, 1998. Springer-Verlag.

[11] J. Kamps, M. Marx, R. Mokken, and M. de Rijke. Using WordNet to measure semantic orientations of adjectives. In *Proceedings LREC 2004*, 2004.

[12] S.-M. Kim and E. Hovy. Determining the Sentiment of Opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, 2004.

[13] M. Koppel, N. Akiva, and I. Dagan. A corpus-independent feature set for style-based text categorization. In *IJCAI'03 Workshop On Computational Approaches And Synthesis*, 2003.

[14] M. Koppel, S. Argamon, and A. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[15] M. Koppel and J. Schler. Authorship verification as a one-class classification problem. In *ICML '04: Twenty-first international conference on Machine learning*, New York, NY, USA, 2004. ACM Press.

[16] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW2005: the 14th international conference on World Wide Web*, pages 342–351, New York, NY, USA, 2005. ACM Press.

[17] H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge. In *IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132, New York, NY, USA, 2003. ACM Press.

[18] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[19] A. McEnery and M. Oakes. *Handbook of Natural Language Processing*, volume 2, chapter Authorship Studies / Textual Statistics. Marcel Dekker, 2000.

[20] T. Nasukawa and J. Yi. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings K-CAP'03: the international conference on Knowledge capture*, 2003.

[21] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings EMNLP 2002*, 2002.

[22] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *The workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, 2000.

[23] J. Read. Recognising affect in text using pointwise-mutual information. Master's thesis, University of Sussex, 2004.

[24] V. Rubin, J. Stanton, and E. Liddy. Discerning emotions in texts. In *The AAAI Symposium on Exploring Attitude and Affect in Text (AAAI-EAAT)*, 2004.

[25] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian Approach to Filtering Junk E-Mail. In *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, 1998. AAAI Technical Report WS-98-05.

[26] B. Say and V. Akman. Current Approaches to Punctuation in Computational Linguistics. *Computers and the Humanities*, 30(6):457–469, 1996.

[27] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[28] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[29] P. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, London, UK, 2001. Springer-Verlag.

[30] P. Turney and M. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4), 2003.

[31] J. Wiebe. Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*. AAAI Press / The MIT Press, 2000.

[32] Yahoo Development Network, URL: `http://developer.yahoo.net`.

[33] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, New York, NY, USA, 1999. ACM Press.

User-aware page classification in a search engine

Rachel Aires
NILC/ICMC, University of São Paulo
Caixa Postal 16668 13560-970
São Carlos/SP, Brazil
raires@icmc.usp.br

Sandra Aluísio
NILC/ICMC, University of São Paulo
Caixa Postal 16668 13560-970
São Carlos/SP, Brazil
sandra@icmc.usp.br

Diana Santos
Linguateca, SINTEF ICT
Pb 124 Blindern, 0314
Oslo, Norway
diana.santos@sintef.no

ABSTRACT
In this paper we investigate the hypothesis that classification of Web pages according to the general user intentions is feasible and useful. As a preliminary study we look into the use of 46 linguistic features to classify texts according to genres and text types; we then employ the same features to train a classifier that decides which possible user need(s) a Web page may satisfy. We also report on experiments for customizing searching systems with the same set of features to train a classifier that helps users discriminate among their specific needs. Finally, we describe some user input that makes us confident on the utility of the approach.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]

General Terms
Algorithms, Design, Reliability, Experimentation, Human Factors, Languages.

Keywords
Natural language processing, machine learning, information retrieval, text categorization, web search, stylistic features, personalised search.

1. STYLE IN INFORMATION ACCESS
One of the goals of the present workshop is to discuss, among others, the question: Can stylistic information be used profitably e.g. in information access interfaces? We believe our work in text categorization for IR shows that this is definitely the case, although a consensual definition of style is a pre-requisite to agreement on this matter.

For us, stylistic information is that part of linguistic data that is not related to content, but rather with the way the content is conveyed. Content and style are, however, intricately mixed and are related in complex ways. While some style features are individual and often involuntary, others are taught and described in style manuals and professional training. Style is a set of elusive properties that allow scholars to talk about genres, and humans in general to assess what is appropriate or not in specific (con)texts.

Style, as most linguistic concepts, can be approached at least from two angles: as a macro property of full texts (and/or collections of texts), something that can only be predicated of a large collection, or as a micro property that is operational in every minor linguistic choice a speaker or writer makes. The challenge in automatic style categorization is to connect the two, and, using easy to compute features, provide a classification in terms of recognizable kinds.

The kinds of texts we looked into, or rather the classification we wanted to get at, was not of style in itself, but of what kind of user's need a particular Web page was supposed to satisfy. There were two reasons for this somewhat radical move:

1) automatic style categorization is generally motivated by the belief that style is a good indicator of different user needs, as in Kesser et al. (1997:32): "in information retrieval, genre classification could enable users to sort search results according to their immediate interests" [1]. Why go indirectly through genre and not directly to user's interests?
2) there was not a well-established off-the-shelf classification scheme for the Web, with some people claiming there are many new and evolving genres in it. For example, [2] and [3] Web typologies are significantly different.

Also, genre and style are often connected with the text producers, while our view was a consumer's view. Given the well-know mismatch between what users want and what producers offer, it is not obvious that one should start by looking at the collection instead of looking at what the presumable goals of users might be. Inspired by Broder [4] and previous work in detecting user's goals in Web search [5], we devised a user need typology from a qualitative analysis of the TodoBr logs.

Our assumption is that categorizing the results by user need (i.e., which would user need they would satisfy) would improve information access, something related, but not necessarily equivalent to what was reported and tested by Bretan et al [2].

Currently, users are faced with information overflow: the problem is too large quantity, not scarcity of information. Information Retrieval is now advancing towards different ways of organizing information. Several search engine companies have recently announced initiatives to improve search results by allowing users to customize their searches by delimiting the search context, in several ways. We are currently aware of 1) the saving of previous queries and more general user behaviour tracking; 2) the ability for the user to define subject profiles of interest, and 3) the offer of vertical search within the web, as in geographically-aware search, news lookup and search for famous people. All of these developments have, in addition to privacy issues, a particular problem, namely how to maintain profile accuracy across different tasks and over extended periods, something arguably difficult to deal with. In fact, the user's focus and interests can

change, and repeated frequent searches may be too specific, as in tasks from the user's work. Cognitive aspects of search behaviour have been claimed to deserve close attention, especially because multitasking information seeking seems to be common in the Web as well as in other information seeking environments [6].

So, in [7] we took the path to present to the user the results of her search classified by type of goal, i.e., develop a categorization meant to improve the presentation of the search results. Eventually this kind of categorization might also be employed for indexing, but this has not been the focus of our work.

Our work was inspired by Karlgren's [8] studies of systematic stylistic variation in order to characterize the genre of documents and improve Web search. Later on, we performed also similar experiments on automatic genre classification, but using a different genre classification, developed for Brazilian Portuguese in connection with the freely available reference corpus Lácio-Ref [9].

Our primary goal is to find which stylistic features can be used to classify web pages in Portuguese into understandable and useful classes to the web search task, in order to decrease the user effort to retrieve information.

In this paper, we report on several experiments performed to investigate automatic web pages classification into:

- genre and text types;
- seven general kinds of user needs;
- personalized user needs.

The paper is structured as follows: we start by presenting our explorations in genre and text type classification. Then, we present new experiments carried out to classify texts according to seven users' needs. Finally, we present two studies on the use of stylistic features to create customized classification schemes, one of which for English. We end the paper with a discussion of these results and alleys for further work.

2. CLASSIFICATION IN GENRES AND TEXT TYPES

The genre of a text captures its communicative intention and discourse character. In other words, it classifies the community to which the text is addressed and the human activities that make it relevant. Genres can be told apart by the text types (which are defined by a particular text structure, lexicon, syntax, and adequacy to the main theme) usually associated to each of them.

Karlgren, based primarily in Biber [10], used stable characteristics of texts for genre categorization. According to Karlgren [8], style is the difference between two ways of saying the same thing, and systematic stylistic variation can be used to characterize the genre of documents. In one of his studies ([8]: Chapter 16) he looked into the design of an interactive system with the interface incorporating stylistic information, categorizing retrieval results by genre, and displaying the results using this categorization. In this experiment, eleven categories were employed and a user-centred evaluation was performed. The users were asked to execute two tasks each, using the interface prototype with stylistic features and the web search engine Altavista. Karlgren concluded that most users used the interface as intended and many searched for documents in the genres the results could be expected to show up in.

Biber [10] has studied English text variation using several variables, and found that texts vary along five dimensions. Registers would then differ systematically along each of these dimensions, relating to functional considerations such as interactiveness, involvement, purpose, and production circumstances, all of which have marked correlates in linguistic structure.

Stamatatos et al [3] have also worked with genre classification based on stylometric methods, creating a Web corpus for Modern Greek and automatically categorizing it.

We performed the following experiment: We used the genres scheme presented by Aluísio et al [9] on which the Lácio-Ref[1] corpus was based, in connection with the corpus for training. The corpus has 4,278 files with 8,291,818 words, divided into 5 genres (scientific, informative, law, literary, and instructional) and 30 text types (paper, administrative circular, statement, dissertation, editorial, interview, law, textbook, public notice, decree, short stories, letter, monograph, news, legal opinion, report, review, abstract, provisional measure, official letter, ordinance, receipt, news reporting, resolution, government body rules, court management measure, court decision, superior court decision, poem, and other).

To make text classification even more flexible an option is to allow the user to get the search results classified by text types. For example, it is improbable that the same information need can simultaneously be satisfied with a poem about lonely hearts and a recipe using chicken heart as an ingredient. Therefore, we investigated whether classification of Web pages into text types could mirror somehow user's intentions.

We computed the 46 features for each text that had been suggested in [7] (shown in Figure 1), and used them to train a genre classifier. These features, which are mainly closed lists, were inspired by those proposed by Biber [10] and Karlgren [8], but checked in grammars and textbooks for Portuguese.

Word-based statistics
Type/token ratio
capital type token ratio
digit content
average word length in characters
long words (>6 chars) count
Text-based statistics
Character count
average sentence length in characters
sentence count
average sentence length in words
text length in words
Other statistics
the subjective markers "acho", "acredito que", "parece que" and "tenho impressão que" ("I think so", "I believe that", "it seems that", "have the impression that")

[1] http://www.nilc.icmc.usp.br/lacioweb/

the present forms of verb to be "é/são" ("is/are")
the word "que" (can be: noun, pronoun, adverb, preposition, conjunction, interjection, emphatic particle)
the word "se" ("if/whether" and reflexive pronoun)
the discourse markers "agora", "da mesma forma", "de qualquer forma", "de qualquer maneira" and "desse modo" ("now", "on the same way", "anyway", "somehow" and "this way")
the words "aonde", "como", "onde", "por que", "qual", "quando", "que" and "quem" on the beggining of questions (wh-questions)
"e", "ou" and "mas" as sentence-initial conjunctions ("and", "or", "but")
amplifiers. Amplifiers scale upwards (Quirk et al, 1992), denoting either an upper extreme of a scale or a high degree, high point on the scale. Some examples are: "absolutamente" (absolutely), "extremamente" (extremely), "completamente" (completely) and "longe" (far).
conjuncts. Most conjuncts are adverbs and prepositional phrases (Quirk et al, 1992). Some examples are: "além disso" (moreover), "conseqüentemente" (accordingly), "assim" (thus) and "entretanto" (however).
downtoners. Downtoners have a lowering effect on the force of the verb and many of them scale gradable verbs, they can have a slight lowering effect, scale downwards considerably or serve to express an approximation to the force of the verb (while indicating its non-application) (Quirk et al, 1992). Some examples are: "com exceção" (with the exception), "levemente" (slightly), "parcialmente" (partially) and "praticamente" (practically).
emphatics. Emphatics (emphasizers) have a general heightening effect (Quirk et al, 1992). Some examples are: "definitivamente" (definitely), "é óbvio que" (it is obvious that), "francamente" (frankly) and "literalmente" (literally).
suasive verbs. Some examples are the verbs: *aderir* (to adhere), *distinguir* (to distinguish), *crer* (to believe) and *dar* (to give).
private verbs. Some examples are the verbs: *partir* (to leave), *ter* (to have), *averiguar* (to check) and *guardar* (to keep).
public verbs. Some examples are the verbs: *abolir* (to abolish), *promulgar* (to promulgate), *mencionar* (to mention) and *declarar* (to declare).
number of definite articles
number of indefinite articles
first person pronouns
second person pronouns
third person pronouns
number of demonstrative pronouns
indefinite pronouns and pronominal expressions
number of prepositions
place adverbials
time adverbials
number of adverbs
number of interjections
contractions
Causative conjunctions
Final conjunctions
Proportional conjunctions
Temporal conjunctions
Concessive conjunctions
Conditional conjunctions
"conformative" conjunctions
comparative conjunctions
consecutive conjunctions

Figure 1. The 46 features selected

We used the Weka J48, Sequential Minimal Optimization (SMO) and Logistic Model Tree (LMT) algorithms [11]. J48 is the Weka implementation of the decision tree learner C4.5. C4.5 was chosen for several reasons: it is a well-known classification algorithm, it had already been used in similar studies [8], and it produces easily understandable rules. LMT [12] is a classification algorithm for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves. SMO implements Platt's [13] sequential minimal optimisation algorithm for training a support vector classifier using scaled polynomial kernels, transforming the output of SVM into probabilities by applying a standard sigmoid function that is not fitted to the data. The implementation used does not perform speed-up for linear feature space and sparse input data. It globally replaces all missing values, transforms nominal attributes into binary ones, and normalizes all numeric attributes.

The results[2] for precision, recall and F-measure are shown in Table 1.

Table 1. Results for genres and text types

Algorithms	**J48**	**SMO**	**LMT**
Classification in Genres			
Precision	0.82	0.81	0.89
Recall	0.77	0.85	0.85
F-measure	0.79	0.82	0.87
Classification in Text Types			
Precision	0.65	0.55	0.76
Recall	0.67	0.91	0.74
F-measure	0.65	0.69	0.75

Results for genres confirm that stylistic features can also be used in the classification of Portuguese texts, as it was done in studies for English and other languages. The best result was achieved with LMT. The results for text types were poorer, even for the

[2] In all experiments presented in this paper we used 10-fold cross validation.

best algorithm. Reasons for this may include the fact that the corpus is not balanced in terms of text types (there are 300 texts for some types, while for others there are only 6), or that the text types themselves do not really stand apart in linguistic terms.

3. CLASSIFICATION IN SEVEN USERS' NEEDS

As documented in [7], the classification scheme based on the seven users' needs was the outcome of a qualitative analysis of the most common users' needs for the period between November 1999 and July 2002, provided by TodoBr[3] logs — a major Brazilian search engine from Akwan Information Technologies. This classification reflects what the user wants:

1) A definition of something or to learn how or why something happens. For this need, dictionaries, encyclopaedias, textbooks, technical articles, reports and texts of the informative genre would present the best results.

2) To learn how to do something or how something is usually done, as in finding a recipe of cake or learning to make gift boxes and installing Linux. Typical results are texts of the instructional genre, such as manuals, textbooks, readers, recipes and even some technical articles or reports.

3) A comprehensive presentation about a given topic. In this case, the best results should be texts of the instructional, informative and scientific genres, e.g. textbooks, essays and long articles.

4) To read news about a specific subject, as the news about the current situation in a given part of the world, or the latest results of soccer games. The best answers in this case would be texts of the informative genre, e.g. online newspapers and magazines.

5) To find information about someone or a company or organization. A typical example would be the user interested in more information about his/her blind date or to find the contact information of someone he met in a conference. Typical answers here are personal, corporation and institutional web pages.

6) To find a specific web page whose URL the user does not remember. For this type of need the results could be from any type of text or genre. The only way to identify this need would be if the interface asked the user what type of page he/she is looking for.

7) To find URLs where for accessing online services, such as buying clothes or downloading software. The best answer to this kind of request is commercial text types (companies or individuals offering products or services).

In a previous experiment (see [7] for more details) we created a corpus with 511 texts extracted from the Web, 73 for each type of need[4] plus additional 73 texts that would not answer any of the six types used (we call it "others"), in order to have a balanced corpus. The resulting corpus had 640,630 words. For comparison, note that Biber's 481 texts amounted to a corpus with approximately 960,000 words, which is larger in number of words because Web texts tend to be smaller.

For this experiment we used the 46 features shown in Figure 1. We computed these statistics for all texts, and trained classifiers using 2, 3 (2 categories plus "others"), 4, 5 (4 categories plus "others"), 6 and 7 categories (6 categories plus "others") (Table 2). In [14] we used mainly the J48 algorithm.

Table 2. Used categories

2 categories	4 categories	6 categories
1) the union of needs 1, 2, 3, 4 and 5 2) need 7	1) the union of needs 1, 2, 3 2) need 4 3) need 5 4) need 7	Need 1 Need 2 Need 3 Need 4 Need 5 Need 7

The classification with 2 categories decides whether a page gives any kind of information about a topic or gives access to an online service. The classification with 4 categories distinguishes among information about something, someone or some company/institution/organization, news, and online services. Finally, the classification with 6 categories is the most comprehensive presented here, which excludes only category 6 that can be of any type of text or genre. The class "others" contains text types like blogs, jokes, poetry, etc, that are examples of text types not covered by the seven users' needs.

The results were encouraging: We got 90.93% of correct classification for 2 categories, 76.97% for 3, 65.06% for 4; 56.56% for 5; 52.01% for 6 and 45.32% for 7 categories. We then replicated these experiments using all 44 Weka algorithms which could deal with non-nominal features, with non-numerical classes, with the number of classes we needed (maximum 7) and which did not present errors related to the standard deviation of our features for any of our classes. Fourteen algorithms achieved the same or better results than J48, regarding the percentage of correct decisions. The best ones were LMT and SMO. The best result for 2, 3, 4, 5, 6 and 7 categories were, respectively 93.83%, 82.97%, 73.74; 67.90%, 63.69% and 58.31% (see [14] for a full description of the results, such as precision and recall per class).

In spite of these good results, there were problems in this approach, particularly the assumption that any given text could only satisfy one user's need. So we created a new and larger corpus, "Yes, user!",[5] which was reclassified in as many of 22 classes (see [15] for corpus description), some with only a few texts. In order to have a balanced corpus, it was enlarged to 1,703 texts (2,159,491 words).

We carried out 3 experiments using the reclassified corpus with: (i) the 46 features of Figure 1; (ii) the 46 features from (i) plus 5 functions to measure vocabulary richness, taken from [3] and shown in Figure 2, resulting in 51 features; (iii) the features from (i) plus features dealing with the most frequent words in the corpus, after eliminating stop-words, linking verbs, adverbs, domain related words (terminology) and further grouping some words together (108 features).

[3] www.todobr.com.br

[4] Except for type 6, which, as explained above, can correspond to any kind of text.

[5] http://www.linguateca.pt/Repositorio/YesUser/

In Figure 2 V_i is the number of words used exactly i times and α is fixed as 0.17.

$$K = \frac{10^4 (\sum_{i=1}^{m} i^2 V_i - N)}{N^2} \qquad W = N^{V^{-\alpha}}$$

$$S = \frac{V_2}{V}$$

$$R = \frac{(100 \log N)}{(1 - (\frac{V_1}{V}))} \qquad D = \sum_{i=1}^{V} V_i \frac{i(i-1)}{N(N-1)}$$

Figure 2. functions to measure vocabulary richness

In the first experiment we generated 3 classification schemes[6]: one with all the six needs, another distinguishing among pages which offer services, pages which offer information and pages which offer both, and the last one which distinguishes between services and information. The results are shown in Table 3.

Table 3. Correct classifications using the 46 features

	J48	SMO
Full classification in 6 needs plus "others"	69.7%	72.52%
Information x service x information and service plus "others"	72.17%	73.58%
Information x service plus "others"	85.11%	86.37%

The second and the third experiments were done only with the full classification scheme (six needs plus "others") and the results are shown in Table 4.

Table 4. Correct classifications using 51 or 108 features

	J48	SMO
51 features	70.38%	73.77%
108 features	73.17%	77.02%

The results from table 3 and 4 are significantly better than those in [10] which presented a precision of 45.32% for the classification in six categories plus "others" and 82.97% for the classification in two categories plus "others". For 6 categories plus "others" the best result was with 108 features and SMO; for 2 categories plus "others" the best result was with SMO.

4. CREATION OF CUSTOMIZED CLASSIFICATION SCHEMES

Obviously, the seven types of user needs explained in Section 3 do not cover all kinds of user intentions, as users may do all kinds of unpredictable searches and it is unlikely that one can recover their intentions by looking only at the logs. However, the very features used to generate rules and classify texts can be used to build customized schemes for other tasks. For example, a doctor can create a classification scheme to distinguish between web pages with technical articles about a disease and web pages that deal with the subject without scientific rigor. However, it is not possible to use the same features we have studied to distinguish among subjects, for example, to tell cardiology technical texts apart from other medical technical texts. We plan to offer customized schemes to the user in a desktop web search prototype being developed, where the user can select examples of text types that often make his/her searches difficult. In the doctor's example, he/she would give to the system samples of technical and non-technical material that would be used as training material. The system would then automatically calculate the features for the given text set, train a classifier and present an estimation of the system efficacy to the user personal scheme. The generated classification model would be saved as a new option of the classification task. Summing up, we would offer predefined options (genres, text types and seven user's needs) as it is provided by search engines shortcuts and tabs, but we will also allow the user to create his/her own shortcut specific to the binary text type related problematic tasks that he/she often performs.

In the following sections we show three case studies regarding the use of stylistic features to create customized classification schemes.

4.1 Legal texts

We created a corpus with 200 texts in the law domain, extracted from the Web. Half of them are meant for experts, the other half for laymen. In order to find out how many texts are necessary for training personalized schemes, in this experiment we have used: (i) an increasing number of texts in the training sets; (ii) the algorithms J48, SMO and LMT; (iii) the 46 features from Figure 1. Results of each classifier appear in Table 5.

Table 5. Results for legal texts in Portuguese

	J48	SMO	LMT
20 texts			
Precision	0.43	0.61	0.42
Recall	0.60	0.79	0.56
F-measure	0.48	0.67	0.47
100 texts			
Precision	0.67	0.78	0.81
Recall	0.68	0.75	0.75
F-measure	0.66	0.75	0.76
200 texts			
Precision	0.77	0.83	0.84
Recall	0.76	0.84	0.84
F-measure	0.76	0.83	0.83

The best results were achieved with a training set with 200 texts and the algorithms SMO and LMT.

We have also trained a classification scheme for texts in English using: (i) a corpus with 200 texts extracted from www.findlaw.com; (ii) the algorithms J48, SMO and LMT and (iii) 52 features taken from Biber and Karlgren [1, 2] which are the original features for English that were adapted for Portuguese (Figure 1) plus 3 types of modals, 2 of negation, nominalizations, besides reflexive and possessive pronouns. Results of each

[6] In all three schemes the class "others" was considered.

classifier appear in Table 6. Figure 3 shows the J48 decision tree when trained with the 200 texts.

Table 6. Results for legal texts in English

	J48	SMO	LMT
20 texts			
Precision	0.82	0.78	0.77
Recall	0.88	0.78	0.80
F-measure	0.84	0.77	0.78
100 texts			
Precision	0.87	0.95	0.91
Recall	0.86	0.91	0.86
F-measure	0.85	0.92	0.87
200 texts			
Precision	0.89	0.96	0.94
Recall	0.87	0.92	0.92
F-measure	0.87	0.94	0.93

The best results were achieved with a training set with 200 texts and the algorithm SMO.

```
second person pronoun <= 0.062305
|  capital type token ratio <= 106
|  |  predictive modals <= 0.295683: laymen (4.0)
|  |  predictive modals > 0.295683: expert (2.0)
|  capital type token ratio > 106
|  |  second person pronoun <= 0.022763: expert (82.0)
|  |  second person pronoun > 0.022763
|  |  |  definite article <= 3.540519: laymen(4.0/1.0)
|  |  |  definite article > 3.540519: expert (7.0)
second person pronoun > 0.062305
|  interjections <= 0.006979
|  |  prepositions <= 8.235294: laymen(92.0/1.0)
|  |  prepositions > 8.235294
|  |  |  second person pronoun <= 0.160128: expert (17.0/1.0)
|  |  |  second person pronoun > 0.160128
|  |  |  |  prepositions <= 11.081323
|  |  |  |  |  definite article <= 5.350978
|  |  |  |  |  |  synthetic negation <= 0.26178: laymen(48.0/4.0)
|  |  |  |  |  |  synthetic negation > 0.26178: expert (3.0/1.0)
|  |  |  |  |  definite article > 5.350978
|  |  |  |  |  |  type token ratio <= 0.357788: expert (8.0)
|  |  |  |  |  |  type token ratio > 0.357788: laymen(2.0)
|  |  |  |  prepositions > 11.081323
|  |  |  |  |  first person pronouns <= 0.151976: expert (9.0)
|  |  |  |  |  first person pronouns > 0.151976: laymen(2.0)
|  interjections > 0.006979
|  |  prepositions <= 4.71464: laymen(2.0)
|  |  prepositions > 4.71464: expert (17.0)
```

Figure 3. Decision tree for English texts classification scheme

4.2 Finding product descriptions

The second study concerned finding out whether E-commerce pages described products on sale or not. We used a collection provided by Martins & Moreira [16] containing 1,252 pages.

Table 7. Results for e-commerce pages

	J48	SMO	LMT
Precision	0.87	0.90	0.90
Recall	0.85	0.69	0.86
F-measure	0.86	0.78	0.88

The best results were achieved with LMT.

5. EVALUATING THE SCHEMES

In order to have some feedback from potential users, we applied a questionnaire to undergraduate students of computer science, linguistics, medicine and to graduate photography students. The goals were to find out:

- How clear to the users was the seven user's needs scheme
- How clear was the genre classification scheme. This was done in 2 ways: (i) asking if any of the three genre schemes presented in [9, 8:16, 3] was helpful for the search task; (ii) presenting the genre scheme from [9] through text type examples and calling it text types schemes (we did not present it as 30 classes mentioned in Section 2, we presented it in 9 classes)
- Which schemes were easier to use
- Whether the user would spend one day collecting text samples to generate a system that would be specific for the tasks that often trouble him

Sixty three students answered the questionnaire. At least two students believed that one of the schemes above was not helpful, specifically: 2 for the seven user's needs, 3 for the text types, 8 for the genre scheme presented in [8:16], 12 for the genre scheme presented in [9] and 13 for the genre scheme presented in [3]. At least six students believed that one of the schemes was easier to use: 25 for the seven user's needs, 29 for the text types, 15 for the genre scheme presented in [8:16], 13 for the genre scheme presented in [9] and 6 for the genre scheme presented in [3].

The hypothesis behind our work was that it is easier for a user to choose among types of needs than between genres. From the questionnaire we realized that the students did not completely understand the genres labels, since the difference between the genres scheme and the text types scheme was only the label and only 3 considered it not useful while 12 considered the genres scheme one not useful. As an example for the labelling, the label for the instructional genre was changed to "text book, culinary recipe, course notes, etc."

The number of students which considered the scheme based on the seven user's needs useful was also larger than those that preferred the genre scheme. However this has to be confirmed using a user-centred evaluation of our prototype. Figure 4 shows a screen dump of our desktop meta searcher prototype, named Leva-e-Traz ("takes and brings").

The results seem to indicate that there is something to be gained classifying Web pages using the a priori schemes: seven users' needs, genres scheme and text type scheme (the one presented in Section 2). All 41 users who had reported having frequent problems in their searches answered that they would spend a day

creating personalised schemes, which apparently confirms the feasibility of the option described in Section 4.

Figure 4. Leva-e-Traz main screen

6. DISCUSSION AND ONGOING WORK

In this paper we have presented results for genre, text types, seven user needs and personalized classification of texts on the Web.

On the one hand, we confirmed that the use of stylistic features to classify texts in genre and text types, as advocated and used for other languages, also works for Portuguese.

In addition, we believe that our attempt to automatically categorise, in terms of user needs, texts on the Web – first reported in [7] – had not been tried before, for any language. We applied it to Brazilian Portuguese, and the experiments reported here improved precision from 0.45 reported in [7] to 0.77, which seems to indicate that this 7 types can be reliably enough identified to help the user.

Finally, we have also obtained some first results for personalized classification, achieving a precision of at least 0.84. As far as we know, the decision of how to classify and rank the texts has not been previously put in the user's hands, although adaptive systems learning from user choices exist in the literature [17]. We are currently conducting more experiments like the ones presented in Section 4 to find out how many texts the user has to collect, and how the selection can be improved. If we confirm, with future experiments, that a small number of texts, such as 200, is sufficient to achieve good results, we may have found a cost-effective way to solve a user's specific text type related problem, as well as sharpened our knowledge of which relevant features to add.

We are also currently investigating the addition of structural clues (such as those in HTML) and of more deep linguistic features (such as those provided by syntactically parsing the text) to our classifiers, and hope to report on these experiments soon [18].

7. ACKNOWLEDGEMENTS

Our thanks to Akwan Information Technologies for the TodoBr logs. This work was partially supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI.

8. REFERENCES

[1] Kessler, B.; Nunber, G.; Schütze, H. Automatic Detection of Text Genre, in Proceedings of the 35th annual meeting on Association for Computational Linguistics (Morristown, NJ, USA, 1997), ACL, 32-38.

[2] Bretan, I.; Dewe, J.; Hallberg, A.; Wolkert, N.; Karlgren, J.: Web-Specific Genre Visualization, *WebNet '98* (Orlando, Florida, November 1998).

[3] Stamatatos, E.; Kakotakis, N.; Kokkinakis, G. Automatic text categorization in terms of genre and author. Computational linguistics (2001), vol 26, number 4, 271-295.

[4] Broder, A. "A Taxonomy of Web Search", SIGIR Forum 36 (2), Fall 2002, p.3-10.

[5] Aires, R.; Aluísio S.: Como incrementar a qualidade das máquinas de busca: da análise de logs à interação em Português. *Revista Ciência da Informação*, **32** (1), 2003, pp. 5-16.

[6] Spink, A.; Ozmutlu , H. C.; Ozmutlu, S.: Multitasking information seeking and searching processes. *Journal of the American Society for Information Science and Technology* **53** (8), 2002, pp. 639-652.

[7] Aires, R.; Manfrin, A.; Aluísio, S.; Santos, D. (2004) What is my Style? Using Stylistic Features of Portuguese Web Texts to classify Web pages according to Users' Needs. In LREC 2004. Lisbon - Portugal, May 2004, p. 1943-1946.

[8] Karlgren, J.: Stylistic Experiments for Information Retrieval. PhD Thesis, Stockholm University, Department of linguistics, 2000.

[9] Aluisio, S.; Pinheiro, G.; Finger, M.; Nunes, M.G.V.; Tagnin, S.E. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In Proceedings of Corpus Linguistics (2003), Lancaster, UK. v. 16, p. 14-21.

[10] Biber, D.: Variation across speech and writing. Cambridge University Press. Cambridge, UK (1988)

[11] Witten, I. H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. San Francisco: Morgan Kaufmann, 2000.

[12] Landwehr, N, Hall, M, Frank, E. (2003) Logistic Model Trees. ECML 2003, p. 241-252.

[13] Platt, J.: Fast training of support vector machines using sequential minimal optimization. In Advances in kernel methods: support vector learning. B. Schölkopf, C. Burges, and A. Smola, eds. MIT Press, 1999, p. 185-208.

[14] Aires, R.; Manfrin, A.; Aluísio, S.; Santos, D.: Which classification algorithm works best with stylistic features of Portuguese in order to classify web texts according to users' needs? Relatório técnico nº 241, outubro de 2004, ICMC/USP.

[15] Aires, R.; Aluísio, S.; Santos, D.: "Yes, user!": compiling a corpus according to what the user wants. Proceedings of Corpus Linguistics 2005 (Birmingham, UK, July 14-17 2005).

[16] Martins Junior, J.; Moreira, E. S. Using Support Vector Machines to Recognize Products in E-commerce Pages. In Proceedings of The IASTED International Conference, February 2004, p. 212-217.

[17] Ciravegna, F.; Wilks, Y.: Designing Adaptive Information Extraction for the Semantic Web in Amilcare, in S. Handschuh and S. Staab (eds), *Annotation for the Semantic Web*, Amsterdam: IOS Press, 2003.

[18] Aires, R.: O uso de características lingüísticas para a apresentação dos resultados de busca na Web de acordo com a intenção da busca do usuário – uma instanciação para o português. PhD Dissertation, Computer Science Department (ICMC), University of São Paulo, forthcoming.