ISSN 2613-0073 Vol 276
http://ceur-ws.org/V276
CEUR-WS.org

SIGIR 2007 Workshop

Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection

held in conjunction with the 30th Annual International ACM SIGIR Conference

27 July 2007, Amsterdam

Organizers

Benno Stein *Bauhaus University Weimar, Germany*
Moshe Koppel *Bar-Ilan University, Israel*
Efstathios Stamatatos *University of the Aegean, Greece*

Preface

This workshop brings together experts and prospective researchers around the exciting and future-oriented topic of plagiarism analysis, authorship identification, and high similarity search. This topic receives increasing attention, which results, among others, from the fact that information about nearly any subject can be found on the World Wide Web. At first sight, plagiarism, authorship, and near-duplicates may pose very different challenges; however, they are closely related in several technical respects:

- Plagiarism analysis is a collective term for computer-based methods to identify a plagiarism offense. In connection with text documents we distinguish between corpus-based and intrinsic analysis: the former compares suspicious documents against a set of potential original documents, the latter identifies potentially plagiarized passages by analyzing the suspicious document with respect to changes in writing style.
- Authorship identification divides into so-called attribution and verification problems. In the authorship attribution problem, one is given examples of the writing of a number of authors and is asked to determine which of them authored given anonymous texts. In the authorship verification problem, one is given examples of the writing of a single author and is asked to determine if given texts were or were not written by this author. Authorship verification and intrinsic plagiarism analysis represent two sides of the same coin.
- Near-duplicate detection is mainly a problem of the World Wide Web: duplicate Web pages increase the index storage space of search engines, slow down result serving, and decrease the retrieval precision. Near-duplicate detection relates directly to plagiarism analysis: at the document level, near-duplicate detection and plagiarism analysis represent also two sides of the same coin. For a plagiarism analysis at the paragraph level, the same specialized document models (e.g. shingling, fingerprinting, hashing) can be applied, where a key problem is the selection of useful chunks from a document.

The development of new solutions for the outlined problems may benefit from the combination of existing technologies, and in this sense the workshop provides a platform that spans different views and approaches.

Benno Stein
Moshe Koppel
Efstathios Stamatatos

Program Committee

Shlomo Argamon, Illinois Institute of Technology
Yaniv Bernstein, Google Switzerland
Dennis Fetterly, Microsoft Research
Graeme Hirst, University of Toronto
Timothy Hoad, Microsoft
Heiko Holzheuer, Lycos Europe
Jussi Karlgren, Swedish Institute of Computer Science
Hans Kleine Büning, University of Paderborn
Moshe Koppel, Bar-Ilan University, Israel
Hermann Maurer, University of Technology Graz
Sven Meyer zu Eissen, Bauhaus University Weimar
Efstathios Stamatatos, University of the Aegean
Benno Stein, Bauhaus University Weimar
Özlem Uzuner, State University of New York
Debora Weber-Wulff, University of Applied Sciences Berlin
Justin Zobel, RMIT University

Content

The Class Imbalance Problem in Author Identification

Efstathios Stamatatos
University of the Aegean
stamatatos@aegean.gr

Abstract

Author identification can be seen as a single-label multi-class text categorization problem. Very often, there are extremely few training texts at least for some of the candidate authors or there is a significant variation in the text-length among the available training texts of the candidate authors. Moreover, in this task usually there is no similarity between the distribution of training and test texts over the classes, that is, a basic assumption of inductive learning does not apply. Previous work [3] provided solutions to this problem for *instance-based* author identification approaches (i.e., each training text is considered a separate training instance). This work [4] deals with the class imbalance problem in *profile-based* author identification approaches (i.e., a profile is extracted from all the training texts per author). In particular, a variation of the Common N-Grams (CNG) method, a language-independent profile-based approach [2] with good results in many author identification experiments so far [1], is presented based on new distance measures that are quite stable for large profile length values. Special emphasis is given to the degree upon which the effectiveness of the method is affected by the available training text samples per author. Experiments based on text samples on the same topic from the Reuters Corpus Volume 1 are presented using both balanced and imbalanced training corpora. The results show that CNG with the proposed distance measures is more accurate when only limited training text samples are available, at least for some of the candidate authors, a realistic condition in author identification problems.

References

[1] Juola, P. “Ad-hoc Authorship Attribution Competition”. *Proc. of ALLC/ACH Joint Conf.*, pp. 175-176, 2004.

[2] Keselj, V., F. Peng, N. Cercone, and C. Thomas, “N-gram-based Author Profiles for Authorship Attribution”. *Proc. of the Conf. of Pacific Association for Computational Linguistics*, 2003.

[3] Stamatatos, E., “Text Sampling and Re-sampling for Imbalanced Author Identification Cases”, In *Proc. of the 17th European Conference on Artificial Intelligence (ECAI'06)*, 2006.

[4] Stamatatos, E. “Author Identification Using Imbalanced and Limited Training Texts”, In *Proc. of the 4th International Workshop on Text-based Information Retrieval*, 2007.

Serial Sharers: Detecting Split Identities of Web Authors

Einat Amitay, Sivan Yogev, Elad Yom-Tov
IBM Research, Haifa, Israel
{einat;sivany;yomtov}@il.ibm.com

ABSTRACT

There are currently hundreds of millions of people contributing content to the Web. They do so by rating items, sharing links, photos, music and video, creating their own webpage or writing them for friends, family, or employer, socializing in social networking sites, and blogging their daily life and thoughts. Of those who author Web content there is a group of people who contribute to more than a single Web entity, be it on a different host, on a different application or under a different username. We name this group *Serial Sharers*. In this paper we analyze patterns in the contributions of Serial Sharers. We examine the overlap between their individual contributions and propose a method for detecting their pages in large and diverse collections of pages.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering, Information filtering.

General Terms

Algorithms, Measurement, Experimentation, Human Factors, Standardization.

Keywords

Web authorship, profiling Web authors, publicly shared spaces.

1. INTRODUCTION

The idea for this paper stemmed from reading an interesting visualization paper about authorship in Wikipedia [12] in which the authors, Holloway et al., describe the contribution patterns of the top 10 most zealous Wikipedians. The thought that such productive contributors can actually change or influence a domain like "law" or "science" to an extent that they dictate the structure of the whole domain was intriguing.

Taking this thought even further, how many people dedicate their writing on the Web to advocate "open source" and what is their influence on current trends by merely expressing their stand in online forums, in blogs, and in virtual communities like Wikipedia? For example, Figure 1 demonstrates that there are nearly 1000 single authors who contributed over 1000 edits (contribution to a single Wikipedia entry in a given time) to the English portion of Wikipedia. Some people annotated the collection with over 100,000 text edits. This small group of people who contribute so much content to a single collection like Wikipedia may create either intentionally or maliciously a distortion in the way information is interpreted.

ACM SIGIR 2007 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection

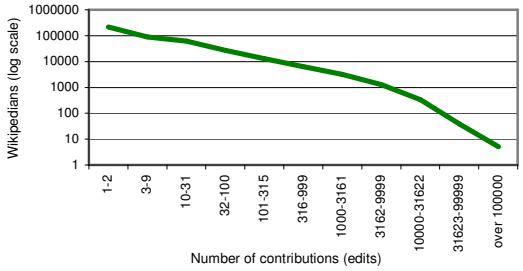

Figure 1 – a histogram of the number of contributions per single author to the English portion of Wikipedia, until September 2006.

Another anecdotal example is the size of the entry for each country in the English Wikipedia plotted alongside the population size of the country, as shown in Figure 2. The trend line traces the decrease in entry size in kb with the decrease in country size. Assuming that there are certain facts that should be common to the description of all countries, like size, population, government, etc., this decrease may be explained by the fact that there are many more social and cultural aspects to describe, but it may also be explained by the number of authors who contribute to each entry. This assumption is supported by the nearly equal size of the entries in the CIA Factbook online (around 100 kb for each country). This authorship "voting" system is a democracy in which the one with the loudest voice wins. Being loud on the Web simply means producing a lot of content on many different pages.

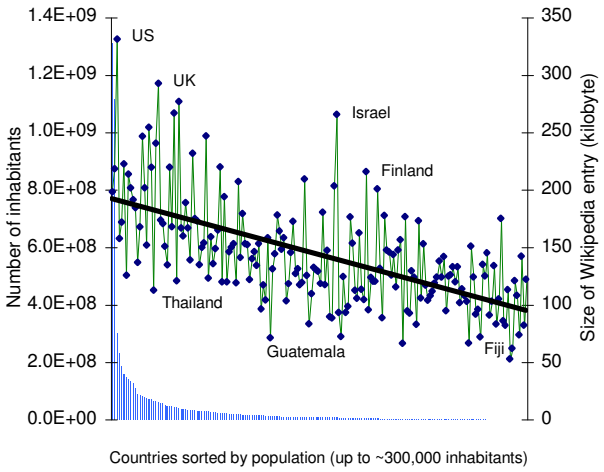

Figure 2 – a comparison of country population size and Wikipedia entry size in kilobytes.

According to recently published studies, about 35% of American Web users contribute some form of content to the Web [12].

Similarly, 31% of urban Chinese Web users create or update Web pages [18], while 20% of British users publish content on the Web [11]. Although the ratio between those who contribute content to those who do not contribute seems quite low, the real numbers translate to tens of millions of Web authors who constantly create and publish new content.

Table 1 lists the various forms in which people contribute content to the Web. They rate products, share links, photos, music and video, create their own webpage or write them for friends, family, or employer, socialize in social networking sites, and blog their daily life and thoughts. The younger the users the more zealous this diverse activity becomes. A recent study shows that 61% of 13 to 17 year-olds in the US have a personal profile on sites such as MySpace, Friendster, or Xanga. Half have posted pictures of themselves online and 37% of those teens maintain a blog [9].

Table 1 – Web authorship – what % of Web users actually contribute content and essentially writes the Web. Source: Pew Internet & American Life Project Surveys [19].

% of Web users who have done this	Activity	Survey Date
35%	Posted content to the internet	December 2005
30%	Rated a product, service or person using an online rating system	September 2005
14%	Created or worked on own webpage	December 2005
11%	Used online social or professional networking sites like Friendster or LinkedIn	September 2005
8%	Created or worked on own online journal or blog	February-April 2006

Among those who hide behind the numbers in Table 1 there are people who produce several types of content. Good examples for these are university professors and students who maintain their own personal Web page on a different host and also a page on their faculty site. This paper is about those authors who shout the loudest on the Web. They not only contribute content to the Web but do so on several different hosts and in various different forms, be it by tagging public material, through their homepage, by blogging, by contributing portions to Wikipedia, and the likes. These authors are not spammers in the trivial sense. Most have no intention of manipulating search results, or influencing world-wide information. They simply enjoy utilizing everything the virtual world offers. We call them *Serial Sharers*.

1.1 Serial Sharers

In a recently published study [21] it was found that 37% of American bloggers had a personal website before they started blogging and that 43% of all bloggers maintain at least two blogs. The actual numbers show that several millions of people in the US alone have authored more than a single page of content and published it online. The portions of content produced by such prolific authors may be considered as a distribution of their online identity. Overall if we took the sum of all the content contributed by a single author we may better describe the interests and thus better profile such a user. The example shown in Figure 3 is a real collection of eight different pages authored by the same person. The pages have some features in common such as the name of the author, some links, some images, some sentences or words, but the overall layout is different, the amount of information provided varies from page to page, the purpose and audience of the pages are different, and so are the hosts where those pages reside.

1.2 Possible Applications

Knowing that the same person authored a collection of not trivially-related pages may be used to enhance and create new applications where knowledge about users is essential. Analyzing and using information about a single author which is extracted from different sources may add new dimensions to user information, such that is not easily available today.

1.2.1 User profiling

Analyzing the identified set of pages written by the same author may help in tailoring user profiles for personalization or for expertise location. Such user profiles may be derived from information the author chose to include in some or all of the pages. For ***personalization*** the profile may be modeled according to the choice of publication media and the information presented in each media; by the shared structure of the documents; by color choice; by syntactic and lexical choice; by layout decisions, by the inclusion of images, etc.

Such information may be used to create user-driven defaults of color and layout choices tailored for each individual user. It may also be used to display advertisements that match the profile of who the user's readership is across all sites, which is the readership most likely to visit the documents in the set. Looking at profiling the audience of a whole site, such collections of authorship-driven profiles spread over several media types and may help to better understand use patterns. For example, what information people choose to share in blogs versus what information they choose to publish on their homepages. It may also help determine the influence of individuals on a collection, to better track a community and those who shape its shared content.

For ***expertise location*** profiling the whole set may reveal and strengthen evidence for knowledge repeating itself in several documents. Also, by using link analysis techniques it may be possible to better reflect the interest the author attracts by looking at all the incoming links to the whole set of documents rather than to a single document. Analyzing social networks based on the whole set of pages written by the same author reveal different patterns than those networks found in homogenous collections consisting only of blogs or of online forum messages. Such information may serve businesses like recruiting services, online dating services, people search indices, and so on.

1.2.2 Noise reduction

Serial sharers may also affect search engine ranking since a single author may produce the same idea in identical or similar forms on some or all of the published pages. This may introduce quite considerable noise to the index of a search engine that relies on any kind of host counting, link counting, term counting or even clickthrough counting.

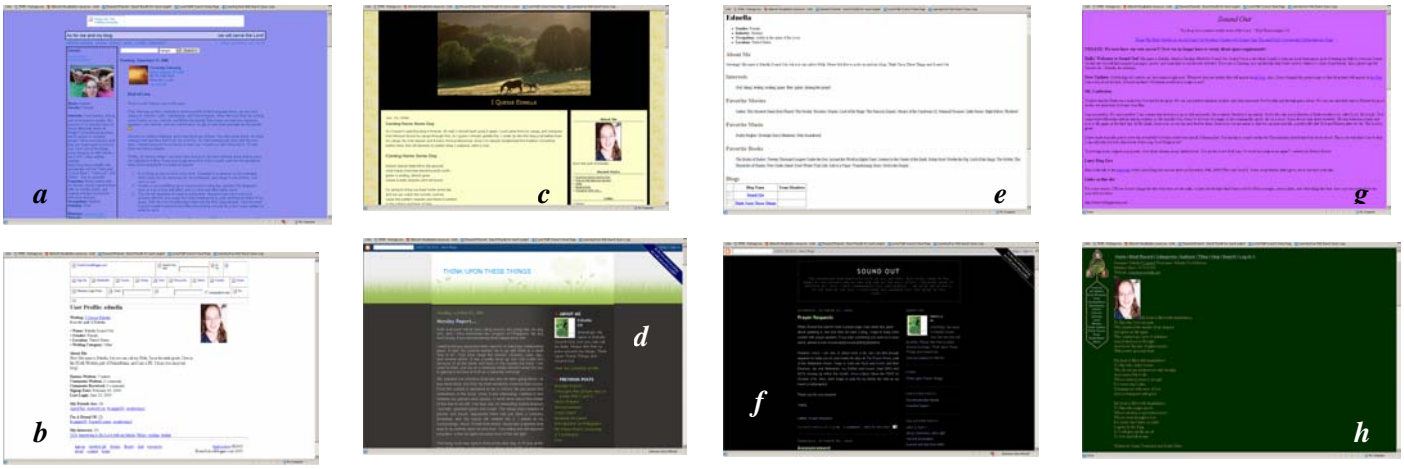

Figure 3 – Eight pages written by the same author and hosted on different sites (*a, c, d, f* blogs; *b, e, h* profiles; *g* unknown type)

On narrow scale or esoteric topics the phenomenon may even influence content per subject. So, assume that there is a band with only 50 content references created by online fans. One specific fan has authored several of them, describing a specific favorite song on two blogs, a homepage, a social networking profile and also on the same fan's YouTube page along with the appropriate link to the mp3 file of that song. Thus, a tenth of the content about the band was produced by a single author. Even if all the other fans disagree with the author on which is the favorite song, the prolific author's voice is loud enough to make a difference. The content contributed is definitely not spam and should not be considered spam.

Serial sharers do not produce spam. They simply use the media in the way it was intended to be used. As demonstrated earlier, today's youth have a higher percentage of users contributing blogs and general content to the Web's collection. When those teens grow up, being a serial sharers will most probably be the norm. This will eventually lead to the Web being a collection of many voices associated with many echoes. The echoes introduce noise into search engine indices. The noise may skew results retrieved for certain topics like "open source" where few people write a lot of content distributed on different hosts.

There are some solutions that come to mind for using author detection to reduce noise in search engine indices. The first is similar to the idea of site collapse where results coming from the same author may be displayed in a cluster or appear after a "show more results by this author" button is pressed.

Another option, which is harder to implement, is to reduce the set to a single file, sort of a summary file that will represent the whole set written by the same author as a single entity in the collection. Creating a single file or a connected set of files may also help aggregate clickthrough data received for a set of same-author pages to better reflect the interest in the whole set rather than in portions of it.

1.2.3 Sizing Web site's unique user community

A different usage for collecting the whole set of pages written by the same author is size estimation of user communities publishing on Blogger, YouTube or Facebook. This will allow for more realistic calculation of the number of unique users who contribute content to the site compared to a different site. Such a comparison may provide stronger evidence about the adoption of certain applications and the rejection of others. For example, if a smaller hosting site is able to prove that its audience consists solely of artists who usually do not publish in any other space this makes the site unique and marketable for advertisement to art supplies companies. On the other hand, a site that has most of its authors publish similar content elsewhere has less value in terms of uniqueness and targeted marketing offerings.

Owners of Web sites may be able to produce a seed of documents labeled with their respective authors taken from the collection and compare those samples with those of other sites. This will help create a benchmark against which user community sizing may be performed.

2. RELATED WORK

In a search system, the problem of author detection resembles, in a sense, the problems of Duplicate Page Detection [6] and Mirror Site Detection [5], both of which use multi-dimensional aspects of the page to describe duplication in features such as size, structure, content, similar naming of the URL, etc. Duplication and mirroring are artifacts of hosting similar information on different machines or hosts in order to facilitate access to those pages in a desired context (e.g. hosting a mirror of a software library on a public university server). Author Detection is somewhat similar in the sense that information written by the same author, such as a user profile or a homepage, is sometimes partially duplicated by mentioning similar topics, expressing similar opinions, repeating the same links or usernames, etc.

However, sometimes each page written by the same author comprises of exclusively unique segments. In the collection we describe in section 4.1.1 there are authors who make a clear distinction between pages about their hobbies such as mountain biking, and their professional pages where they write about academic research or their family.

Many studies explore the field of author detection or author attribution in restricted domains. For instance, Argamon et al. [2], Li et al. [16] and Zheng et al. [25] employ machine learning and shallow parsing methods to detect authors in various collections of newsgroups. Using similar methods, Novak et al. [20] cluster

short messages on online message boards for detecting users who mask their identity. Abbasi and Chen [1] analyze online forums in Arabic and English, employing machine learning techniques to learn a distinctive and large set of linguistic features for each user. Others have studied author detection using similar methods in blogs [14] and in emails [10].

However, there have been very few papers published about author detection across several different collections and domains. Rao & Rohatgi [22] tried to align authors from both mailinglists and newsgroups. They report that the stylistic conventions practiced by users of the different media resulted in very poor detection rates with learning and shallow parsing methods.

In this paper we intend to show the feasibility of performing author detection over several media types such as blogs, user profiles, personal tagging spaces, professional and personal homepages and any other identifiable personal information that can be attributed to a single author. Figure 3 is an example for the kind of variety we seek to explore. The set of eight different pages all written by the same author and published on different hosts consists of several traits that are visually similar, like images and layout, and several traits that are different like title, length, and intended readership.

3. DETECTION BY COMPRESSION

The studies described in section 2 all look at very controlled and contained domains. However, to solve the problem of author detection on the Web it is very costly to employ methods of shallow parsing and machine learning for several reasons. First, feature extraction is a costly process which requires analyzing many aspects of the page and then producing large data structures for storing such information. Secondly, feature extraction in such an uncontrolled environment cannot scale up, as observed by Keogh et al. [14]. The authors follow the work of Benedetto et al. [4] who applied off-the-shelf compression software to extract the compression distance for each pair of pages. Benedetto et al. managed to cluster the world languages by using this feature alone. They have also tried to detect similar authors in a small pool (90 documents) of academic papers. Their reported success rate on this restricted domain is over 95% for pairing texts by the same author. Kukushkina et al. [16] explain the linguistic motivation behind using compression to represent author specific repetition frequencies. Recently, Cilibrasi & Vitanyi [8] explained the theoretical rational behind using compression to represent and then compare entities with complex features.

Using compression instead of textual and structural feature extraction is advantageous for our task since there are so many ways in which two pages written by the same author can be similar. They may share themes, content terms, relative URL path, linking patterns, page layout, color scheme, image filenames, etc. Encoding such a feature set for a collection of pages is a very subjective task. If the feature set is large enough to describe all possible aspects its usage will not scale to large collections such as the Web. Compression captures all of the features that repeat themselves in a single page and treats them as information redundancy. So it may capture HTML structure redundancies as well as stylistic redundancies. The final size of the compressed page is determined by the repeating patterns detected in the compression. By using compression for author detection we hypothesize that every author has a unique compression signature that is similar across all the pages of the same author.

3.1 Compression Distance

The *Normalized Compressor Distance* (NCD) was suggested in [4] (with formal justification in [8]) as a tool for detecting document similarity. Given a compressor *C* and two documents *x*, *y*, we define:

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

Where $C(x)$, $C(y)$ and $C(xy)$, are the bit-wise sizes of the result sequences when using *C* to compress *x*, *y* and the concatenation of *x* and *y*, respectively[1]. NCD assesses the similarity between a pair of documents by measuring the improvement achieved by compressing an information-rich document using the information found in the other document.

In this paper we use a variation of NCD which we term 2-sided NCD (2NCD), with the following definition:

$$2NCD(x, y) = \frac{[C(xy) - C(x)] \cdot [C(xy) - C(y)]}{C(x) \cdot C(y)}$$

2NCD measures separately how much the compression of **each** of the documents is improved by using the information included in the other document. The compression distance assigned to the document pair is the product of these two measurements.

4. EXPERIMENT

We designed an experiment to test whether authors can be detected using only their compression signature, even across different types of writing styles and Web publication types. We collected nearly 10,000 pages including blogs, user profiles, del.icio.us spaces, Flickr photo collections, Wiki style pages, personal homepages, etc., written by 2201 different authors. We then conducted several experiments based on this collection.

4.1.1 Data Collection

In order to collect data for such a large scale experiment it is necessary to ask people to provide a list of Web pages that they have authored. It is not possible to simply crawl the Web for such information without prior knowledge as pointed out by Bar Ilan [3], since it may be that a person is described by others such as in the case of corporate executives, and famous movie actors. Obviously, people also write under pseudonym and it will be difficult to detect them without prior knowledge. We first tried to ask people to send us their collection by email, however we received only several dozens of replies which is not enough for our task. One of the replies stated that the full list of his authored pages can be found on ClaimID.com. ClaimID is an experimental site set up by two students from the University of North Carolina. The site is described by Stutzman & Russell [23] as a system for managing online personal identities. ClaimID allows its users to list URLs that were authored by them and/or about them. The site is a list of user profiles with detailed lists of what information was produced by the author and what was not. We crawled the site,

[1] Assuming that C is a normal compressor (see [8]), and therefore $C(xy) = C(yx)$.

which is publicly available to search engine crawlers, and collected over 8000 unique user information. We then filtered this list and stored only authors who had at least two pages authored by them hosted on two different hosts. We also removed those who had simply duplicated the content of one site and put it up as a mirror on another host (assuming this will be revealed by simple duplicate- or mirror- site detection).

We ended up with 2201 users who authored 9834 different pages. Figure 4 describes the distribution of page types in our collection. This is a very crude division, based on the occurrence of terms in the URL, the anchor or the short description appearing in the ClaimID profile. For example, we labeled a page with the term "blog" if any if the fields contained, even partially, any of the terms *blog*, *livejournal*, *typepad*, *wordpress*, and *fotolog*. "Community-Share" label was assigned to social-space pages marked with *del.icio.us*, *simpy.com*, *blinklist.com*, *ma.gnolia.com*, *connotea.org*, *scuttle.org*, *wists.com*, *shadows.com*, *digg.com*, *slashdot.org*, *myspace.com*, *deviantart.com*, *youtube.com*, etc. "Unknown type" means that there was no trivial way to automatically detect the type of the page from its host name or from the description provided by its author on ClaimID. Manually inspecting some "unknown type" pages revealed that many came from sources such as professional or work-related sites, newspaper articles, contributions to school projects, etc.

We left the files intact, including all HTML and scripts. This was done in order to achieve realistic results that could potentially be applied to any collection of Web pages without any pre-processing. Also, removing HTML markup may have affected the detection of structure and layout characteristics unique to individual authors.

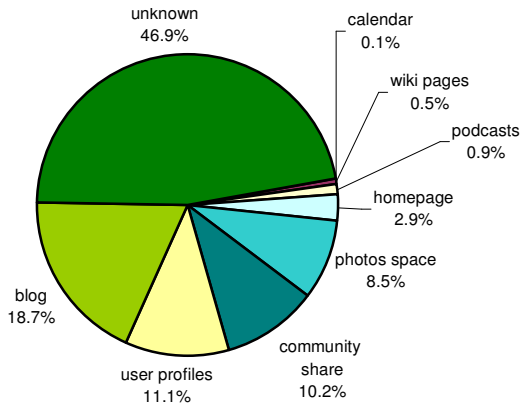

Figure 4 – The percentage of each detected page type in our collection of 9834 pages coming from 2201 different authors.

4.1.2 Common links as baseline comparison

Following Calado et al. [7] who recently tested linkage similarity measures and found link co-citation to yield the best results for topic similarity between documents, we decided that our baseline comparison should be link co-occurrence between each pair of documents.

As a first step to test the existence of link co-occurrence between sets of documents known to be produced by the same author we calculated the amount of shared links for each set. It turned out that about 60% had common links while 40% had no common links between the different pages they have written. The most prolific author had 1283 links appearing repeatedly in the set of the pages he authored. We did not compare against shared textual content since it was not a measure that could scale up to our collection. We also considered using duplicate detection methods, however, after inspecting the documents it seemed that this approach will not yield better results than simply comparing common links.

4.1.3 Detection by Compression Experiment

Motivated by efficiency considerations, we sampled our collection and extracted two smaller sets comprising 1043 documents for the first set and 1109 documents for the second set. The sampling was arbitrary and was designed to sample authors rather than pages. All the pages written by the same author were grouped together and the two samples did not include the same author twice. We worked with these samples to compare each possible pair of documents using link co-occurrence and compression distances.

For each document we computed its shared links with every other document in the sample. For each such pair we also calculated their compression distance by first compressing each document on its own and then compressing the pair together.

For the compression task we used 7za.exe[2], an open source free compressor, which has a relatively large buffer. We found the large buffer to be advantageous for Web pages. The large buffer size also supports our assumption that the compressor is symmetric. We also tried MATLAB's built-in ZIP compressor but found it to be less effective.

4.1.4 Detection by Compression Results

The results of the compression distances computed for each document pair (using 2NCD) are shown in Figure 5 and Figure 6. The figures are the histogram of the values received for each comparison. The green bars represent pairs that actually belong to the same author, while the red bars indicate pairs that were written by different authors. For both samples it is obvious that the green bars accumulate on the left-most side of the chart. This accumulation clearly demonstrates the strength of the compression distance as a method for representing authorship encoded information.

In Figure 5 and Figure 6, the green bars display a bimodal distribution, which is typical to cluster-containing data [23]. In our studied domain we contend that there are two types of relations between documents written by the same author. The first type consists of the cases where a person writes several Web pages with a similar motivation, such as a professional blog and a professional homepage. Since the underlying function of these documents is the same, and they reflect the same purpose, the resulting documents are very similar and therefore the compression distance is very low. This may explain the green slopes on the left end of Figure 5 and Figure 6.

The other type of relation consists of documents which were written by the same author but serve different purposes, such as a a personal calendar and a dig.com entry. These pages will have many dissimilar features. However, since the author is the same the resemblance between these documents will remain. Those documents probably comprise the green hills which spread from compression distance 0.01 to 0.035 in the above figures. Between

[2] http://en.wikipedia.org/wiki/7-Zip

the two types of pages lays a continuum of similarity values, some overlapping with those of unrelated authors.

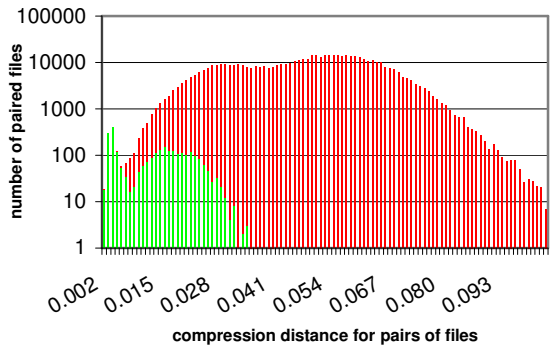

Figure 5 – A histogram of the compression distances computed for each pair of documents in the first sample consisting of 1043 documents. The green bars represent true document pairs. The red bars represent false document pairs.

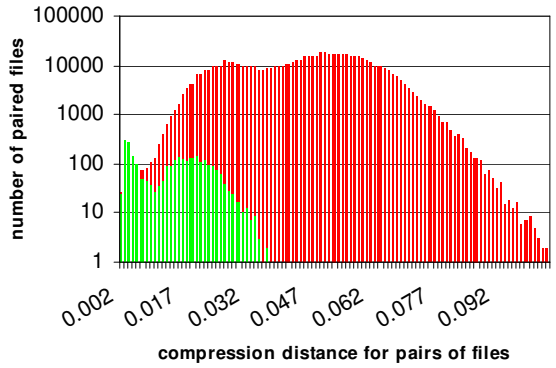

Figure 6 – A histogram of the compression distances computed for each pair of documents in the second sample consisting of 1109 documents. The green bars represent true document pairs. The red bars represent false document pairs.

In order to better visualize the results of the compression-based similarity, we generated a graph known as the Receiver Operating Characteristic (ROC) curve. This curve plots the sensitivity versus the specificity of a system. In our case, each point on the curve plotted in an ROC is a threshold similarity. The horizontal axis of the ROC curve represents the probability that two pages that have a compression similarity index smaller than the threshold will not be from the same author. The vertical axis shows the probability that two pages which have a compression index smaller than the threshold will indeed be from the same author. The ideal curve would touch the upper left corner of the graph, while a random decision will result in a curve from the bottom left corner to the upper right-hand corner. An ROC is usually parameterized by the area under the curve, where 0.5 represents random decision and 1.0 an ideal system.

Figure 7 and Figure 8 show the results of compression-based similarity compared to using the number of co-occurring links as a method for detecting authorship. The area obtained by the latter method is 0.6, only slightly better than chance. Compression-based similarity achieves an area of greater than 0.97, which is close to the ideal detector. Thus, the compression-based similarity offers a superb method for identifying authorship.

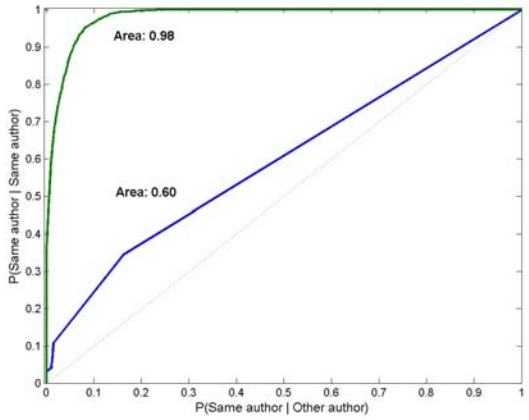

Figure 7 - Receiver Operating Characteristic (ROC) curve plotted for the first experiment. The grey line represents equal chance, blue line represents probability of being correct using common links, and green line represents the probability of being correct using compression sizes.

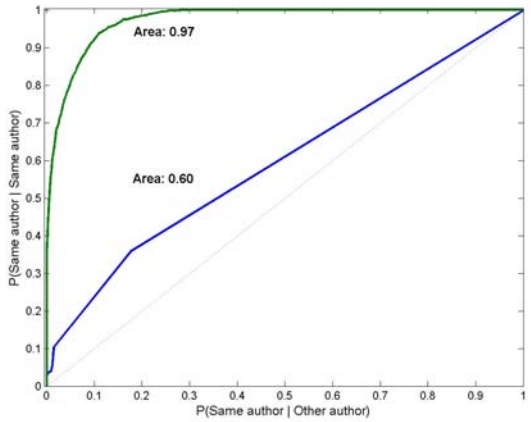

Figure 8 - Receiver Operating Characteristic (ROC) curve plotted for the second experiment. The grey line represents equal chance, blue line represents probability of being correct using common links, and green line represents the probability of being correct using compression sizes.

Table 2 is a color-coded matrix of compression distances calculated for the eight document examples displayed in Figure 3. All the document pairs were assigned low compression distance values which means they were considered similar.

There were no falsely paired documents of that same-author set until the compression distance value doubled from the last true pair. The falsely paired document, appearing in Figure 9, was matched to documents g (0.016), d (0.018), f (0.018), and h (0.018). This brings us to the problem of chaining or clustering together all the scored pairs to create the original set of pages produced by the same author. The next section describes a naïve attempt to cluster the paired documents using only the information provided by the compression distance.

Table 2 - color-coded matrix of compression distances calculated for the pages presented in Figure 3

	a	b	c	d	e	f	g	h
a		0.0051	0.0048	0.0051	0.0065	0.0051	0.0028	0.0028
b	0.0051		0.0033	0.0036	0.0041	0.0036	0.0036	0.0036
c	0.0048	0.0033		0.0038	0.0044	0.0038	0.0033	0.0033
d	0.0051	0.0036	0.0038		0.0041	0.0036	0.0036	0.0036
e	0.0065	0.0041	0.0044	0.0041		0.0041	0.0041	0.0041
f	0.0051	0.0036	0.0038	0.0036	0.0041		0.0036	0.0036
g	0.0028	0.0036	0.0033	0.0036	0.0041	0.0036		0.0036
h	0.0028	0.0036	0.0033	0.0036	0.0041	0.0036	0.0036	

Figure 9 – a page which was the first to be falsely correlated with several of the pages in Figure 3 (with g: 0.016, with d: 0.018, with f: 0.018, and with h: 0.018)

4.1.5 Document clustering

In order to cluster the paired documents we used a naïve clustering algorithm as follows: Given a distance function D and a threshold t, let G = (V, E) be a graph whose vertices are all of the documents in a collection, with an edge connecting every pair of documents (x, y) such that $D(x, y) \leq t$. A cluster of single-author documents is a connected component in G.

The results of applying this algorithm using 2NCD with different thresholds on the two sample sets are given in Figure 10. It should be noted that the data was not manually verified and therefore it may include some noise (for instance a person who registered on ClaimID under two different usernames). The number of same-author pairs is presented along with the error rates produced by using different thresholds. The lines show the number of detected same-author pairs while the bars show the error rate for each threshold. We labeled the document pairs whose compression distance is below the threshold "Original", and the pairs resulting from running the clustering script "Clusters". The total number of true same-author pairs is 2705 and 2745 in sample sets 1 and 2, respectively.

An important observation from this figure is that up to a threshold of 0.008, both error rate and the number of pairs added by the clustering algorithm are relatively small (approximately 10% or lower). This means that given a set of very similar documents, the compression distance identifies almost every pair in the set as related, with relatively few errors. At threshold 0.008, the number of clustered pairs is approximately 3/8 (37.5%) from the total number of truly related pairs.

Estimating the number of those authors who have more than a single Web page to be half of those who maintain blogs yields about 6 million users with at least 12 million pages in the US alone. Detecting nearly 40% of the pages authored by such serial sharers reveals a newly detected community which calls for new methods of exploration and research.

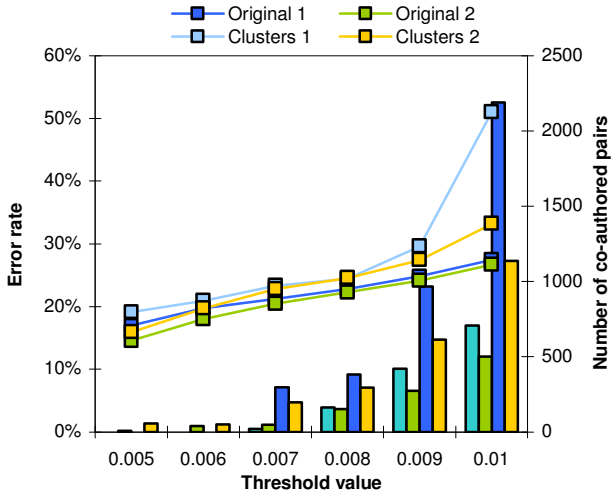

Figure 10 - The number of detected same-author pairs according to compression distance (Original) and clustering algorithm (Clusters), along with the error rates using different thresholds. The lines show the number of detected same-author pairs (out of approx. 2700 real co-authored pairs in each sample), while the bars show the error rate.

5. CONCLUSION & FUTURE WORK

We presented the problem of author detection over a collection of pages originating from different sources and written to serve different online functions. We applied a detection-by-compression algorithm to compute the compression distance for each pair of documents in a collection of pages with a known author. We then showed that it is possible to correctly determine authorship for a considerably large portion of the Web pages based on such a distance, and went on to chain the pairs into document clusters.

It is evident from the studies presented earlier that the youth of today is much more likely to have authored multiple Web pages. When those teens become adults they will probably share much more content on the Web than today's adults. If this prediction is correct then the title "serial sharer" will apply to many more people around the world. Hundreds of millions of people will have their contributions stored all over the Web, managing their personal archiving and memoirs online. Search engines need to prepare for that day with a mechanism for automatically detecting and labeling such individual productivity.

The good news is that search engines already use compression in storing cached versions of documents. The only caveat is the fact that in order to calculate the compression distance for each pair, both files need to be compressed together. This challenge may give rise to new solutions for candidate file pairing that will allow search engines to reduce the number of paired files to be compressed. Such solutions may take usernames found in the URL as a first "rule of thumb" comparison candidacy. Similarly, solutions may be found in computing the probabilities of people co-publishing in certain places, for instance, if a person publishes in del.icio.us they are likely to also have a page on blogger.com, etc.

Such solutions will lead to finding patterns in cross domain adoption of Web applications. It will be easier then to decide which application attracts a larger number of unique users by aligning sites like del.icio.us with blogger and myspace to find common authors. This alignment may also provide insight about what content people choose to publish on one site and not on the other, and why people decide to split their identity and write in several different places.

Incorporating author identification into search engines will advance features such as profiling, expertise location, finer granularity in trend analysis, and may help generating better insights about the sources and motivation for the publication of the retrieved results.

6. REFERENCES

[1] Abbasi, A. and Chen, H. 2005. Applying Authorship Analysis to Extremist-Group Web Forum Messages. IEEE Intelligent Systems 20(5):67-75.

[2] Argamon, S., Šarić, M., and Stein, S. S. (2003). Style mining of electronic messages for multiple authorship discrimination: first results. In Proceedings of ACM KDD'03, pp. 475-480.

[3] Bar–Ilan J. (2006). False Web memories: A case study on finding information about Andrei Broder. First Monday, volume 11, number 9 (September 2006), URL: http://firstmonday.org/issues/issue11_9/barilan/index.html

[4] Benedetto D., Caglioti E., Loreto V. (2002). Language trees and zipping. Physical Review Letters, 88(4):048702.

[5] Bharat, K. and Broder, A. 1999. Mirror, mirror on the Web: a study of host pairs with replicated content. WWW8, appeared in Computer Networks & ISDN Systems, 31(11-16):1579-1590.

[6] Broder A. Z., Glassman S. C., Manasse M. S., Zweig G. (1997). Syntactic clustering of the Web. WWW6, appeared in Computer Networks & ISDN Systems, 29(8-13):1157-1166.

[7] Calado P., Cristo M., Moura E.S., Gonçalves M.A., Ziviani N., Ribeiro-Neto B. (2006). Link-based Similarity Measures for the Classification of Web Documents. Journal of the American Society for Information Science and Technology (JASIST) 57(2):208-221.

[8] Cilibrasi R., Vitanyi P.M.B. (2005). Clustering by compression. IEEE Transactions on Information Theory, 51(4):1523-1545.

[9] Cox Communications, Press Release, March 2006. Available online: http://www.cox.com/takecharge/survey_results.asp

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

[11] Dutton W. H., di Gennaro C., Hargrave A. M. (2005). Oxford Internet Survey 2005 Report: The Internet in Britain. Available online: http://www.worldinternetproject.net/publishedarchive/oxis2005_report.pdf

[12] Holloway T., Bozicevic M., Börner K. (2005). Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. under review. Available online: http://arxiv.org/ftp/cs/papers/0512/0512085.pdf

[13] Horrigan J.B. (2006). Broadband Adoption 2006. Pew Internet & American Life Project. Available online: http://www.pewinternet.org/pdfs/PIP_Broadband_trends2006.pdf

[14] Keogh E., Lonardi S., Ratanamahatana C. A. (2004). Towards parameter-free data mining. In Proceedings of ACM KDD 2004, pp. 206-215.

[15] Koppel M., Schler J., Argamon S., Messeri E. (2006). Authorship Attribution with Thousands of Candidate Authors. ACM SIGIR 2006, pp. 659 - 660.

[16] Kukushkina O.V., Polikarpov A.A., Khmelev D.V. (2000). Using Literal and Grammatical Statistics for Authorship Attribution. Problemy Peredachi Informatsii, 37(2):96-108. Also translated in “Problems of Information Transmission”, pp. 172-184. Available online: http://www.math.toronto.edu/dkhmelev/PAPERS/published/gramcodes/gramcodeseng.pdf

[17] Li J., Zheng R., Chen H. (2006). From fingerprint to writeprint. Commun. ACM 49(4):76-82.

[18] Liang G. (2005). Surveying Internet Usage and Impact in Five Chinese Cities. Report of the Research Center for Social Development, Chinese Academy of Social Sciences (November 2005). Available online: http://news.bbc.co.uk/1/shared/bsp/hi/pdfs/10_02_06_china.pdf

[19] Madden M., Fox S. (2006). Riding the Waves of "Web 2.0". Pew Internet & American Life Project's Report, October 2006. Available online: http://www.pewinternet.org/pdfs/PIP_Web_2.0.pdf

[20] Novak, J., Raghavan, P., and Tomkins, A. (2004). Anti-aliasing on the web. In Proceedings of WWW '04. pp. 30-39.

[21] Princeton Survey Research Associates International for the Pew Internet & American Life Project. (2006). Blogger Callback Survey Final Revised Topline 7/6/06, Data for July 5, 2005 – February 17, 2006. Available online: http://www.pewinternet.org/pdfs/PIP_Bloggers_Topline_2006.pdf

[22] Rao J.R., Rohatgi P. (2000). Can pseudonymity really guarantee privacy? In Proceedings of the 9th USENIX Security Symposium, pages 85–96.

[23] Steinbach M., Ertoz L., Kumar V. (2004). Challenges of clustering high dimensional data. In New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition. Springer, 2004.

[24] Stutzman F., Russell T. (2006). ClaimID: a system for personal identity management. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL), p. 367.

[25] Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 57(3):378-393.

Authorship attribution for small texts: Literary and forensic experiments

Ol'ga Feiguina
Cherches and Associates
Montreal, Canada
olga82@gmail.com

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, Canada M5S 3G4
gh@cs.toronto.edu

ABSTRACT

To capture syntactic structure as a feature for the classification of short texts by their authorship, we use the frequencies of bigrams in the stream of syntactic labels produced by a partial parser. We experimented on literary data (from the Brontë sisters) and simulated forensic data. Syntactic label bigrams were found to be helpful with the former but not the latter.

Categories and Subject Descriptors

I.2.7 [**Artificial intelligence**]: Natural language processing—*Language models, text analysis*

Keywords

Authorship attribution, partial parsing, literary data, forensic data, text classification

1. INTRODUCTION

Methods of authorship attribution developed for literary analysis typically require the document to be long and the comparison corpus to be large (did Shakespeare or Marlowe write this play?). However, in many applications, these assumptions are not valid. In literary analysis, the texts might be relatively short poems or stories. In forensic situations, the documents are likely to be short and the corpus small. For example, the document might be an anonymous letter whose authorship is to be compared with samples from suspects. In the detection of plagiarism, short segments of a longer work may be compared with one another to see if they bear evidence of diverse authorship. Previous approaches to authorship attribution have been unsuccessful on short texts (Burrows 2002) or have succeeded only in narrow domains by using customized and highly domain-specific features (Zheng, Li, Chen, and Huang 2006).

In this paper, we present a method for authorship attribution that is suitable for texts as short as around 200 words. The method makes better use of the syntactic information in the texts than prior approaches.

Earlier research on the use of syntax in authorship attribution has shown both the strengths and the limitations of using either full parses or simple syntactic chunking. Baayen, van Halteren, and Tweedie (1996), working on long, hand-parsed literary texts, represented a text as the bag of rewrite rules used in the syntactic derivation of each of its sentences, and applied vocabulary-richness measures to this bag. The results were better than those obtained by applying the same measures directly to the vocabulary of the text, but Baayen et al despaired of using their method with automatic parsing because its accuracy would be insufficient. Stamatatos, Fakotakis, and Kokkinakis (2000), working with news texts averaging around 1100 words, used simple non-embedded chunking of the text to derive a number of quantitative features for authorship classification. While their results were quite good, the method was dependent on artifacts of their particular chunker, and the error rate was high for shorter texts.

2. BIGRAMS OF SYNTACTIC LABELS

To obtain the strengths of using syntax while minimizing the weaknesses, we use partial parsing (Abney 1996), which produces an embedded but not recursive syntactic structure for sentences. We then represent a sentence as a sequence of the syntactic labels of its bracketed substructures and words, ignoring the brackets and the words themselves; this can be thought of as an approximation to the syntactic structure of the sentence (Hirst and Feiguina 2007). As an example, Figure 1 shows a fragment of a sentence, the corresponding structure from partial parsing, and the stream of labels that we take as its representation. A document can then be represented by the frequencies of bigrams of these syntactic labels; Figure 1 shows the bigrams extracted for the example fragment. These frequencies are then used as features for text classification by authorship. In addition, following Baayen et al, we also use the rewrite rules from the partial parser as a feature — both by simple frequency of use and by vocabulary richness.

3. TESTS WITH LITERARY DATA

We first tested the method on text by the Brontë sisters, selecting them because Koppel, Schler, and Mughaz (2004) had found them to be very hard to discriminate even by sophisticated authorship attribution methods. We took 250,000 words each from novels by Anne Brontë and Char-

SIGIR '07 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection.

Sentence [fragment]:
Let it be theirs to conceive the delight of joy ...

Partial parse:
```
[vp [vx [vb Let]]]
[c [c0 [nx [prp it]] [vx [be be]]] [nx [prp theirs]]]
[infp [inf [to to] [vb conceive]]
      [ng [nx [dt the] [nn delight]] [of of] [nx [nn joy]]]] ...
```

Stream of syntactic labels:
```
vp vx vb c c0 nx prp vx be nx prp infp inf to vb ng nx dt nn of nx nn ...
```

Bigrams of syntactic labels:
```
vp-vx vx-vb vb-c c-c0 c0-nx nx-prp prp-vx vx-be be-nx nx-prp prp-infp infp-inf inf-to to-vb vb-ng ng-nx
nx-dt dt-nn nn-of of-nx nx-nn ...
```

Figure 1: Example of partial parse, with corresponding label stream and bigrams.

lotte Brontë, and divided them at sentence boundaries into fragments of approximately 100, 500, or 200 words. (That is, we pretended that instead of writing novels they had written many independent short texts.) We then tried to classify these short texts by author, using frequencies of syntactic label bigrams as features. As a baseline for comparison, we also tried the task with a number of standard lexical features commonly used in authorship attribution (Graham, Hirst, and Marthi 2005), including frequencies of function words, part-of-speech tags, and word lengths, and vocabulary richness measures. Lastly, we tried combinations of these features with the syntactic label bigram frequencies. For classification, we used a support-vector machine, with ten-fold cross-validation for testing.

Our results, which we present in greater detail in Hirst and Feiguina 2007, are shown in Table 1. These results are based on the complete 500,000-word dataset, but accuracies started to level off when the size of the training set reached about 40,000 words; we discuss training-set size in Hirst and Feiguina 2007. The random baseline (the accuracy that would be achieved by guessing randomly) is 50%.

We observe immediately that, contrary to the results of Baayen et al, vocabulary-richness measures on rules give poor results by themselves, and we exclude them from further discussion, although they do in all cases give a boost to the other syntactic features. For the 1000-word texts, there is a ceiling effect: all conditions do quite well, though syntactic label bigram frequencies alone achieve a 99% accuracy, effectively the same as all lexical features combined. For smaller text sizes, not surprisingly, accuracy drops. However, for both 500-word and 200-word texts, the accuracy achieved by the combination of all feature sets exceeds that of any single set; that is, our label bigram frequencies increase accuracy compared to the use of standard lexical features alone. An examination of the nine label bigrams that were most discriminating found that seven of them involved non-terminal labels, indicating that a sensitivity to syntactic structure is indeed making a difference to the classification.

	Text size		
Features	1000	500	200
Syntactic features			
Label bigram freqs	99.0	93.4	84.9
Rule freqs	93.2	93.4	83.8
Vocab richness of rules	76.6	76.7	70.3
Bigram and rule freqs	98.4	95.8	87.4
All syntactic features	**99.5**	94.2	87.5
Lexical features			
PoS freqs	93.8	93.4	82.7
Other lexical features	97.5	90.5	85.6
All lexical features	98.9	95.0	89.5
All features	99.2	**96.8**	**92.4**

Table 1: Average accuracy (in percent) in 10-fold cross-validation on pairwise classification of Brontë texts, by text size and features used. Boldface indicates best results for each text size.

4. TESTS WITH FORENSIC DATA

Given this success on short literary texts, we then turned to using the method on forensic data — for example, anonymous threatening letters, tip-off notes, etc. We assume that attested writing samples from suspects are available for comparison. Because there is no readily available corpus of such data for use in experiments, we used simulated forensic data: Chaski's (2005a,b) model forensic dataset of short texts. Chaski asked 11 different authors to each write approximately 2000 words in total, choosing from topics such as a threatening letter, an apology, or a complaint. There are a total of 73 texts, ranging from 4 to 10 texts per author and varying widely in length (average length, 265 words). Depending on the method of data analysis, Chaski's own syntactically-aware method achieves 95% accuracy (Chaski 2005a) or 81.5% accuracy (Chaski 2005b) in pairwise author identification on this dataset; we take the latter, the more-recent publication, as definitive.

We applied our method to this dataset, classifying both complete texts, regardless of length, and fragments of approximately 200 words. We tried both pairwise authorship classification (with a dataset size of about 4000 words) and multiclass (1 in 11) classification (with a dataset size of about 22,000 words). For pairwise classification, the random baselines are 50% for uniformly guessing an author and 50 to 70% (depending on the pair) for always guessing the author with the greater number of texts; for multiclass classification they are 9% and 14% respectively. The results are shown in Tables 2 and 3. (We have dropped vocabulary richness of rules as a feature and added some new combination feature sets.)

Features	Text size Whole	200
Syntactic features		
Label bigram freqs	86.1	78.8
Rule freqs	87.3	72.4
Label bigram and rule freqs	88.3	75.4
Lexical features		
PoS freqs	89.2	84.1
Other lexical features	84.4	83.2
All lexical features	**91.2**	**85.6**
Combinations		
Label bigrams and other lexical	88.3	83.3
Label bigrams and all lexical	89.3	80.0
All features	88.7	75.6

Table 2: Average accuracy (in percent) in 10-fold cross-validation on pairwise classification of simulated forensic texts, by text size and features used. Boldface indicates best results for each text size.

Features	Text size Whole	200
Syntactic features		
Label bigram freqs	57.5	39.6
Rule freqs	56.2	25.5
Label bigram and rule freqs	56.2	34.9
Lexical features		
PoS freqs	**60.3**	34.0
Other lexical features	41.4	**50.9**
All lexical features	51.0	49.1
Combinations		
Label bigrams and other lexical	**60.3**	48.1
Label bigrams and all lexical	58.9	50.0
All features	**60.3**	37.7

Table 3: Average accuracy (in percent) in 10-fold cross-validation on multiclass (1 in 11) classification of simulated forensic texts, by text and features used. Boldface indicates best results for each text size.

For pairwise classification (Table 2), the accuracy is lower overall than for the Brontë data, but is of course based on much smaller training data; in fact, the accuracy is much higher than for a Brontë dataset of the same size. But we observe quite a different pattern in the results compared to those for the Brontë texts. Here, the standard lexical features do better than the syntactic features and better than Chaski's method, especially for the smaller texts. Moreover, adding the syntactic features to the lexical features degrades performance. Examination of the most-discriminating label bigrams shows that, contrary to the Brontë case, almost all of them involved terminal symbols, and the method's sensitivity to syntactic structure was hardly used at all.

The results for multiclass classification also showed a superiority for standard lexical features, although the overall pattern was quite different yet again (Table 3), with the efficacy of a feature set or combination varying widely with text size. For whole texts, while label bigrams perform better than lexical features in general, frequencies of part-of-speech tags alone do best, and combinations do worse or no better. For the 200-word fragments, however, the performance with part-of-speech tags degrades severely, while that of other lexical features actually improves.

These results can be explained in part simply by the small and unbalanced dataset that was used. However, it is also clear that the writing styles in the simulated forensic data were more distinct from one another than the styles of the Brontë sisters are, and their differences were more at the lexical level than the syntactic level. That is, "ordinary writers" are an "easier problem" than the Brontë sisters. This suggests that the use of an intermediate feature such as PoS-tag bigrams might be more successful for this kind of data. (Spassova and Turell (2006) have carried out experiments with high-frequency PoS-tag trigrams for authorship attribution and reported promising results.)

5. CONCLUSION

Syntactic label bigrams were found to be a helpful feature in discriminating authorship of short texts by the Brontë sisters, but were not helpful on simulated forensic data in which syntactic distinctions seemed to be less necessary. This can be attributed in part to the imbalance and small size of the forensic dataset, but it is also a reminder that features for authorship attribution can be very genre- or situation-specific.

6. ACKNOWLEDGEMENTS

Our research is funded by the Natural Sciences and Engineering Research Council of Canada. We are grateful to Carole Chaski for allowing us to use her dataset of simulated forensic texts, to Neil Graham for his assistance and for the use of his software for the lexical features that were used in his work, and to Nadia Talent for helpful comments on an earlier draft of this paper.

7. REFERENCES

[1] Abney, Steven (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4): 337–344.

[2] Baayen, R. Harald; van Halteren, Hans; and Tweedie, Fiona J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3): 121–131.

[3] Burrows, John (2002). 'Delta': A measure of stylistic difference and likely authorship. *Literary and Linguistic Computing*, 17(3): 267–287.

[4] Chaski, Carole E. (2005a). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1).

[5] Chaski, Carole E. (2005b). Computational stylistics in forensic author identification. *SIGIR Workshop on Stylistic Analysis of Text for Information Access.*

[6] Graham, Neil; Hirst, Graeme; and Marthi, Bhaskara (2005). Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4): 397–415.

[7] Hirst, Graeme and Feiguina, Ol'ga (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, to appear.

[8] Koppel, Moshe; Schler, Jonathan; Mughaz, Dror (2004). Text categorization for authorship verification. *Eighth International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale, Florida.

[9] Spassova, Maria S. and Turell, M. Teresa (2006). The use of morpho-syntactically annotated tag sequences as markers of authorship. *Proceedings of the Second European IAFL Conference on Forensic Linguistics / Language and the Law*, Barcelona.

[10] Stamatatos, Efstathios; Fakotakis, Nikos; and Kokkinakis, George (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4): 471–495.

[11] Zheng, Rong; Li, Jiexun; Chen, Hsinchun; and Huang, Zan (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3): 378–393.

Authors, Genre, and Linguistic Convention

Jussi Karlgren and Gunnar Eriksson
Swedish Institute of Computer Science
jussi@sics.se, guer@sics.se

Authorship, Language, and Individual Choice

The basic premise underlying authorship attribution studies is that while the form of expression in language is in some respects strictly bound by linguistic rule systems and in others somewhat constrained by topic and genre, it is in some other respects freely available for configuration or preferential choice by author or speaker. This individual variation can be observed, detected, and predicted to some extent, using traditional stylostatistic measures. For instance, word length varies from author to author [Mendenhall, 1887, e.g.]; sentence length likewise; and some forms of lexical expression are characteristic of speakers, either on an individual level or on a community level [Book of Judges].

Common to most computation of individual difference in authorship is that the features used to characterise and distinguish authors are based on the repeated measurement of some, often clause-internal, property at independent positions in the text and then *aggregating* these *pointwise* measures by averaging or normalising the result. In this position paper we claim that by measuring *local* clause- or even word-internal properties, and by aggregating in such a way that the relation between individual observations is destroyed, we obtain features that are most likely to have been subject to pressure from conventionalisation and grammaticalisation processes in language. Instead, we want to examine features that capture differences between authors on a level of textual structure where the space for individual choice is wide: the organisation of informational flow and narrative frame. Such features can be obtained by studying configurations and progressions of observable properties above the clause level. We will call this family of aggregated features *configurational* in contrast to the typical pointwise measurements.

Rules, Constraints, and Conventions

The patent regularities of linguistic expression are formed by constraints – rules, conventions, and norms which can be of a biological, social, psychological, or communicative character. While language use is regular to a great extent, the rules that govern it change continously. Observations and descriptions of language from an earlier time can become obsolete; early samples of language can be all but incomprehensible to the modern reader (and presumably, to the listener). The origin of linguistic constraints, their ontological nature, and their life span or life cycle is much debated in linguistics, but grammaticalisation, the process whereby optional linguistic behaviour becomes a norm, is assumed to proceed sequentially, with many partially counteracting motivating factors and driving forces, variously ascribed to economy of expression, redundancy, tolerance towards noise, and factors related to social cohesion [Dahl, 2006, e.g.].

Many obligatory grammatical items are likely to have started their life as idiosyncratic choice, become accepted in some community as markers of some function, informational or social, and thence migrated to broader linguistic usage.

Given this progress from characteristics of individual usage to conventionalisation, and further to grammatical constraints, the claim underlying these first experiments is that the degree of leeway or freedom for the individual user varies, not only between some specific lexical or syntactic item, but between some *types* of observable items: text-global patterns, e.g. being less rule-bound than local clause-internal structure.

In brief, linguistic items grammaticalise, but first conventionalise. Some choices are optional, some non-optional. All such choices are not as accessible to the process of grammaticalisation. The linguistic items most studied in the fields of linguistics, information access, and stylostatistics are lexical or syntactic on a local level. These are precisely the situation-independent, topic-independent, speaker-independent features most susceptible to linguistic control and grammaticalisation.

The different levels of constraints are shown in Figure 1. There is good reason for syntacticians to study the local and rule-bound variation – the task of linguistics is to generalise from observations to rules; for the purposes of authorship attribution the converse is the case – the task is to find the characteristic and the special. Global textual patterns are available for author choice and will provide better purchase for discrimination of individual style than choice on a level where conventions are strong, observable usage for language users less sparse, and grammar and grammaticalisation hold fast.

SIGIR '07 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection

Free	Author	Repetition, organisation, elaboration
Convention	Genre	Lexical patterns, patterns of argumentation, tropes
Rule	Language	Syntax, morphology

Figure 1: Levels of constraints.

Observanda — Features

What sort of features do we, as authorship attribution experimentalists, then have recourse to? Typically, text categorisation studies compute observed frequencies of some lexical items, or some identifiable constructions. An observed divergence in a text sample from the expected occurrence of that specific item (with prior information taken into account) is a mark of individuality and can be used in the process of identifying author (or, indeed, similarly, genre or topic).

This type of measurement only aggregates local statistics in texts and is not as likely to yield as much individual variation as will variation as measured over the length of the text, on the level of information organisation: examples might be term recurrence [Katz, 1996] or term patterns [Sarkar, 2005]; type-token ratio [Tallentire, 1973]; rhetorical structure; measures of cohesion and coherence [Halliday, 1978]; measures of lexical vagueness, inspecificity, and discourse anchoring; and many other features with considerable theoretical promise but rather daunting computational requirements.

Our hypothesis is that author (and speaker) choice on the level of informational structuring and organisation is less subject to pressure from conventionalisation and grammaticalisation processes. This both by virtue of wide scope, which limits the possibilities of observers to track usage, as well as the many degrees of freedom open for choice, which makes rule expression and rule following inconvenient.

In the present first experiment two simple binary features were calculated:

advl the occurrence of more than one adverbial expression of any type in a sentence, and

clause the occurrence of more than two clauses of any type in a sentence.

Each sentence was thus given the score 1 or 0 for each of the two features. The choice of features was purposely kept simple – both these features are simple to compute, but have pertinence to informational and topical organisation, "clause" being a somewhat more sophisticated proxy for syntactic complexity than the commonly used sentence length measure, and "advl" an estimate of topical elaboration and narrative anchoring of the text. An example analysis is given for reference in section .

Aggregation

Returning to the main claim of this paper, we investigate whether the introduction of configurational features spanning over text rather than local measurements might improve the potential for categorisation of authors. We wish to find an aggregation method which allows us to preserve some of the sequential information of author choice progression: as a candidate we measure the two features studied over a sequence of sentences, and record the resulting transition pattern for each feature over each text.

The experiment is designed to investigate whether using such longitudinal patterns improves the potential for author identification *more* than it improves the potential for genre identification: these transition patterns can then be compared for varying window lengths — the operational hypothesis being that a longer window length would better model variation over author rather than over genre.

Experimental data

We performed an experiment using

The method shown above example was performed on a larger set of genre-categorised texts. For the full experiment, one year of newsprint from the Glasgow Herald was used, about 34 000 articles in all. About half of the articles are tagged for "Article type", and 28 000 have a byline. 8 article types, as given in Figure 2, are found in the collection, and 244 authors with more than 500 sentences. The texts were preprocessed by the Connexor tool kit for English morphology, surface syntax, and syntactic dependencies.

ARTICLETYPE	n
advertising	522
book	585
correspondence	3659
feature	8867
leader	681
obituary	420
profile	854
review	1879
total	17467

Figure 2: Sub-genres of the Glasgow Herald.

Measurements and metrics

The measurements for the two chosen variables are given in Figure 3 for all genres and for some authors – the number of authors is large; only the authors with the highest and lowest scores for each variable are shown. The table shows, somewhat unsurprisingly, that the genres is more consistent with each other than are authors: some authors have really very few clauses ($clause_{min} = 0.52$) and very few adverbials ($advl_{min} = 0.39$) in their sentences, but all genres have a somewhat consistent density of subclauses and adverbials, spanning from 0.866 to 0.899 and from 0.601 to 0.735, respectively.

Transition patterns

To obtain the longitudinal patterns discussed above, each item, "clause" and "advl", was measured over sliding windows of one to five sentences along each text, and the occurrence of the feature was recorded as a *transition pattern* of binary occurrences, marking the feature's absence or presence in the sentences within the window. The first and last bits of text where the window length would have extended

feature	clause	advl
advertising	0.899	0.682
book	0.832	0.637
correspondence	0.918	0.705
feature	0.929	0.689
leader	0.931	0.735
obituary	0.784	0.601
profile	0.921	0.712
review	0.866	0.646
author $clause_{max}$	0.96	
author $clause_{min}$	0.52	
author $advl_{max}$		0.81
author $advl_{min}$		0.39

Figure 3: Relative presence of features "clause" and "advl" in sentences.

over the text boundary were discarded. The feature space, the possible values of the feature with a certain window size is thus all the possible transition patterns for that window size. For windows of size two, the feature space consists of four possible patterns, for windows of size five, thirty-two, as shown in Figure 4.

window size	patterns	number patterns
1	$1, 0$	2
2	$11, 10, 01, 00$	4
3	$111, 110, 101, 100$ $011, 010, 001, 000$	8
4	$1111, ..., 0000$	16
5	$11111, ...,$ $11101, 11100, ...,$ $..., 00000$	32

Figure 4: Feature space for varying window size.

Models of probability

The observed presence of a feature in a pattern, normalised for sentence frequency, yields a crude estimate of probability of recurrence of any observed pattern in further texts in the same category – the same genre or same author. Such a probability distribution describes the density of occurrence over the different features values – how often some feature is likely to occur, compared to the others.

Thus, as an example, any text in the category "correspondence", using a feature space for the feature "clause" based on a window size of three, has the relative probabilities describable as a vector of probability estimates – and is likely to have about two thirds of sentences in runs without multiple clauses, which can be seen from the last position in the vector below. Likewise, the first position of the vector tells us that the probability of finding three sentences in sequence with multiple clauses in a text in this category is 0.0069:

$$p_3(correspondence) =$$

$$= \{p_{111}, p_{110}, p_{101}, p_{100}, p_{011}, p_{010}, p_{001}, p_{000}\} =$$

$$= \{0.0069, 0.0654, 0.00903, 0.155, 0.00454, 0.0363, 0.0486, 0.674\}$$

Evaluation

In categorisation tasks, an unknown item – in this case, a text – with an observation or estimate of feature values, is matched to the category closest to it – in some way, using some algorithm, and some definition of "closest". In these experiments we choose not to test our probability distributions applied to categorisation, to avoid the noise necessarily introduced by the categorisation methodology itself, but instead use an intrinsic assessement of the probability distributions over the target categories.

The assumption we make is that if the set of target categories is well distributed over the feature space, matching unknown items to it will be easier than if the categories are clustered together. Or, in other words, one wishes to find features which separate categories well. So, given a particular feature space we wish to use some measure for how widely it separates the target categories at hand. Figure 5 shows the probability estimates for the eight genres and some randomly picked authors in the material with a window size of 2 for the feature "clause". The question is how distinct this estimate finds the categories to be.

Distance between probability distributions are commonly measured or assessed using the Kullback-Leibler divergence measure[Kullback and Leibler, 1951]. Since the measure as defined by Kullback and Leibler is asymmetric, we use a symmetrised version, a harmonic mean given by [Johnson and Sinanović, 2001].

$$d_{kls} = \frac{1}{\frac{1}{\sum_{i=0}^{n} p_i \times log_2(p_i/q_i)} + \frac{1}{\sum_{i=0}^{n} q_i \times log_2(q_i/p_i)}}$$

The divergence is a measure of distance between two categories. In this experiment, for each condition, we report a sum of pairwise divergences for the set of categories. A large figure indicates a greater separation between categories – which is desirable from the perspective of a categorisation task, since that would indicate better potential power for working as a discriminating measure between the categories under consideration.

The cateory set is vastly different for authors and genres. As there are eight genres and 244 authors with more than 500 sentences, the sums of pairwise divergences for the two category sets are of different orders of magnitude, and in order to facilitate comparison between authors and genres, we randomly select eight authors, compute the sum of pairwise differences for that set, repeat this fifty times (with replacement), and use the mean of the resulting divergences as the result.

For each window length, the sum of the symmetrised Kullback-Leibler measure for all genres or authors is shown in Figure 6. The figures can only be compared horizontally in the table — the divergence figures for different window sizes (the rows of the table), cannot directly be related to each other, since the feature spaces are of different size. This means that we cannot directly say if window size improves

genre	p_{11}	p_{10}	p_{01}	p_{00}
feature	0.022	0.078	0.056	0.84
review	0.041	0.13	0.072	0.76
advertising	0.011	0.072	0.039	0.88
profile	0.016	0.056	0.040	0.89
leader	0.016	0.055	0.023	0.91
correspondence	0.066	0.15	0.051	0.73
obituary	0.0079	0.072	0.023	0.90
book	0.038	0.084	0.069	0.81
author	p_{11}	p_{10}	p_{01}	p_{00}
Stephen McGinty	0.013	0.071	0.052	0.86
Ian Paul	0.021	0.050	0.018	0.92
James O'Brien	0.018	0.11	0.088	0.78
Hugh Dan MacLennan	0.19	0.097	0.032	0.68
Tom McConnell	0.013	0.11	0.052	0.82
William Tinning	0.0062	0.071	0.020	0.90
Andrew Mackay	0.018	0.063	0.038	0.88
Charlie Allan	0.0067	0.047	0.032	0.91
Robert Ross	0.010	0.064	0.027	0.90

Figure 5: Probability estimates for genres and some authors, window size 2, feature "clause".

Window size	Genre		Author	
	"clause"	"advl"	"clause"	"advl"
1	0.5129	0.1816	0.7254	0.4033
2	0.8061	0.3061	1.3288	0.8083
3	1.1600	0.4461	2.1577	1.2211
4	1.4556	0.6067	2.3413	1.8111
5	1.7051	0.7650	3.0028	2.2253

Figure 6: Occurrence patterns' effect on features "clause" and "advl".

the resulting representation or not, in spite of the larger divergence values for larger window size. Bearing that caveat in mind, the relative difference between the features can be compared, and the table gives us purchase to make two claims.

Firstly, comparing both features for each window size between genre and author representations we find that the *difference* between genre categories and author categories is greater for large window sizes. This speaks to the possibility of our main hypothesis holding: a larger window size allows a better model of individual choice than a shorter one.

Secondly, we find that the feature "advl" seems to make relative gains compared to feature "clause" for the larger window size, for the author case: while "clause" still shows a larger value, the relative difference is smaller for the larger window size. This speaks to the possibility of finding better informed feature sets for the larger contextual models to improve distinction between individuals rather than genres.

Conclusions

This experiment was a first shot at finding whether more sequential features might not be better than local ones for distinguishing between genres and authors.

Our *topical goal*, for these experiments, restated, is that lenghtier text spans might give better purchase for finding features open to author choice as compared to locally computed features, mostly determined by syntax. Adverbials, as an example, might be expected to have a certain occurrence frequency in any genre or topic, but the placement of them in text and their resulting distribution can be assumed to be up to individual choice rather than genre or topical convention or syntactic constraint.

The results of our experiment show that configurational features do give different results from pointwise features; they also support our contention that author categories and genre categories should be identified and discriminated in different ways – in the one case, identifying conventions, in the other, avoiding them.

At this juncture, the task is finding more (and more informative) features and factors of the less-conventionalised levels of the linguistic system, measuring them, evaluating them, and understanding and diagnosing their import on the knowledge representation we choose for an application. The features we expect to study are intended to reach beyond occurrence statistics, to measure presence or repetition rather than frequency, to avoid notions such as average and mean and instead to model patterns, trends, bursts and variation.

The *methodological goal* of the experiment is to build an experimental process based on hypotheses informed by some sense of textual reality, rather than computational expediency, and to evaluate the results by discriminatory power, not by application to noisy task. We will further investigate the diagnostic power of e.g. divergence measures, rather than sample-based experiments, to study the potential usefulness of competing knowledge representation schemes.

Choice points left by the wayside

Some questions clamor for attention in this specific experimental setting:

- Is Kullback-Leibler divergence (and its current sym-

metric implementation) the right measure to determine distance between observed occurrence patterns?

- Is summing pairwise divergences the best way of modelling the consistency of a set of category models? Maybe measuring the separation between the two closest neighbours in a set would be better?
- If we would happen to be convinced that transitional patterns are better than local singularities as a feature base – is the model presented here a step in the right direction?
- What better kernel features – beyond adverbial and clause count – should we try utilising?

Acknowledgments

This experiment was funded by the Swedish Research Council. We are grateful to our colleague Anders Holst for providing us with formulæ and valuable intuitions and starting points.

Example analysis

The following three texts describe the same event and were taken from various newsfeeds on August 26, 2007. Feature measurements are given in Table 7.

Example: Text 1

A powerful earthquake [jolted]$_{clause}$ eastern Indonesian islands [in North Maluku province]$_{advl}$ [Thursday]$_{advl}$, prompting government authorities to a tsunami warning. The quake, measuring 6.6 [on the Richter scale]$_{advl}$, [took place]$_{clause}$ [at about 0540 GMT]$_{advl}$, shaking Halmahera and nearby islands [in North Maluku province]$_{advl}$, [said]$_{clause}$ Fauzi, an official at Jakarta's Meteorology and Geophysics Agency. According [to the US Geological Survey USGS]$_{advl}$, the quake [was measured]$_{clause}$ [at 7.0 on the Richter scale]$_{advl}$. "We have [issued]$_{clause}$ a warning that the quake [could [potentially]$_{advl}$ trigger a tsunami]$_{clause}$," Fauzi [told]$_{clause}$ Deutsche Presse-Agentur dpa. He [said]$_{clause}$ the quake [took place]$_{clause}$ [about 57 kilometres beneath the seabed]$_{advl}$. No immediate casualties or injuries [were reported]$_{clause}$ [from the quake]$_{advl}$. Indonesia [is]$_{clause}$ located [in the Pacific volcanic belt]$_{advl}$ known as the "Ring of Fire," where earthquakes and volcanoes are common. [On December 26, 2004]$_{advl}$, a massive 9.0-magnitude earthquake, which [triggered]$_{clause}$ gigantic tidal waves, [devastated]$_{clause}$ thousands of homes and buildings [along the coastline of northern Sumatra]$_{advl}$, leaving around 170,000 people dead or missing [in Indonesia]$_{advl}$ and thousands more dead and injured [along the Indian Ocean coastline]$_{advl}$.

Example: Text 2

A powerful earthquake [rocked]$_{clause}$ eastern Indonesia [on Thursday]$_{advl}$, sending residents fleeing [from swaying homes and hospitals]$_{advl}$, authorities and witnesses [said]$_{clause}$. There [were]$_{clause}$ no immediate reports of damage. The quake, which [had]$_{clause}$ a preliminary magnitude of 7, [triggered]$_{clause}$ a tsunami warning but the alert [was]$_{clause}$ [quickly]$_{advl}$ lifted after it [became]$_{clause}$ clear no destructive waves [had been]$_{clause}$ generated, the country's geophysics agency [said]$_{clause}$. The earthquake [struck]$_{clause}$ [under the Maluku Sea]$_{advl}$ [at a depth of 20 miles]$_{advl}$, the U.S. Geological Survey [said]$_{clause}$ [on its Web site]$_{advl}$. The quake's epicenter [was]$_{clause}$ more than 130 miles [north of Ternate city]$_{advl}$. "We [felt]$_{clause}$ a strong tremor [for almost a minute]$_{advl}$, people [ran]$_{clause}$ [in panic]$_{advl}$ [from buildings]$_{advl}$, [said]$_{clause}$ George Rajaloa, a resident in Ternate. "Children [are]$_{clause}$ crying and their mothers [are]$_{clause}$ screaming, but there is no damage [in my area]$_{advl}$." Indonesia, the world's largest archipelago, [is]$_{clause}$ prone [to seismic upheaval]$_{advl}$ [due to its location on the so-called Pacific "Ring of Fire,"]$_{advl}$ an arc of volcanoes and fault lines encircling the Pacific Basin. [In December 2004]$_{advl}$, a massive earthquake [struck]$_{clause}$ [off Sumatra island]$_{advl}$ and triggered a tsunami that [killed]$_{clause}$ more than 230,000 people [in a dozen countries]$_{advl}$, including 160,000 people [in Indonesia's westernmost province of Aceh]$_{advl}$. [Just over a year ago]$_{advl}$, another quake-generated tsunami [killed]$_{clause}$ around 600 people [on Java island]$_{advl}$.

Example: Text 3

[According to the United States Geological Survey USGS]$_{advl}$ a strong magnitude 6.9 earthquake [has struck]$_{clause}$ Indonesia [in the Molucca Sea]$_{advl}$ [approximately 220 kilometers 135 miles north of Ternate, Maluku Islands, Indonesia]$_{advl}$ [at a depth of 44.6 kilometers 27.7 miles]$_{advl}$. The Japan Meteorological Agency [reports]$_{clause}$ the quake at a magnitude 7.0 with a depth of 50 km. An unnamed official with the USGS [says]$_{clause}$ "there [is]$_{clause}$ a potential that a tsunami [might develop]$_{clause}$, [judging from the magnitude]$_{advl}$," but no tsunamis [were]$_{clause}$ reported. "We [have]$_{clause}$ lifted the warning. [After monitoring]$_{advl}$, there [were]$_{clause}$ no signs of tsunami," [said]$_{clause}$ the Indonesian head of the country's geology agency, Fauzi.[Initially]$_{advl}$, Fauzi [issued]$_{clause}$ a tsunami warning saying "we [have issued]$_{clause}$ a warning that the quake [could]$_{clause}$ [potentially]$_{advl}$ trigger a tsunami."There [are]$_{clause}$ no reports of injuries, deaths or damage. One resident in Ternate [said]$_{clause}$ that he "[felt]$_{clause}$ a strong tremor [for almost a minute]$_{advl}$, people [ran]$_{clause}$ [in panic]$_{advl}$ [from buildings]$_{advl}$. Children [are]$_{clause}$ crying and their mothers [are]$_{clause}$ screaming but there [is]$_{clause}$ no damage [in my area]$_{advl}$." [Earlier]$_{advl}$ the National Oceanic and Atmospheric Administration NOAA [had issued]$_{clause}$ a tsunami bulletin stating that local high waves [could]$_{clause}$ be possible, but a widespread tsunami [is]$_{clause}$ "not expected [based on historical earthquake data]$_{advl}$."

	Text 1	*Text 2*	*Text 3*
Sentences	8	10	10
Words	175	213	203
wps	6.6	6.2	6.2
cpw	21.9	21.3	20.3
clause	4	6	5
advl	4	6	4
1	- +	+ +	- +
2	+ +	- -	- -
3	- +	+ -	+ -
4	+ -	+ +	- -
5	+ -	- -	+ -
6	- -	+ +	+ +
7	- -	+ -	- -
8	+ +	- +	+ +
9		+ +	+ -
10		- +	+ +

Figure 7: Example texts, measurement of features

1. REFERENCES

In *Book of Judges, King James Version, Old Testament*, chapter 12, pp. 5–6.

Östen Dahl. 2006. *The Growth and Maintenance of Linguistic Complexity*. John Benjamins, Amsterdam, Philadelphia.

M A K Halliday. 1978. *Language as social semiotic*. Edward Arnold Ltd, London.

Don H Johnson and Sinan Sinanović. 2001. "Symmetrizing the Kullback-Leibler distance". *IEEE Transactions on Information Theory*.

Slava Katz. 1996. "Distribution of content words and phrases in text and language modelling". *Natural Language Engineering*, 2:15–60.

S Kullback and R A Leibler. 1951. "On informmation and sufficiency". *Annals of Mathematical Statistics*, 22:79–86.

T.C. Mendenhall. 1887. "The Characteristic Curves of Composition". *Science*, 9:237–249.

Avik Sarkar, A de Roeck, and P H Garthwaithe. 2005. "Term re-occurrence measures for analyzing style". In *Textual Stylistics in Information Access. Papers from the workshop held in conjunction with the 28th International Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil, August. ACM SIGIR.

D. Tallentire. 1973. "Towards an Archive of Lexical Norms: A Proposal". In A. Aitken, R. Bailey, and N Hamilton-Smith, editors, *The Computer and Literary Studies*. Edinburgh University Press.

Investigating topic influence in authorship attribution

George K. Mikros
Department of Italian and Spanish Language and Literature
University of Athens
Panepistimioupoli Zografou - 15784
Athens, GREECE
+30 210 6511344
gmikros@isll.uoa.gr

Eleni K. Argiri
Department of Linguistics
University of Athens
Panepistimioupoli Zografou - 15784
Athens, GREECE
eleniargiri@hotmail.com

ABSTRACT
The aim of this paper is to explore text topic influence in authorship attribution. Specifically, we test the widely accepted belief that stylometric variables commonly used in authorship attribution are topic-neutral and can be used in multi-topic corpora. In order to investigate this hypothesis, we created a special corpus, which was controlled for topic and author simultaneously. The corpus consists of 200 Modern Greek newswire articles written by two authors in two different topics. Many commonly used stylometric variables were calculated and for each one we performed a two-way ANOVA test, in order to estimate the main effects of author, topic and the interaction between them. The results showed that most of the variables exhibit considerable correlation with the text topic and their exploitation in authorship analysis should be done with caution.

Keywords
Authorship Attribution, Stylometry, Topic-neutral features.

1. Introduction
Authorship attribution research is based on the "authorship fingerprint" notion. According to this view, each person possesses an idiosyncratic way to utilize their linguistic means, which are unique and their quantitative description can discriminate him/her among every other possible author. In order to find which parts of the human linguistic behavior reflect authorship, researchers have investigated a large number of text characteristics in many linguistic levels. We now know that there are at least 1000 textual attributes relevant to authorship [24]. The selection of these variables is based on their ability to reveal subconscious mechanisms of language variation, which are unique to each author. Therefore, authorship analysis is based on detecting and counting unconscious linguistic habits that are directly related to the text author.

2. Related work

2.1 Corpora controlled for topic in authorship attribution studies
Recently, text metadata influence has been acknowledged as a serious bias in authorship attribution studies. Rudman [24] provides a systematic exposition of the various pitfalls of authorship research and cites specifically that the corpora used for authorship analysis should be matched for genre and time period. Since then, many studies appeared, systematically using corpora that are controlled for topic, genre, medium etc. Baayen et al. [3] created a balanced corpus of written essays in 3 different genres and in 3 topics for each genre. Graham [8] used the Risks corpus, a one-topic corpus, which consists of nearly 1 million words of postings on the Forum on Risks to the Public in Computers and Related Systems (comp.risks). Koppel & Schler [14] used an e-mail discussion group concerning automatic information extraction. It included 480 e-mails written by 11 different authors, during a period of one year. All posts were about the same subject, forming a highly homogeneous corpus with regard to topic. Luyckx & Daelemans [16], in order to isolate the effects of topic and genre, collected 300 texts on the same topic and genre, distributed in 3 author categories (2 separate authors and 1 author category named "Others" with texts of 10 different authors and some collaborative articles of the previous two authors). Argamon et al. [1] developed a benchmark collection of electronic messages for experimentation on author attribution. The collection was based on three Usenet groups with different topics (books, computer theory, programming language C). In each topic, four subcorpora were created, based on different numbers of authors for attribution. In Mikros [22], authorship attribution was attempted in a highly homogeneous newswire corpus, controlled for topic, genre and medium. In total, 1200 texts were collected, written by four different authors in the same topic (Politics).

2.2 Topic independent features
Stylometric variables used in authorship attribution should be independent of any metalinguistic entity, that is genre, topic, medium, chronological era etc. At the same time, they should have a reasonable frequency of occurrence, in order to facilitate their statistical analysis. The above characteristics are fulfilled in the lexical level by the well-known class of function words.

Mosteller & Wallace [23] were among the first to search for text attributes that were systematically topic-neutral. They ended up using specific function words, which have high frequency of occurrence and at the same time remain corpus independent. Recently, Koppel et al. [15], using experimental methodology, found that function words are indeed the best candidates for a universal, corpus-independent feature set for authorship attribution. They used the measure of "stability", which represents quantitatively the degree of available synonymy of a specific linguistic item. Function words are unstable, in the sense that they can be substituted easily in a passage, without affecting the meaning of the text.

Although the frequency of function words has been proved a reliable author discriminator feature in many studies, there are

SIGIR '07 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection

many other stylometric variables which have been used extensively and at least in theory are topic-neutral. Many of them are smaller than the word units, such as characters. At this sub-word level we can safely assume that it is very difficult to trace conscious linguistic usage. Other variables attempt to capture the vocabulary size used in a text, such as Yule's K and Language Density. These measures should also be topic independent, and since vocabulary "richness" is an author's characteristic it should not correlate with topic information. Readability measures, such as word length and sentence length, are also some of the oldest and most common features used in authorship attribution studies and are used extensively as topic-neutral variables.

2.3 The effects of stylistic choices in topic categorization

Although most stylometric features used in authorship attribution studies are considered to be topic independent, recent advances in text topic categorization have shown that topic categorization accuracy can be further improved, if we add stylistic information to the classifier models. Relevant research of stylistic analysis in text categorization has shown that stylistic markers, utilized notably for authorship attribution studies, play at least an auxiliary role in topic classification. The first important attempts to construct text classification systems for recognizing text genres and thus set the foundations for further research were the works of Kalgren & Cutting [11] and Karlgren [12], who used Biber's [4] feature set and Discriminant Function Analysis (DFA) to classify documents according to genre. Kessler, et al.[13] used cue words for the same purpose.

The reliability of style markers as topic discriminators was investigated by Argiri [2] in experiments involving the categorization of Internet articles into predefined thematic categories, with the use of machine learning schemes. The results proved that stylistic features may have subject-revealing power and significantly enhance topic classification.

Mikros & Carayannis [20] used exclusively non lexical features in order to classify 1200 texts in four topic categories. The feature set used was based exclusively on stylometric variables such as lexical "richness" and various sentence and word level measures including specific sociolinguistic attributes. Overall topic classification accuracy reached 81%, providing evidence that these features carry content information.

Mikros [19] used DFA and compared various features, lexical and non lexical in topic categorization using a corpus of 900 newswire articles. Each variable's contribution was measured using Wilks'L and the results showed that stylometric variables like the Average Word Length and frequency of the Punctuation Marks were among the most influential variables in the analysis.

Tambouratzis, et al. [28] carried out style-based text classification tests for the Greek language, focusing on polysemy and grammatically equivalent word forms. They counted morphological, as well as structural features of the texts and deployed cluster analysis on three categories (Fiction, History, Politics), with high accuracy results.

Another study was effected by Michos, et al. [18], focusing on functional rather than literary style. In their automatic text categorization experiments, they used syntactic and verbal identifiers, such as adjective/noun and adverb/verb ratios, and studied the positive/negative effects of linguistic features in real-life texts.

Overall, more and more text categorization studies seem to focus on the discriminatory role of stylistic attributes within various topics, producing interesting results, that should be further explored.

2.4 The effects of topic in authorship attribution

The increasing number of topic-controlled corpora used in authorship attribution studies, described in 2.1, reveals an awareness of topic bias in author discrimination accuracy. However, a small number of studies that have directly researched this issue report contradicting results.

Corney [5] investigated the effect of e-mail topics in authorship classification. The corpus used in this study consisted of e-mails written by a small closed group of authors on a specific set of topics. To measure the topic effect, classifier models were built for each of these authors, using the e-mails of one of the topics. Other topics' e-mails were then used as the test data for the classifier learning models from the original topic. The obtained results showed that authorship attribution accuracy was unaffected by e-mail topic and that function words were consistently the best individual feature set independent of topic.

Madigan et al. [17] also underlined the need to research topic effect in authorship attribution using cross-topic corpora. In order to test the effect of topic in authorship attribution, they used a corpus of Usenet postings compiled from two users, who systematically post many messages in discussion groups of different topic. Results showed that topic interacts with authorship and the Bag of Words (BoW) representation, which was the most successful feature set in data sets of multitopic authorship attribution, performed poorly on this experiment.

De Vel et al. [6] used a corpus of 1259 Usenet postings in four topics written by four authors. Results showed that inter- and intra- topic authorship attribution is possible but authorship categorization precision is not stable across all authors. In specific cases, the categorization obtained was biased towards the e-mail document topic content rather than on its author.

Finn & Kushmerick [7] investigated genre classification corpora controlled for topic. They evaluated their classifiers using two text collections. The first experiment calculated the accuracy of the classifier in a single subject domain. The second experiment measured the classifier accuracy, when trained on one subject domain, but tested on another. This specific task was used as a measure of the performance of a genre classifier across multiple subject domains and gave an indication of the classifier's ability to generalize to new domains. The results showed that topic and genre besides their theoretical distinctiveness, in practice, they partially overlap. The standard stylometric features used in this study were able to discriminate genres but the models built were partially topic dependent.

3. Methodology

3.1 The topic-controlled authorship corpus

In order to study the topic effect in authorship attribution we compiled a small-scale corpus consisting of 200 newspaper articles written by two authors (Dimitris Maronitis, who is actually a scholar and Pantelis Boukalas, who is a philologist) for

the electronic editions of two major Greek newspapers, TO VIMA and KATHIMERINI, during the period 1997-2006. All articles were downloaded from the websites of the newspapers in question.

We collected articles from two topic categories, Culture and Politics, keeping in mind the authors' similar writing style. A special criterion for the selection of the specific articles was the authors' natural register, as well as their overlapping in terms of writing within the same genre, but also each one's similar style when writing for different topics. Another interesting aspect of the texts is that their authors mix various topics while analysing certain political aspects of these topics and vice versa. For instance, they may write about a political subject and use historic or cultural examples to illustrate their point, or they may write about a cultural event or review a book and discuss them in a political context. The latter case is more frequent in the articles written by Pantelis Boukalas. Moreover, each text per author comes from the same column and section in each newspaper, as included in the newspaper supplements consisting of essays and articles regarding culture, history, science, social and political issues etc. In principle, this means that such texts undergo some low-level post-editing, as opposed to editorial or reportage articles, which are subject to a stricter editing, so that they conform to the overall style of the newspaper. Therefore, the style of the specific authors is more personal and independent of outer influences. Similar texts have also been used in a corpus compiled by Stamatatos [27] in his study on ensemble-based author identification.

The corpus size distribution per author and topic is shown in the table below (Table 1):

Table 1: Distribution of words and texts across Topic and Author categories.

	Topics					
	Culture		**Politics**		***Total***	
Authors	Texts	Words	Texts	Words	Texts	Words
Boukalas	50	41,107	50	21,561	***100***	***62,668***
Maronitis	50	30,645	50	28,850	***100***	***59,495***
Total	***100***	***71,752***	***100***	***50,411***	***200***	***122,163***

3.2 Stylometric variables

We used different categories of stylometric variables all of which are in theory topic-neutral:

1) Lexical "richness" variables: Yule's K [Yule's K], Standardized Type Token Ratio [stTTR], Lexical Density (ratio of content to function words) [LexDens], Percentage of hapax-legomena [HapaxL], Percentage of dis-legomena [DisL], Ratio of Dis- to Hapax legomena [Dis_Hap], Relative Entropy [RelEntr], Percentage of numbers in the text [Numbers] - 8 variables

2) Sentence level measures: Average length of sentences measured in words [SL], Standard deviation of sentence length per text [SLstdev] - 2 variables

3) 10 most Frequent Function Words of Modern Greek – 10 variables

4) Word level measures: Average word length per text measured in letters [AWL], Standard deviation of word length per text [AWLstdev], Word length distribution containing frequency of 1 letter word to frequency of 14 letters word [1LW, 2LW… 14LW), - 16 variables

5) Character level measures: Frequency of the letters normalized to 1000 word fixed text length – 32 variables

4. RESULTS

4.1 Classification accuracy in author and topic discrimination

In order to test the discriminatory power of the above-mentioned features, we used DFA, a well documented classification function, which has been used extensively in authorship attribution research (e.g. [3], [25], [29], [22]).

DFA involves deriving a variate, the linear combination of two (or more) independent variables that will discriminate best between a priori defined groups. Discrimination is achieved by setting the variate's weight for each variable to maximize the between-group variance, relative to the within-group variance [9].

If the dependent variables have more than two categories, DFA will calculate k-1 discriminant functions, where k is the number of categories. Each function allows us to compute discriminant scores for each case for each category, by applying the following equation:

$$\mathbf{D_{jk} = a + W_1X_{1k} + W_2X_{2k} + ... + W_nX_n}$$

where

D_{jk}= Discriminant score of discriminant function j for object k

a= intercept

W_i= Discriminant weight for the independent variable i

X_{ik}= Independent variable i for object k

For the validation of the DFA results, we used the U-method, a cross-validation procedure based on the "leave-one-out" principle [10]. Using this method, the discriminant function is fitted to repeatedly drawn samples of the original sample. This procedure estimates k-1 samples, eliminating one observation at a time from a sample of k cases.

We first applied DFA using Author as dependent variable and obtained the cross-validated classification results. In the second phase, we applied DFA again using the same stylometric variables, but we used Topic as dependent variable. Both DFA's were computed using the stepwise method. The confusion matrix of both DFAs is presented below (Table 2):

Table 2: Cross-validated classification results in Author and Topic categorization.

Overall Author classification accuracy = 96%	***Predicted author***	
Author	Boukalas (%)	Maronitis (%)
Boukalas	97	3
Maronitis	5	95
Overall Topic classification accuracy = 79.5%	***Predicted topic***	

Topic	Culture (%)	Politics (%)
Culture	76	24
Politics	17	83

The authorship attribution achieved an overall 96% accuracy, showing that the selected feature set was indeed useful in capturing authorship information. However, the topic categorization accuracy was also very high (79.5%), especially if we consider that we used only stylometric variables and no content words at all. This result indicates that the features used, at least some of them, correlate with topic information and are not topic-neutral.

4.2 Testing the topic-neutral hypothesis of common stylometric variables

In order to explore further which features are truly topic independent, we performed a series of two-way ANOVA with dependent variable each time a specific stylometric variable and factors, the Author and the Topic of the text. Two-way ANOVA can reveal not only the main effects of Author and Topic in the dependent variable, but also the interaction effect between them. We examined the distribution of all the variables using Kolmogorov-Smirnov test and we found 30 variables that were not normally distributed. In these variables we used additionally the non-parametric Mann-Whitney U test in order to validate the p values of the ANOVA. In all these cases ANOVA results were confirmed although the normality assumption was violated.

The ANOVA results are reported in the following tables organized by feature sets. Grey cells are statistically significant ($p<0.05$):

Table 3: ANOVA significance in main and interaction effects with dependent variables Lexical "richness" features.

Lexical "richness" variables	*Author*	*Topic*	*Author~Topic*
Yule's K	0.00	0.02	0.08
stTTR	0.00	0.2	0.00
LexDens	0.00	0.31	0.21
DisL	0.07	0.00	0.23
RelEntr	0.57	0.00	0.05
HapaxL	0.7	0.00	0.57
Dis_Hap	0.12	0.27	0.4
Numbers	0.67	0.01	0.00

The lexical "richness" variables displayed above (Table 3), exhibit considerable variation regarding their correlation with topic. Lexical Density seems to be the only variable that discriminates authorship exclusively. All the others have some interaction with topic. In particular, four of them, appear to discriminate only topic (Hapax Legomena, Dis Legomena, Relative Entropy, Numbers). Yule's K, one of the most widely used stylometric variables in authorship attribution, relates both to authorship and topic. Standardized TTR discriminates authors, but at the same time exhibits author~topic interaction effect.

Table 4: ANOVA significance in main and interaction effects with dependent variables Sentence level features.

Sentence level variables	*Author*	*Topic*	*Author~Topic*
SL	0.00	0.84	0.03
SLstdev	0.00	0.92	0.04

The two sentence level variables have similar behavior as can be seen in the above table (Table 4). They discriminate authors and not topics, but they present statistical significance in author~topic interaction, as can be seen in Figure 1:

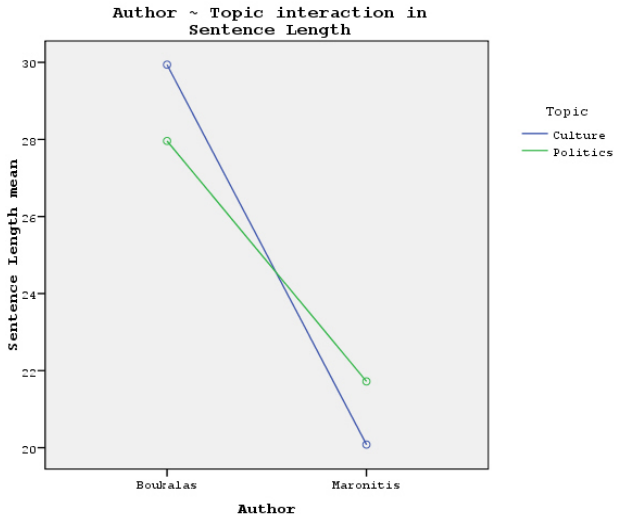

Figure 1: Author ~ Topic interaction in Sentence Length.

Sentence length mean is not statistically different between the two topics. However, Boukalas is using statistically significant larger sentences than Maronitis in Culture texts and smaller sentences than Maronitis in Politics texts. This kind of interaction reveals that each author manipulates this variable in a different way, according to the topic of the text. In general, an author~topic statistically significant interaction in a stylometric variable falsifies its topic-neutral character.

Table 5: ANOVA significance in main and interaction effects with dependent variables Frequent Function Words features. In parenthesis a rough translation in English.

Frequent Function Words variables	*Author*	*Topic*	*Author~Topic*
kai (and)	0.00	0.61	0.13
na (to)	0.00	0.00	0.64
tha (will)	0.00	0.01	0.25
den (don't)	0.00	0.00	0.06
oti (that)	0.00	0.00	0.03
apo (from)	0.06	0.43	0.83
pou (where ~ who/m)	0.12	0.97	0.63
gia (for)	0.37	0.09	0.24
se (in)	0.5	0.45	0.93
me (with)	0.73	0.05	0.68

From the ten most frequent function words of Modern Greek displayed in the above table (Table 5), half of them do not have any discriminatory power over author or topic (apo, pou, gia, se, me). From the remaining five, only "kai" discriminates exclusively authorship, while the others distinguish both author and topic. These results show that, although function words are indeed semantically free, they do however contribute indirectly to the meaning of the text. This is happening probably through syntax and discourse level, since many function words construct phrase complexity and build cohesion patterns, which can indirectly be linked with topic information.

Table 6: ANOVA significance in main and interaction effects with Word level features as dependent variables.

Word level variables	*Author*	*Topic*	*Author~Topic*
AWL	0.00	0.00	0.38
2LW	0.00	0.6	0.93
7LW	0.00	0.00	0.51
8LW	0.00	0.03	0.72
9LW	0.00	0.05	0.38
10LW	0.00	0.5	0.86
11LW	0.00	0.08	0.97
12LW	0.00	0.18	0.72
14LW	0.11	0.00	0.34
3LW	0.13	0.07	0.77
4LW	0.22	0.24	0.14
AWLstdev	0.36	0.00	0.31
1LW	0.4	0.23	0.71
13LW	0.55	0.04	0.44
6LW	0.75	0.03	0.13
5LW	0.9	0.82	0.5

The word level variables discriminate both author and topic, as shown in the above table (Table 6). Authorship is exclusively distinguished by 2, 9, 10, 11, 12 letters words and topic by 6, 13, 14 letters words plus Average Word Length standard deviation. Discrimination of both author and topic is observed by Average Word Length and 7 and 8 letters words. The influence of topic on word level variables is important. A possible explanation could be that long words tend to be terms with specific topic meaning. Furthermore, average word length standard deviation is higher in texts with many long words, which make this variable topic-dependent.

Table 7: ANOVA significance in main and interaction effects with Character level features as dependent variables. In parentheses, the character in Modern Greek.

Character level variables	*Author*	*Topic*	*Author~Topic*
gh (γ)	0.00	0.00	0.00
f (φ)	0.00	0.00	0.00
s (σ)	0.00	0.00	0.05
k (κ)	0.00	0.00	0.09
dh (δ)	0.00	0.00	0.16
u (υ)	0.00	0.06	0.14
n (ν)	0.00	0.1	0.73
i_st (ί)	0.00	0.14	0.63
r (ρ)	0.00	0.2	0.13
h (η)	0.00	0.3	0.15
sfin (ς)	0.00	0.41	0.98
e (ε)	0.00	0.5	0.26
ks (ξ)	0.00	0.62	0.9
h_st (ή)	0.00	0.69	0.26
th (θ)	0.00	0.75	0.46
m (μ)	0.00	0.84	0.03
a (α)	0.00	0.9	0.87
bh (β)	0.00	0.99	0.08
l (λ)	0.02	0.07	0.67
omg (ω)	0.03	0.34	0.34
e_st (έ)	0.04	0.18	0.05
a_s (ά)	0.07	0.19	0.38
x (χ)	0.07	0.9	0.00
t (τ)	0.25	0.04	0.77
ps (ψ)	0.31	0.02	0.51
u_st (ύ)	0.33	0.12	0.02
z (ζ)	0.6	0.15	0.7
o_st (ό)	0.68	0.82	0.17
i (ι)	0.78	0.02	0.07
p (π)	0.83	0.00	0.06
omg_st (ώ)	0.94	0.92	0.88
o (ο)	0.95	0.18	0.55

From the above table (Table 7), we conclude that letter frequencies are not topic-neutral feature. From the 32 measured characters, 12 correlate with topic either as a main effect (gh, f, s, k, dh, t, ps, i, p) or as interaction with the Author variable (m, x, u_st). This result is particular interesting since the letters, which present statistically significant main effects in topic, are among the most frequent consonants in Modern Greek. A partial explanation of this could be found if we inspect more closely the distribution of the specific consonants at the word level. Mikros et al. [21], found that dh, p, k, t, gh, f, s are the most frequent letters in the beginning of a word. This could reveal a covert relation to the topic of a text, since specific topics contain terms, which begin with specific characters. If this is true, then letter frequencies should not be used as topic-neutral authorship attribution variables, since different topics will change dynamically the correlation with specific characters. As a result, each authorship attribution corpus will present different character~topic correlations in an unpredictable way.

We repeated author and topic classification with 22 features that have been found to be really topic-neutral (that is, features that present statistically significant main effect to Author). The confusion matrix of both DFA's is presented below:

Table 8: Cross-validated classification results in Author and Topic categorization using only topic-neutral features.

Overall Author classification accuracy = 93%	***Predicted author***	
Author	Boukalas (%)	Maronitis (%)
Boukalas	93	7
Maronitis	7	93
Overall Topic classification accuracy = 49%	***Predicted topic***	
Topic	Culture (%)	Politics (%)
Culture	50	50
Politics	52	48

The results reported in the above table (Table 8), show that authorship attribution accuracy remained high (93%), while topic categorization dropped to baseline percentage (49%). Although accuracy in authorship attribution dropped 3% relating to the stepwise DFA reported in Table 2, the feature set that obtained this attribution is far more robust and can be used reliably in measuring author's style, excluding text topic influence.

5. Conclusions and future work

This study investigated the topic-neutral character of some widely used stylometric variables in authorship attribution studies. In order to research the influence of topic in author discrimination, we compiled a balanced corpus of two authors, whose articles are equally divided in two distinctive topics, culture and politics. In this corpus, we measured five feature sets that in theory are topic independent. Using DFA, we showed that the same feature set could provide author and topic classification with high accuracy. A more detailed study, using a series of two-way ANOVA, revealed that many stylometric variables are actually discriminating topic rather than author. Among them, we found Frequent Function Words, specific characters, word lengths, and commonly used lexical "richness" measures, such as Yule's K. The main conclusion is that, when we apply these stylometric variables for authorship attribution to multitopic corpora, we should be extremely cautious. Authorship attribution could become a by-product of the correlation of authors with specific topics. Although this could be a useful parameter, when the set of possible authors is large, or have specific aims [17], it should be avoided in authorship attribution problems with a limited number of authors, where the analysis is focused in identifying the real person behind a text. The reported results are based on a limited corpus in both author and topic categories but they are indicative of the complex interaction between an author's style and the text topic he writes.

Future research will be directed in other languages than Greek, as well as testing other variables, such as bigrams, trigrams, Part of Speech tags, Part of Speech bigrams etc. In addition, a larger experiment is under preparation, containing more author and topic categories.

6. REFERENCES

[1] Argamon, S., Šarić, M., and Stein, S. Style mining of electronic messages for multiple author discrimination. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining*, 2003.

[2] Argiri E., *Style-based topic categorisation with the use of machine learning techniques*. MSc Dissertation, University of Athens/National Technical University, Greece, 2006.

[3] Baayen, H., van Halteren, H., Neijt, A., Tweedie, F. An experiment in authorship attribution. In *Proceedings of JADT 2002* (St. Malo 2002). 2002, 29-37.

[4] Biber D. *Variation across speech and writing*. Cambridge: Cambridge University Press, 1988.

[5] Corney, Malcolm. *Analysing E-mail Text Authorship for Forensic Purposes.* MA thesis, Queensland University of Technology, 2003.

[6] de Vel, O., Anderson, A., Corney, M., Mohay, G. Multi-Topic E-mail Authorship Attribution Forensics. In *Proceedings of ACM Conference on Computer Security - Workshop on Data Mining for Security Applications*, (November 8, 2001), Philadelphia, PA, USA.

[7] Finn, A. & Kushmerick, N. Learning to classify documents according to genre. In S. Argamon, (ed.), *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, 2003, 35-45.

[8] Graham, Neil. *Automatic detection of authorship changes within single documents.* MSc thesis, Graduate Department of Computer Science, University of Toronto, 2003.

[9] Hair, J., Anderson, R., Tatham, R., and Black, W. *Multivariate data analysis*. New Jersey: Prentice Hall, 1995.

[10] Huberty, C., Wisenbaker, J. and Smith, J. Assesing predictive accuracy in discriminant analysis. *Multivariate Behavioural Research*, 1987, 22:307-329.

[11] Karlgren, J, Cutting, D. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *Proceedings of the 15th International Conference on Computational Linguistics*, NJ: ACM Press, 1994, 1071-1075.

[12] Karlgren, J. Stylistic Experiments for Information Retrieval Experiment. In: T. Strzalkowski (ed.), *Natural Language Information Retrieval*. Norwell: Kluwer Academic Publishers, 1999, 147-166.

[13] Kessler, B., Nunberg, G., Schutze, H. Automatic detection of text genre. *Proceedings of the 35th Annual Meeting of the ACL and the 8th Meeting of the European Chapter of the ACL*. San Francisco: Morgan Kaufmann, 1997, 32-38.

[14] Koppel, M., and Schler, J. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, 2003, 69-72.

[15] Koppel, M. Akiva, N. and Dagan, I. A Corpus-Independent Feature Set for Style-Based Text Categorization. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[16] Luyckx, K. and Daelemans, W. Shallow Text Analysis and Machine Learning for Authorship Attribution. In

Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands (CLIN 2004), 2005, 149-160.

[17] Madigan, D., Genkin, A., Lewis, D., Argamon, S., Fradkin, D., and Ye, L. Author identification on the large scale. In *Proceedings of Joint Annual Meeting of the Interface and the Classification Society of North America, 2005*.

[18] Michos, S.E., Stamatatos, E., Fakotakis, N., and Kokkinakis, G. An empirical text categorizing computational model based on stylistic aspects. *Proceedings of the 8th International Conference on Tools with Artificial Intelligence*. Washington: IEEE Computer Society, 1996, 71-77.

[19] Mikros, G. Statistical approaches to automatic text categorisation in modern Greek: A pilot study for evaluating stylistic markers and statistical methods. Paper for the 6th International Conference on Greek Linguistics, in CD-ROM, 2003.

[20] Mikros, G., and Carayannis, G. Modern Greek corpus taxonomy. In: *Proceedings of the Second International Conference on Language Resources and Evaluation.* Athens: National Technical University of Athens, 2000, 129-134.

[21] Mikros, G., Hatzigeorgiu, N., and Carayannis, G. Basic quantitative characteristics of the Modern Greek language using the Hellenic National Corpus. *Journal of Quantitative Linguistics*, 2005, 12: 167-184.

[22] Mikros, G. Authorship attribution in Modern Greek newswire corpora. In Uzuner, O., Argamon, S. & Karlgren, J. (eds), *Proceedings of the SIGIR 2006 Workshop on Directions in Computational Analysis of Stylistics in Text Retrieval*, Seattle, USA, August 10, 2006, 43-47.

[23] Mosteller, F. and Wallace, D. *Applied bayesian and classical inference. The case of The Federalist Papers.* New York: Springer – Verlag, 1964.

[24] Rudman, J. The State of Authorship Attribution Studies: Some Problems and Solutions. *Computer and the Humanities, 31,* 1998, 351–365.

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

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

[27] Stamatatos, E. Ensemble-based author identification using character N-grams. In: *Proceedings of the 3rd International Workshop on Text-based Information Retrieval* (TIR'06), 2006, 41-46.

[28] Tambouratzis, G., Markantonatou, S., Hairetakis, N., and Carayannis, G. Automatic Style Categorisation of Corpora in the Greek Language. In: *Proceedings of the Second International Conference on Language Resources and Evaluation.* Athens: National Technical University of Athens, 2000, 135-140.

[29] Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G., and Tambouratzis, D. Discriminating the Registers and Styles in the Modern Greek Language – Part 2: Extending the feature Vector to Optimize Author Discrimination. *Literary & Linguistic Computing*, 2004, 19: 221-242.

Adaptation of String Matching Algorithms for Identification of Near-Duplicate Music Documents

Matthias Robine, Pierre Hanna, Pascal Ferraro and Julien Allali
LaBRI - Universite de Bordeaux 1
F-33405 Talence cedex, France
firstname.name@labri.fr

ABSTRACT

The number of copyright registrations for music documents is increasing each year. Computer-based systems may help to detect near-duplicate music documents and plagiarisms. The main part of the existing systems for the comparison of symbolic music are based on string matching algorithms and represent music as sequences of notes. Nevertheless, adaptation to the musical context raises specific problems and a direct adaptation does not lead to an accurate detection algorithm: indeed, very different sequences can represent very similar musical pieces. We are developing an improved system which mainly considers melody but takes also into account elements of music theory in order to detect musically important differences between sequences. In this paper, we present the improvements proposed by our system in the context of the near-duplicate music document detection. Several experiments with famous music copyright infringement cases are proposed. In both monophonic and polyphonic context, the system allows the detection of plagiarisms.

1. INTRODUCTION

The number of music documents available on the World Wide Web is highly increasing. Each year, over 10000 new albums of recorded music are released and over 100000 new musical pieces are registered for copyright [19]. For example, the total number of musical pieces registered in France by the French professional association SACEM, protecting artist rights, reached 250000 pieces [7] in 2004. One of the role of this organization is to help justice to take decision about plagiarism complaints. Plagiarism is the act of copying or including another author idea without proper acknowledgment. It is important to note that a plagiarism can only be decided by justice. Some famous proceedings about plagiarism happen in the last few years: Madonna and Salvatore Acquaviva in Belgium, Georges Harrison and The Chiffons in UK, *Les feuilles mortes* and *La Maritza* in France, etc. In 2004, SACEM had only verified 18000 (out of 250000) musical pieces in order to determine their originality. A complete musical analysis is performed by experts only if a complaint is lodged. Considering the important number of new music documents registered every year, it is difficult to check for possible plagiarism. For example, a SACEM member recently registered a piece that was the perfect copy of a Ravel's piece. However, it is impossible to listen and manually compare all the music document registered.

Some studies in the context of the Music Information Retrieval research area deal with computer-based techniques that may help listeners to retrieve near-duplicate music documents and may help justice to determine plagiarisms. These investigations mainly concern the open problem of the estimation of the music similarity. The notion of similarity is very difficult to define precisely and the music similarity remains one of the most complex problem in the field of the music information retrieval. This notion may strongly depend on the musical culture, on personal opinion, on mood, etc.

From a computational point of view, evaluating the similarities consists of computing a similarity measure between a pair of musical segments. Several algorithms have been proposed for achieving such a task between audio signals. But the main of these approaches are based on timbre similarity, mainly evaluated with statistics on low-level audio features. For example, Music Browser (Sony CSL, Paris) computes a similarity measure according to Gaussian models of cepstrum coefficients [13]. However, since this information about timbre is not relevant for the copyright protection of music documents, SACEM considers musical elements such as melody, harmony or rhythm. Therefore, computer-based systems should be able to study these musical elements. Then, two problematics are raising: the extraction of musical elements from audio signals in order to define symbolic data, and comparing these data.

In this paper, we present new techniques based on edit alignment algorithms. In Section 2, we present some of the existing string matching algorithms that have been adapted to the musical context. Then in Section 3, we describe some improvements dedicated to music documents. In Section 4, we introduce different options for estimating music similarity. We present finally in Section 5 some perspectives and remaining problems in the context of the detection of near-duplicate music documents or plagiarisms.

SIGIR '07 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection

2. MEASURING SIMILARITY BETWEEN SEQUENCES

Musical pieces can be described as sequences of elements (notes) [12]. Measuring similarity between sequences is a well-known problem in computer science which has applications in many fields such as text processing, data compression, bio-informatics [9, 15]. In this section, we treat the string matching algorithms that can be adapted to the musical context.

2.1 Musical Sequences

Several techniques for evaluating symbolic music similarities have been introduced during the last few years. Geometric algorithms consider geometric representations of melodies and compute the distance between objects. Some of these systems [20] are closely linked to the well-known piano-roll representation. Other ones represent notes by weighted points [17].

We propose here to investigate adaptations of string matching algorithms, since experiments show their accuracy and their flexibility in the musical context [8]. Such adaptation requires a representation of musical pieces as sequence. In the case of monophonic music (no more than one note is sounded at any given time), a musical piece can be associated to a sequence of integers, representing pitches of successive notes.

2.2 String Matching Algorithms

In [11], Levenshtein defines the notion of edit distance between two strings. This distance is defined as the minimum cost of all possible sequences of elementary operations (edit operations) that transform one string into the other. This distance can be computed in quadratic time $O(|S_1| \cdot |S_2|)$ and linear space using a dynamic programming algorithm [21]. A dual problem of edit distance is to compute alignment of two strings. The alignment of two strings consists in computing a mapping between the symbols of the strings. Symbols not involved in the mapping are designed as gap. The main difference between alignment and edit distance is that alignment computes a score of similarity: the highest is this score the highest is the similarity.

In many applications, two strings may not be highly similar in their entirety but may contain regions that are highly similar. In this case, the problem is to find and extract a pair of regions, one from each of the two given strings, that exhibits high similarity. This is called *local alignment* or *local similarity problem* [16]. The computation of a local similarity allows us to detect local conserved areas between both sequences. Experiments show that considering local alignment improves the quality of symbolic melodic similarity systems [8].

3. ALGORITHMIC IMPROVEMENTS FOR MUSIC DOCUMENTS

Experiments during the first Music Information Retrieval Evaluation eXchange (MIREX 2005) [6] clearly show that the accuracy of direct application of the existing string matching algorithms is limited. That is the reason why several improvements have been recently proposed which are presented in this section.

3.1 Representations of Music

Musical pieces are associated to sequences of notes. The representation of notes is therefore an important problem. Symbolic music analysis systems generally consider the information about pitch and duration [12] which are assumed to be the two main characteristics of musical notes. Several alphabets of characters and set of numbers have thus been proposed to represent these parameters [18]. The vocabulary chosen highly depends on the application. For applications like near-duplicate music document detection, some music retrieval properties are expected. For instance, since a musical piece can be transposed and played faster or slower without degrading the melody, such systems have to be transposition invariant and tempo invariant. In the monophonic context, only a few representations enables systems to be transposition and tempo invariant: representing pitches by the difference between successive pitches (*interval*) or in the case of tonal music, by the difference between the pitch and the key of the musical piece for example.

Experiments have been performed in [8] which confirm that the *interval* parameter leads to the most precise symbolic melodic similarity system. Moreover, other experiments show that taking into account the duration of notes significantly improves such systems.

3.2 Edit Operations specific to Music

Substitution is the main edit operation and mainly determines the accuracy of the music similarity algorithm. For some applications, the substitution score is assumed as constant. However, in the musical context, this assumption must be discussed [18]. It is obvious that substituting one pitch with another one has not always the same influence on the general melody. For example, substituting a C note with a G note (fifth) slightly modifies a melody in comparison with substituting a C note with a D note. As introduced by [12] the substitution score may be correlated to the consonance interval. It has to be determined according to consonance: the fifth (7 semitones) and the third major or minor (3 or 4 semitones) are the most consonant intervals in western music. Experiments show that this choice significantly improves algorithms [8].

Other improvements have been experimentally shown. For example, considering the note duration for the calculation of the insertion/deletion scores improves the quality of the similarity systems. Indeed, the insertion of a half note may disturb more significantly a melody than the insertion of a sixteenth note.

3.3 Weighting by Taking into Account Music Theory

We think that a preliminary music analysis may highlight the properties that help listeners to perceptually discriminate two musical patterns. This analysis may therefore lead to the modification of edit operations specific to music. For example, the notes located on the stronger beats in a bar can be considered as more important than the other ones

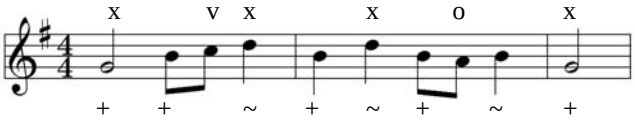

Figure 1: Analysis of a musical piece allows to identify the different functions of the notes and their placement inside the bar. Above the notes, "x" tags the importance of the note regarding the tonality limited to the tonic and the dominant tones (respectively G and D for a G Major tonality here). "v" is used to identify the passing note and "o" for a note on the weak part of the beat (which is not a passing note). Under the staff, "+" stands for the strong beats and "~" for the weak ones.

and can be weighted more than the notes placed on weak beats.

In [14], we proposed to use some notions of music theory to improve the edit-based systems. A few musical elements are analyzed and taken into account during the calculation of the edit score.

Tonality: One of the most important characteristics of the traditional western music is the tonality. The tonic is the pitch upon which all the other pitches of a piece are hierarchically centered. The scale associated to a tonality begins by the tonic. In western tonal music, the tonic and the dominant are very important. They are often used and their succession composes for example the perfect cadence that commonly ends a musical piece. In the G major or in the G minor key, tonic is the note G and dominant is the note D, like in the example of the Fig. 1. Therefore, the alignment algorithm proposed takes into account the tonic and the dominant: if the difference in semi-tones (modulo 12) between each note of the melody and the tonic equals 0 (the tonic note) or 7 (dominant), the note is assumed to be important and is therefore marked. The musical sequence alignment favours matches between these marked notes.

Passing Notes: The algorithm proposed in [14] detects the passing notes in a musical piece. A passing note is assumed as a note between two others in a constant movement (ascending or descending) which is diatonic or chromatic. There is one occurrence of a passing note in Fig. 1. The edit scores are computed according to the information about the passing notes so that the insertion or the deletion of passing notes is less penalized by the similarity system.

Strong and Weak Beats: The bar is a segment of time in a musical piece defined as a given number of beats of a given duration. In function of their position in the bar, the beats can be strong or weak with parts that are also strong or weak. We have proposed to mark the notes placed on the beats. A weight is associated to each of these notes, depending of the strength of the beat. In 4/4 time, the strong beats are the first (a weight 4 is given), and the third (weight 2) of the bars. Other beats are weighted with 1, and the other notes, which are not on the beats, are not weighted. An example of the different strengths is illustrated in Fig. 1. Our algorithm takes into account these weighted notes by favouring matches between notes on strong beats, and by not penalizing insertion or deletion of notes on the weak part of the beat.

3.4 Adaptation to Polyphony

To take into account the polyphonic nature of musical sequences, we propose to use a quotiented sequence representation. Formally, a quotiented sequence is a sequence graph with an equivalence relation defined on the set of vertices, such that the resulting quotient graph is also a sequence. A quotiented sequence can be considered as a self-similar structure represented by sequences on two different scales. A quotiented sequence can also be modelled by a tree of depth 2 where the leaves represent the support sequence and the interior nodes represent the quotient sequence. In the context of polyphonic music, notes that occur at the same time are grouped to form a quotiented sequence $Q = (S, W, \pi)$ where S is a suite of notes, W the suite of chords and π the application that maps a set of notes to each chord. Each vertex of the quotiented sequence is labelled by the pitch and the duration of each note. [10] has proposed two distances between quotiented sequences based on the computation of an optimal suite of edit operations that preserves equivalence relations on sequence vertices.

Furthermore, as previously explained, since a near-duplicate musical piece can be transposed (one or several times) without degrading the melody, algorithms for detecting near-duplicate music have to be transposition invariant. Thus, [2] proposes an original dynamic programming algorithm that allows edit based algorithms to take into account successive local transpositions and to deal with transposed polyphonic music.

3.5 System for Detecting Near-Duplicate Music Documents

According to the improvements presented in this section, we developed an edit-distance based algorithm for estimating similarity between symbolic melodic fragments. It allows us to consider a musical piece (or a fragment) and compare it to a symbolic music database. The system presented computes an edit score by comparing the musical piece tested and all the pieces of the database. The more important the score is, the more similar the pieces compared are. This system have already been evaluated in the last few years. It obtains the very accurate results with MIREX 2005 training database [8]. It also participated to the MIREX 2006 contest and obtained the best results in the monophonic context. Differences with other edit-distance based algorithms show that the optimizations proposed, specific to the musical context, permit to significantly improve such algorithms.

4. MUSIC SIMILARITY

In this section, we propose to illustrate with examples the different ways for automatically evaluating the musical similarity between musical pieces. We consider some famous examples of plagiarisms in order to show that a computer-based method is able to automatically detect near-duplicate

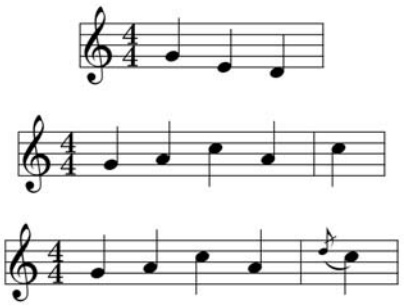

Figure 2: Short musical motifs composing the structure of the two songs *My Sweet Lord* (G. Harrison) and *He's So Fine* (R. Mack): motif A (top), motif B (middle) and motif C (bottom).

music documents. Two different approaches are investigated with systems considering melody and harmony.

4.1 Melodic Similarity

Two of the main characteristics of western music are rhythm and melody. Symbolic musical pieces are here represented by sequences of notes (see Section 2). The presented tests concern music copyright infringement cases in the United States in the last few years [5].

One of the most famous proceedings about music plagiarism concern George Harrison and his song *My Sweet Lord* that was released in 1970 on the album *All Things Must Pass* [1]. He was suspected for plagiarism of the song *He's So Fine* composed in 1963 by Ronald Mack and performed by The Chiffons. Although Harrison explained that he did not knowingly appropriate the melody of this song, the court concluded in 1976 that he had – maybe unconsciously – copied the melody of *He's So Fine*.

In order to take its decision, the court looked at the structure of the two songs. Fig. 3 shows two fragments of each of these songs. *He's So Fine* is composed of four variations of a short musical motif (motif A, Fig. 2), followed by four variations of motif B (Fig. 2). The second use of the motif B series includes a unique grace note, illustrated in motif C (Fig. 2). *My Sweet Lord* has a very similar structure in that it is composed of four variations of motif A, followed by three variations of motif B. The fourth variation of motif B includes the grace note illustrated in motif C.

The first experiments consider these two songs. Fig. 3 shows two excerpts of them. We note that even if the two melodies sound very similar, the excerpts of the melody are really different. The query of the system is defined as a part of the melody of the plagiarism *My Sweet Lord*. The database of musical pieces considered is the database proposed during MIREX 2006, *i.e.* the UK subset of the RISM A/II collection (about 15,000 incipits). The RISM A/II (International inventory of musical sources) collection is composed of one half-million notated real world compositions. The incipits are symbolically encoded music. They are monophonic and contain between 10 and 40 notes. The database also contains the monophonic melodies of *My Sweet Lord* and *He's So Fine*.

The first query corresponds to the structure considered by the court, *i.e.* repetitions of motifs illustrated by Fig 2. The second query is the excerpt of *My Sweet Lord* associated to three repetitions of motif A. The third query is the excerpt associated to the three repetitions of motif B then one motif C. The fourth query is the excerpt associated to motif A followed by motif B. Finally, the two last queries correspond to long excerpts of the monophonic melody of *My Sweet Lord* and *He's So Fine*. Tab. 1 shows the name of the most similar pieces found in the database with these different queries and their corresponding score. The scores associated to the three estimated most similar pieces are presented. The results obtained are the ones expected at the exception of the second query. In this case, the melody of *He's So Fine* is ranked far from the top 3 (the score obtained is 25.5). The little size of the motif A certainly justifies this error. For all the other queries, the most similar piece detected is the melody of *My Sweet Lord* (or *He's So Fine* for the last query), which only shows that the detection system is perfectly able to retrieve a piece from an exact excerpt. More interestingly, the second piece estimated as the most similar is the melody of *He's So Fine* (*My Sweet Lord* for the last query). Although the two sequences representing the two melodies are very different (see Fig. 3), the system proposed is able to detect their musical similarity. The two melodies seem to be also different from the structure composed of the motifs considered by the court (first query). Nevertheless, here again, the system succeeds in retrieving the two melodies. It is also important to note the difference between the scores of rank 2 and 3. As expected it becomes very significant (83 instead of 52 or 45) when the whole melody is considered, since the sequence of notes is longer.

Query	**rank** 1 **score** 1	**rank** 2 **score** 2	**rank** 3 **score** 3
Motif AAABBBC	Sweet Lord 79.6	So Fine 65.4	X 52.6
AAA from My Sweet Lord	Sweet Lord 44.2	X 30.9	X 29.5
BBBC from My Sweet Lord	Sweet Lord 113.3	So Fine 56.6	X 52.9
AB from My Sweet Lord	Sweet Lord 44.7	So Fine 33.3	X 29.8
Sweet Lord melody	Sweet Lord 178.9	So Fine 83.0	X 52.2
So Fine melody	So Fine 199.7	Sweet Lord 83.0	X 45.5

Table 1: Results of experiments about the detection of the near-duplicate monophonic musical pieces *My Sweet Lord* and *He's So Fine* (X indicates a piece that does not sound similar to the query).

In order to confirm the results of these first experiments, we propose to consider another monophonic database, which is composed of long musical pieces. This database groups more than 1650 various MIDI files collected on the internet. All these files are monophonic. Four other music copyright infringement cases are now considered [5]. For each of the five cases, the monophonic melody is proposed as query, and the system computes all the scores for all the pieces of the database (which contains these melodies). Tab. 2 shows the

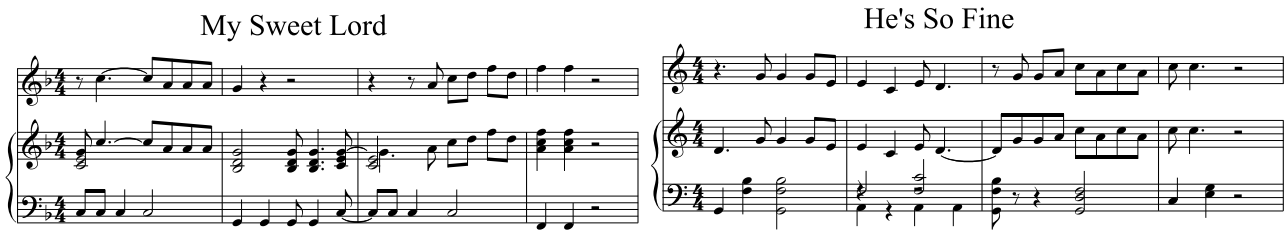

Figure 3: Manual transcriptions of excerpts (corresponding to motif A and motif B) of the two songs *My Sweet Lord* (G. Harrison) and *He's So Fine* (R. Mack).

results obtained by our system (top 3 with associated similarity scores). As expected, the first musical piece of the database estimated as the most similar is the query. The score of the rank 1 thus corresponds to the maximum score. Here, the most important result is the ranked 2 piece. Ideally, it has to correspond to the melody associated to the plagiarism established by he court. Tab. 2 shows that it is always the case, at the exception of the case *Fantasy vs Fogerty*. This error shows the limitations of the current system (see Section 5 for discussion). For all the other cases, the detection system gives the results expected. For cases like *Selle vs Gibb* or *Heim vs Universal* for example, the similarity is evaluated as important. However, the limitations of the system are also shown by the little difference between ranked 2 and ranked 3 scores for the case *Repp vs Webber*. The low score for the rank 2, corresponding to the near-duplicate piece, induces low differences between this score and the other ones obtained with the other pieces of the database. That's why more musical elements have certainly to be considered in order to reduce these differences and to make the system more robust.

We only performed a few experiments with polyphonic musical pieces. The polyphonic database considered is the *MIDI karaoke* database used during MIREX 2006, which is composed of 1000 pieces collected on the internet. The only experiment performed considers the monophonic melody of *My Sweet Lord*. The detection system compares this monophonic melody to all the polyphonic pieces contained in the MIDI karaoke database. Tab. 3 shows that *He's So Fine* has been still detected as the musical piece of the database the most similar to *My Sweet Lord*. However, in the polyphonic context, the limitations of our system are highlighted. The probability of detecting a high similarity with long polyphonic pieces is more important than with monophonic pieces, because all the notes are taken into account by our system. If the similarity score between two corresponding pieces is low in the monophonic context, the system may not correctly evaluate their similarity in the polyphonic context. For example, with *He's So Fine* as query, the system does not succeed in retrieving the corresponding polyphonic piece (*My Sweet Lord* obtains a score equals to 107.6 whereas the ranked 2 score is 141.3). At the contrary, if the similarity is more important in the monophonic context (for example *My Sweet Lord*), the system succeeds in detecting the near-duplicate polyphonic piece. Here again, the main conclusions are that the system succeeds greatly for some cases, but needs improvements. Considering other musical elements may certainly improve the system in both monophonic and polyphonic contexts.

Query	**rank** 1 **score** 1	**rank** 2 **score** 2	**rank** 3 **score** 3
R. Mack vs G. Harrison (1976)			
Sweet Lord	Sweet Lord 178.9	So Fine 83.0	X 77.5
So Fine	So Fine 199.7	Sweet Lord 83.0	X 75.3
Fantasy vs Fogerty (1994)			
Road	Road 168.9	X 87.6	Jungle 75.9
Jungle	Jungle 146.3	Road 75.9	X 75.5
Heim vs Universal (1946)			
Vagyok	Vagyok 248.6	Perhaps 123.5	X 92.8
Perhaps	Perhaps 215.5	Vagyok 123.5	X 76.8
Repp vs Webber (1997)			
Till You	Till You 135.5	Phantom 50.8	X 50.4
Phantom	Phantom 145.8	Till You 50.8	X 49.7
Selle vs Gibb (1984)			
Let It End	Let It End 192.4	How Deep 118.1	X 68.9
How Deep	How Deep 202.8	Let It End 118.1	X 83.8

Table 2: Results of experiments about the detection of the near-duplicate monophonic musical pieces for a few music copyright infringement cases.

4.2 Harmonic Similarity

Taking only the melody into account may not be sufficient to identify near-duplicate music documents. Let us take an example: a famous french case of plagiarism concerns the musical pieces *Les feuilles mortes* (internationally known as *Autumn leaves*) and *La Maritza*. As we can see on Fig. 4, even if the two pieces are perceptively very similar, a lot of notes are inserted in *La Maritza* regarding to *Les feuilles mortes*. The composer of *La Maritza* has been recognized guilty of plagiarism offense by a french court. Algorithms presented in the previous sections could strongly identify this kind of plagiarism. It is a human music expert that influenced this judgment by exposing the similarities between the two different music scores. His conviction was based on a music analysis of the scores and a look for some duplicated

Query	rank 1 score 1	rank 2 score 2	rank 3 score 3
Sweet Lord	Sweet Lord 160.3	So Fine 96.1	X 89.2
So Fine	So Fine 178.7	X 141.3	X 137.8

Table 3: Results of experiments about the detection of the near-duplicate polyphonic musical pieces *My Sweet Lord* and *He's So Fine* from monophonic melody.

motifs. In fact, he highlighted few similar sequences of notes with the same intervals used. He considered that the chord progression is the same for the refrains of the two musical pieces and that all the notes inserted in *La Maritza* could be considered as ornaments (musical flourishes that are not necessary to the overall melodic or harmonic line). Thus, even if few notes are common to the two musical piece, they are important regarding the harmony.

Therefore, we think that one possibility of improvement would be to base the comparison of two musical pieces first on their harmony. It would consist in finding the different chords that compose each piece before to perform a string matching on the sequences of these chords (on their name, as illustrated by the chord sequence on Fig. 4). All the ornaments and non-chord tone which can be added in a copied document from the original would not be considered (we can call it *melodic noise* in this context). The first step consists of extracting the sequence of the chords for a musical piece. In [3] a model for the tonality of a musical piece is proposed, and some methods to analyse the chord progression from the MIDI format are presented. Extracted chord sequences could then be compared with algorithms of string matching presented in Section 2. As these methods had been successfully evaluated in a musical context for the melody, we expect to obtain again some good results. As previously, we could improve the system by taking into account some musical considerations : the sequence may be invariant considering the tonality for example (a chord sequence C D E is similar to F G A) and the notion of consonance interval could be used as presented for the melody in Section 3. On the same way, a different level of matching may concern the key sequence of a musical piece. When the key of a piece is not constant, there are some modulations, and the musical piece can be segmented in different parts regarding the key (each part is composed with several chords). It could be done with methods proposed in [3, 4] to segment a musical piece in key sequences from a MIDI file.

We may therefore match a music document at least on three levels : one for comparing the melodic sequences, one for the chord sequences and the third for the key sequences. Let us imagine what could be the main interest of using all these levels for detecting similarities and near-duplicate documents. All the musical pieces registered in the world, the music inserted in movies, video games or websites constitute a huge music database in which a high level matching could allow to look for similarities as a filter. Only the pieces that would be similar on high levels, with a same chord progression for example, could be compared at the melodic level. It also gives a way to deal easily with polyphonic sounds reduced to a monophonic sequence of chords. Although the harmony of two similar musical pieces is generally very similar, it is not always true and this approach may complement the comparison at the melodic level. On another way, some pieces have the same chord progression without plagiarism. The matching of the chord sequences could therefore be used for looking for musical variations for example.

5. PERSPECTIVES FOR NEAR-DUPLICATE DETECTION

Existing algorithms that can be applied to detect near-duplicate music documents rely on string matching or geometric algorithms. Results obtained with such algorithms are quite good if the musical sequences are nearly the same. When studying a few music copyright infringement cases, it appears that musical sequences composed of very different notes can be musically very similar. Therefore, we have proposed some improvements specific to the musical context. Elements of musical theory have to be taken into account in order to improve the existing systems. The first experiments proposed in the previous section show that, when considering these improvements, edit-based systems are able to detect plagiarisms. Nevertheless, some limitations have been shown with some examples. Therefore, we propose some new perspectives by considering both melody, rhythm and harmony.

We have exposed several representations of a musical piece with the aim of finding similar pieces in a database. Concerning the representation of a melody in a monophonic or polyphonic context, we expect to test the impact of each factor of similarity – intervals, rhythm, harmonic function of the notes – and to evaluate how these parameters are independent and could be combined. The combination which is used for the moment is only a first step. We can also imagine to match sequences for each of these parameters independently. The system could give normalized results as score of similarity which could be used in different ways. One possibility would be to obtain a probability of plagiarism offense which can be finally confirmed by a human. A second possibility would be to test the similarity regarding a special parameter only if the precedent score regarding another parameter was over a threshold of similarity.

Furthermore, other musical rules than in [14] are needed to be implemented for considering and detecting the ornaments and non-chord tones that are less important in a musical piece to detect a near-duplicate document. We also aim at improving and evaluating our methods in the polyphonic context.

We expect to implement the hierarchical model we have presented in Section 4.2 to compare efficiently a great number of music documents using three different levels : melodic, chord and key level. We aim at finding the best method to use this model in the plagiarism domain with using the upper levels as filters in a big music database of polyphonic documents for example.

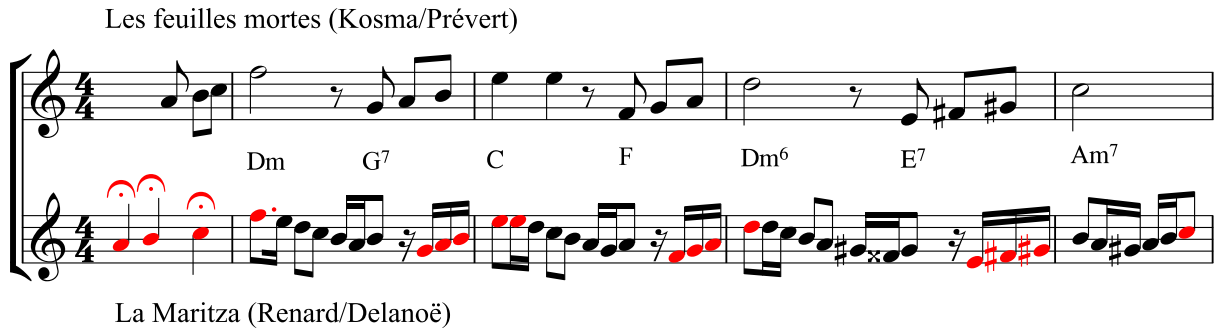

Figure 4: Manual transcriptions of excerpts of the two songs *Les feuilles mortes* and *La Maritza*. All the notes of the melody from *Les feuilles mortes* are also present in the Maritza's melody (red notes). The inserted black notes in *La Maritza* can be considered as ornaments.

6. REFERENCES

[1] *Copyright Website*. `http://www.benedict.com/Audio/Harrison/Harrison.aspx`.

[2] J. Allali, P. Hanna, P. Ferraro, and C. Iliopoulos. Local Transpositions in Alignment of Polyphonic Musical Sequences. 2007. Submitted.

[3] E. Chew. *Towards a Mathematical Model of Tonality*. PhD thesis, MIT Cambridge, MA, 2000.

[4] E. Chew. Regards on Two Regards by Messiaen: Automatic Segmentation using the Spiral Array. In *Proceedings of the Sound and Music Computing Conference (SMC)*, Paris, France, 2004.

[5] Columbia Center for New Media Teaching and Learning. *Music Plagiarism Project*. `http://ccnmtl.columbia.edu/projects/law/library/`.

[6] J. S. Downie, K. West, A. F. Ehmann, and E. Vincent. The 2005 Music Information retrieval Evaluation Exchange (MIREX 2005): Preliminary Overview. In *ISMIR*, pages 320–323, 2005.

[7] P. Emberger. Dossier SACEM : Le Livre Blanc. In *Keyboards Magazine*, volume 200, Sep 2005.

[8] P. Ferraro and P. Hanna. Optimizations of Local Edition for Evaluating Similarity Between Monophonic Musical Sequences. In *Proceedings of the 8th International Conference on Information Retrieval - RIAO 2007, Pittsburgh, PA, USA*, May 2007.

[9] D. Gusfield. *Algorithms on Strings, Trees and Sequences - Computer Science and Computational Biology*. Cambridge University Press, 1997.

[10] P. Hanna and P. Ferraro. Polyphonic Music Retrieval by Local Edition of Quotiented Sequences. In *Proceedings of the 5th International Workshop on Content-Based Multimedia Indexing (CBMI'07)*, Bordeaux, France, June 2007. To appear.

[11] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, 6:707–710, 1966.

[12] M. Mongeau and D. Sankoff. Comparison of Musical Sequences. *Computers and the Humanities*, 24(3):161–175, 1990.

[13] F. Pachet, J.-J. Aucouturier, A. La Burthe, A. Zils, and A. Beurive. The cuidado music browser : an end-to-end electronic music distribution system. *Multimedia Tools and Applications*, 2006. Special Issue on the CBMI03 Conference.

[14] M. Robine, P. Hanna, and P. Ferraro. Music similarity: Improvements of edit-based algorithms by considering music theory. *Internal report RR-1433-07, LaBRI, University of Bordeaux 1*, 2007.

[15] D. Sankoff and J. B. Kruskal, editors. *Time Wraps, Strings Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*. Addison-Wesley Publishing Company Inc, University of Montreal, Quebec, Canada, 1983.

[16] T. Smith and M. Waterman. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[17] R. Typke, R. C. Veltkamp, and F. Wiering. Searching Notated Polyphonic Music Using Transportation Distances. In *Proceedings of the ACM Multimedia Conference*, pages 128–135, New-York, USA, 2004.

[18] A. L. Uitdenbogerd. *Music Information Retrieval Technology*. PhD thesis, RMIT University, Melbourne, Victoria, Australia, July 2002.

[19] A. L. Uitdenbogerd and J. Zobel. Matching Techniques for Large Music Database. In *Proceedings of the ACM International Conference on Multimedia*, pages 56–66, Orlando, Florida, USA, 1999.

[20] E. Ukkonen, K. Lemström, and V. Mäkinen. Geometric Algorithms for Transposition Invariant Content-Based Music Retrieval. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR'03)*, pages 193–199, Baltimore, USA, October 2003.

[21] R. A. Wagner and M. J. Fisher. The String-to-String Correction Problem. *Journal of the association for computing machinery*, 21:168–173, 1974.

Intrinsic Plagiarism Analysis with Meta Learning

Benno Stein
Faculty of Media, Media Systems
Bauhaus University Weimar
99421 Weimar, Germany
benno.stein@
medien.uni-weimar.de

Sven Meyer zu Eissen
Faculty of Media, Media Systems
Bauhaus University Weimar
99421 Weimar, Germany
sven.meyer-zu-eissen@
medien.uni-weimar.de

ABSTRACT

In intrinsic plagiarism analysis we are given a document, allegedly written by a single author, and the task is to nd suf cient evidence either to accept or to reject this hypothesis. Existing research to intrinsic plagiarism analysis tries to quantify changes in the writing style by analyzing the distributions of particular style markers. This way, acceptable detection rates can be achieved if the portion of plagiarized sections is known a-priori and if the document is of a single genre. However, both assumptions may not be ful lled in practice.

In [6] Koppel and Schler propose a new approach to the authorship veri cation problem, where the task is to determine whether two texts are written by the same author. Their approach is ingenious in that it provides a means to detect relatively shallow differences in writing style while being independent of language, period, and genre. Since the approach requires two (relatively large) samples of text to be compared to each other it cannot be applied directly to the intrinsic plagiarism analysis problem.

Main contribution of our paper is the idea to address the shortcomings of existing approaches to intrinsic plagiarism analysis with the technology presented in [6]. We propose a hybrid approach that employs style marker analysis for the purpose of hypotheses generation which then are accepted or rejected by an authorship veri cation analysis. A second contribution of our paper is the evaluation of style markers for German text and their application to a real-world plagiarism case.

Keywords

intrinsic plagiarism analysis, one-class classi cation, meta learning

1. INTRODUCTION

Intrinsic plagiarism analysis is characterized as follows. We are given a document d, allegedly written by a single author, and we want to identify sections in d which stem from another author and which are not labeled as such, e. g. by proper citation.[1] Intrinsic plagiarism analysis is a one-class classi cation problem. The salient property of such classi cation problems is that information of only one class is available. This class is called the target class, all other objects are comprised in the so-called outlier class.

In the context of intrinsic plagiarism analysis all documents or document parts of the pretended author form the target class, and all documents or document parts of an arbitrary other author form the outlier class. Note that the document d is the only source to formulate a writing style model for objects in the target class, whereas the formulation of this model is impeded to the extent at which d is plagiarized. Also note that the documents in the outlier class are so abundant that neither a representative sampling nor the formulation of a writing style model for this class is possible.

D	collection of real-world documents
$d \in D$	real-world document
$\mathbf{d}$	vector space representation of document d
$\mathbf{D}$	collection of vector space representations of $d \in D$
$s \subseteq d$	section of a real-world document
n	number of sections in which d is decomposed
$\sigma : s \mapsto \sigma(s) \in \mathbf{R}$	scalar style marker or style feature of a section s
$\mathbf{s}$	style model representation of a section s = vector of style markers
m	length of a vector $\mathbf{s}$ of style markers
θ	portion of d that is plagiarized

Table 1: Notation used in this paper.

One-class classi cation problems, and hence the intrinsic plagiarism analysis problem, must be solved on the basis of examples from the target class. Tax distinguishes following methods to solve one-class classi cation problems [10]:

1. *Outlier Detection Methods.* These methods are further distinguished with respect to the detection strategy:
   (a) Methods that rely on standard classi cation and learning technology. Outliers are generated arti cially, and a standard classi cation approach is applied to separate outliers from the target class.
   (b) Modi ed methods from the eld of classi cation or regression problem solving. Instead of using the most probable feature weights $\mathbf{w}$ in a classi er, which aims at the minimization of the classi cation error given a training set, the classi er utilizes the probability of the correctness of $\mathbf{w}$.

[1] The intrinsic plagiarism analysis problem becomes harder if d is declared as a multi author document.

SIGIR '07 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identi cation, and Near-Duplicate Detection.

(c) Density methods, which directly estimate the probability distributions of features for the target class. Outliers are assumed to be uniformly distributed, and Bayes rule can be applied to separate outliers from the target class.

2. *Reconstruction Methods.* If we are given both an object's feature vector (which is a style model representation $\mathbf{s}$ here) as well as the original object (which is the document d or its VSM representation $\mathbf{d}$ here), we may be able to reconstruct $\mathbf{s}$ from d as $\alpha(d)$ as well as to measure the reconstruction error $\alpha(d) \ominus \mathbf{s}$. It is assumed that α captures the domain theory underlying the target class, and the smaller the reconstruction error is the more likely $\mathbf{s}$ belongs to the target class.

3. *Boundary Methods.* These methods avoid the estimation of the multi-dimensional density and focus on the de nition of a boundary around the set of target objects. The computation of the boundary is based on the distances between the objects in the target set.

1.1 Contributions

The contributions of this paper are as follows. Section 2 outlines existing as well as, up to now, not applied technology to solve the problem of intrinsic plagiarism detection. The two presented methods rely on a style marker analysis and can be regarded as speci c variants of what Tax terms outlier detection methods [10]. A weakness of the presented plagiarism analysis methods is that they require meta knowledge about the amount and the distribution of the plagiarized text in a document d in order to achieve acceptable values for precision and recall.

To improve the classi cation performance and to become more independent of a-priori knowledge we propose to verify the classi cation results obtained by a style marker analysis with the meta learning approach developed by Koppel and Schler [6]. Section 3 outlines their approach and its application to the intrinsic plagiarism analysis problem. Section 4 presents rst results based on both arti cial data and a real plagiarism case.

Table 1 compiles the notation that is used throughout the paper.

2. INTRINSIC PLAGIARISM ANALYSIS

Intrinsic plagiarism analysis deals with the detection of plagiarized sections within a document d, without comparing d to extraneous sources [8]. To solve this ambitious task the writing style of individual sections has to be analyzed in order to spot those sections whose style differs signi cantly from the rest. There are several subproblems that arise in this connection, including the smart decomposition of d, the identi cation of features that capture style information, the detection of stylistic anomalies or changes in style, or the construction of a corpus with positive and negative examples for plagiarism.

Writing style aspects can be quanti ed with style markers: Let $s_1, \ldots, s_n$ be a decomposition of a document d into n contiguous, non-overlapping sections. Moreover, let $\sigma_1, \ldots, \sigma_m$ denote a set of style markers, each of which assigning a real value to a section $s \subseteq d$ in order to quantify a certain style aspect of the writing. The style model representation $\mathbf{s}$ of a section s is an m-dimensional vector, comparable to an instance of the vector space model or a genre retrieval model:

$$\mathbf{s} = \begin{pmatrix} \sigma_1(s) \\ \vdots \\ \sigma_m(s) \end{pmatrix}, \; s \subseteq d$$

When a section $s^- \subset d$ is plagiarized, the assumption is that its style model representation, $\mathbf{s}^-$, differs signi cantly from other representations $\mathbf{s}^+$ that belong to non-plagiarized sections $s^+ \subset d$. Using an outlier detection method, $\mathbf{s}^-$ may be distinguished from $\mathbf{s}^+$ with acceptable reliably.

In [8] Meyer zu Eissen and Stein proposed and analyzed an outlier detection method of Type (1a). They developed a factory corpus for plagiarism analysis, and generated test corpora with several thousand positive and negative training examples. Based on these corpora different classi ers were constructed, using discriminant analysis and SVM training among others. Input for the training are the relative deviations of 10 carefully selected style markers and about 10 part-of-speech features, whereas for each section $s \in \{s_1, \ldots, s_n\}$ the vector $\mathbf{s}_\Delta$ of relative deviations of its style marker values from the document mean is computed:

$$\mathbf{s}_\Delta = \begin{pmatrix} \frac{\sigma_1(s) - \sigma_1(d)}{\sigma_1(d)} \\ \vdots \\ \frac{\sigma_m(s) - \sigma_m(d)}{\sigma_m(d)} \end{pmatrix}, \; s \subseteq d$$

Meyer zu Eissen and Stein reported precision and recall values of about 80% provided that meta knowledge about the plagiarized portion θ of d is given. In particular, they distinguished for θ the values $0.03 \cdot i, i = 1, \ldots, 6$.

Main contribution of [8] is the analysis of style markers with respect to their robustness, and the identi cation of a new class of robust style markers. In this connection, robustness pertains to the sensitivity ς of a style marker $\sigma(s)$ with respect to the length $|s|$ of a section: $\varsigma(\sigma(s), |s|)$ of a robust style marker has a small variance.

2.1 Improved Style Marker Analysis

The most severe de ciency of outlier detection methods of Type (1a) roots in their dependency on the dimensionality of $\mathbf{s}$: the number of examples must grow exponentially in the number of *relevant* features, in order to apply a machine learning approach without bad conscience. This fact is sometimes termed as curse of dimensionality . The second-worst de ciency relates to the arti ciality of the generated examples: the less we know about the stylistic impacts of plagiarism and the possible means to model these impacts the more unrepresentative the examples will be. It is in the nature of one-class classi cation problems that we have only very restricted knowledge and very few examples to model the outlier class, which are the plagiarized sections here.

By directly modeling the target objects, outlier detection methods of Type (1c) provide a way out for the mentioned problems. In this connection it is reasonable to presume the style markers in the objects of the target group being Gaussian distributed, while being uniformly distributed in the outlier group. Let S^+ denote the event that a section $s \in \{s_1, \ldots, s_n\}$ belongs to the target group (= not plagiarized); likewise, let S^- denote the event that an s belongs to the outlier group (= plagiarized). Given a suspicious document d and a single style marker σ the acceptance or rejection of the hypothesis whether a paragraph $s \subset d$ is plagiarized happens in ve steps:

1. Hypothesizing an a-priori probability, $P(S^-) = \theta$, that some section $s \subset d$ is plagiarized; $P(S^+) = 1 - P(S^-)$.

2. Depending on $P(S^+)$, decomposition of d into sections $s_1, \ldots, s_n$. Note that $P(S^+)$ provides valuable meta knowledge for the estimation of reasonable values for the section lengths $|s_i|$.

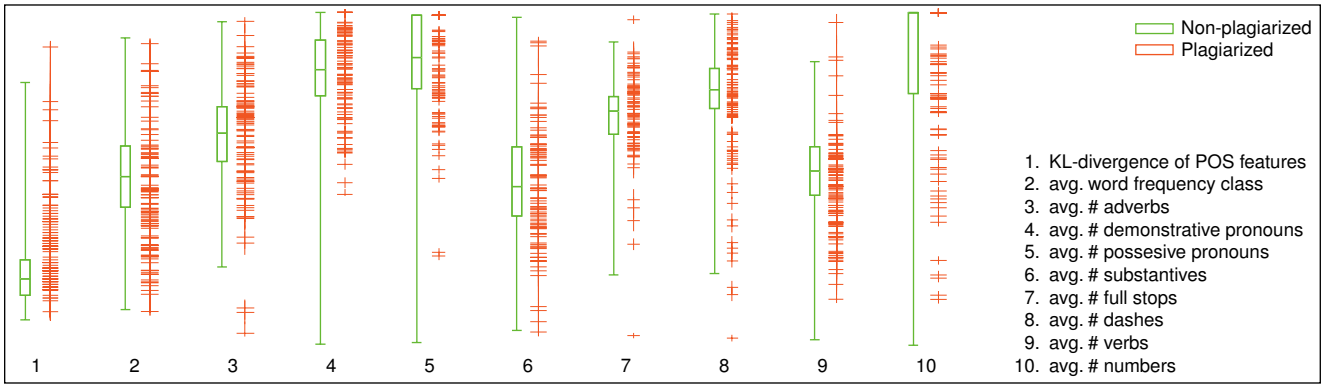

Figure 1: Distribution of 10 style markers in 16,000 non-plagiarized (green) and 1,500 plagiarized (red) sections. The sections have a length of about 400 words and result from an equidistant partitioning of 900 plagiarized documents. The plagiarized portion, θ, of a document ranges between 0.05 and 0.5.

3. Estimation of σ's expectation value and variance with respect to $s_1, \ldots, s_n$.

4. Provided an equidistant segmentation of σ's domain, computation of the conditional probabilities $P(\sigma(s) \mid S^+)$ and $P(\sigma(s) \mid S^-)$, assuming a Gaussian and a uniform distribution respectively.

5. Application of Bayes rule and determination of the maximum a-posteriori hypothesis:

$$P(S^+ \mid \sigma(s)) = \frac{P(\sigma(s) \mid S^+) \cdot P(S^+)}{P(\sigma(s))} \quad \text{and}$$

$$P(S^- \mid \sigma(s)) = \frac{P(\sigma(s) \mid S^-) \cdot P(S^-)}{P(\sigma(s))}, \quad \text{with}$$

$$P(\sigma(s)) = P(\sigma(s) \mid S^+) \cdot P(S^+) + P(\sigma(s) \mid S^-) \cdot P(S^-).$$

The above decision procedure is formulated for a single style marker. Multiple style markers $\sigma_1, \ldots, \sigma_m$ require the accounting of multiple conditional probabilities. Under the conditional independence assumption the naive Bayes approach can be applied; the accepted a-posteriori hypothesis then computes as

$$\underset{S \in \{S^+, S^-\}}{\operatorname{argmax}} \; P(S) \cdot \prod_{i=1}^{m} P(\sigma_i(s) \mid S).$$

An alternative and, dependent on the training corpus more powerful approach is the construction of a Gaussian mixture for the $\sigma_1, \ldots, \sigma_m$. The respective weights, $\mathbf{w}$, can be estimated by the linear model of a discriminant analysis, similar to the construction of a classi er when pursuing an outlier detection method of Type (1a).

The question that remains to be answered is which style markers qualify for intrinsic plagiarism analysis?

2.2 Style Markers

Quantifying the writing style of text is an active eld of research since the 1940s [11, 3]. Several style markers have been proposed to measure writer-speci c style aspects like vocabulary richness [4, 11] or text complexity and understandability [3], as well as to determine reader-speci c requirements that are necessary to understand a text, like grading levels [2, 5, 1]. These style markers have been developed to judge longer texts ranging from a few pages up to book size.

Since plagiarizers often copy sections that are shorter than a page [7], the section decomposition $\{s_1, \ldots, s_n\}$ of a document must not be too coarse, and, it is questionable which of the style markers will work for shorter sections. It should be clear that style markers that employ measures like average paragraph length are not reliable for shorter sections that consist of one or two paragraphs.

The work in [9] investigates the robustness of the vocabulary richness measures Yule's K, Honore's R, and the average word frequency class. The outcome is that only the average word frequency class can be called robust: it provides reliable results even for short sections, which can be explained with its word-based granularity. To get an idea of the usability of different style markers, Figure 1 contrasts their distribution in both original (shown green) and plagiarized (shown red) sections in a collection of 1000 documents.

3. COUPLING STYLE MARKER ANALYSIS AND META LEARNING

With the methods presented in the former section, we are able to identify possibly plagiarized sections in a document d. Let $d^+ \subseteq d$ and $d^- \subseteq d$ denote two auxiliary documents constructed from d, where d^+ is comprised of all allegedly non-plagiarized sections in d, while d^- is comprised of all allegedly plagiarized sections in d. In particular we claim that $d^+ \cup d^- = d$.

Note that, based on the decomposition $s_1, \ldots, s_n$ of d and the quality of the detection approach, $d^- \subseteq d$ may contain non-plagiarized sections, say, its precision is < 1. Likewise, d^- may not be complete, say, the recall of the plagiarized sections is < 1. Moreover, different a-priori probabilities $P(S^-)$ will result in different documents d^- to be synthesized.

Given d^+ and d^- our objective now is to nd further evidence whether d contains plagiarized sections at all. I. e., we will not try to verify whether a single section $s \subset d$ is plagiarized instead we try to answer the following relaxed decision problem:

Is d written by a single author?

For this purpose we employ the unmasking approach of Koppel and Schler, originally developed to solve the authorship veri cation problem [6]. Unmasking is a special meta learning approach, where two documents d_1 and d_2 (likewise d^+ and d^-) are incrementally reduced towards author-speci c writing style essentials. If d^+ and d^- in fact stem from different authors, unmasking is a

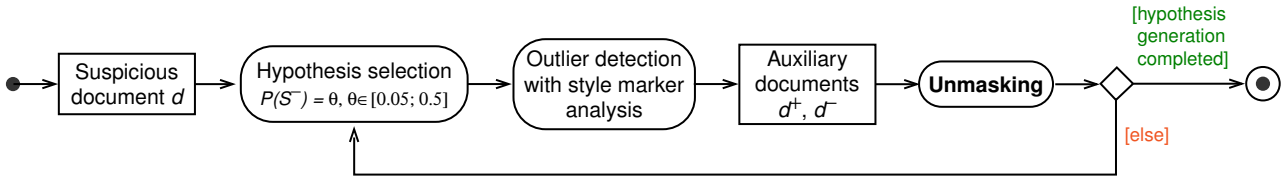

Figure 2: UML activity diagram of a new hybrid approach to intrinsic plagiarism analysis: (a) selection of a hypothesis for the plagiarized portion θ of d, (b) generation of two auxiliary documents $d^+ \subseteq d$ and $d^- \subseteq d$ with style marker analysis, (c) authorship verification with unmasking. See Figure 3 for a detailed description of the unmasking step.

powerful method to discover this fact. Figure 2 shows, in the form of a UML activity diagram, the combination of style marker analysis with subsequent unmasking.

3.1 Authorship Veri cation with Unmasking

In the authorship veri cation problem, one is given examples $d_{1_1}, \ldots, d_{1_n}$ of the writing of a single author, and one is asked to determine if a given document, d_2, were or were not written by this author.

For universal applicability we consider the examples $d_{1_1}, \ldots, d_{1_n}$ being combined into a single document d_1. The basic technology of unmasking is captured in the following procedure (cf. Figure 3):

0. *Chunking and Collection Construction.* Decomposition of d_1 and d_2 into a number of chunks. In [6] Koppel and Schler report on approximately 100 chunks of at least 500 words without breaking up paragraphs. The result of this step are two collections of chunks, D_1, D_2, generated from d_1 and d_2 respectively. The sets D_1 and D_2 are represented under a reduced vector space model, designated as $\mathbf{D}_1$ and $\mathbf{D}_2$. As an initial feature set the 250 words with the highest (relative) frequency in $D_1 \cup D_2$ are chosen.

1. *Model Fitting.* Training of a classi er that is able to separate $\mathbf{D}_1$ from $\mathbf{D}_2$. Koppel and Schler implement a ten-fold cross-validation experiment using an SVM with a linear kernel to determine the achievable accuracy. Within our analyses logistic regression is applied.

2. *Impairing.* Elimination of the most discriminative features with regard to the model obtained in Step 1, and construction of new collections $\mathbf{D}_1$, $\mathbf{D}_2$ which now contain the impaired representations of the chunks. Koppel and Schler achieved convincing results by eliminating the three most strongly-weighted positive features and most strongly-weighted negative features. Note, however, this heuristic depends on the section length which in turn depends on the length of d_1 and d_2.

3. Go to Step 1 until the feature set is suf ciently reduced. Typically about 5-10 iterations are necessary.

4. *Meta Learning.* Analyze the degradation in the quality of the model tting process: if after the last impairing step the sets $\mathbf{D}_1$, $\mathbf{D}_2$ can still be separated with a small error, assume that d_1 and d_2 stem from different authors.

Unmasking operationalizes following observation: two sets of chunks, D_1, D_2, constructed from two different documents d_1 and d_2 of the same author can be told apart easily if a vector space model (VSM) representation for the chunks in $D_1 \cup D_2$ is chosen. The VSM representation considers all words in $d_1 \cup d_2$, and hence it includes all kinds of open class and closed class word sets. If only the 250 most-frequent words are selected, a large fraction of them will be function words and stop words.[2] Among these 250 most-frequent words a small number does the major part of the discrimination job. These words may capture topical differences, differences that result from genre or purpose, and the like. By eliminating them we approach step by step the distinctive and subconscious manifestation of an author's writing style. After several iterations the remaining features are not powerful enough to discriminate two documents of the same author. By contrast, if d_1 and d_2 stem from two different authors, the remaining features will still quantify signi cant differences between the impaired representations $\mathbf{D}_1$ and $\mathbf{D}_2$ of the two chunk sets D_1 and D_2.

Remarks. At heart, unmasking is a representative of what Tax terms reconstruction methods in his taxonomy [10]. Unmasking measures the increase of a sequence of reconstruction errors, starting with a good reconstruction which then is more and more

[2]Function words and stop words are not disjunct sets: most function words in fact are stop words; however, the converse does not hold.

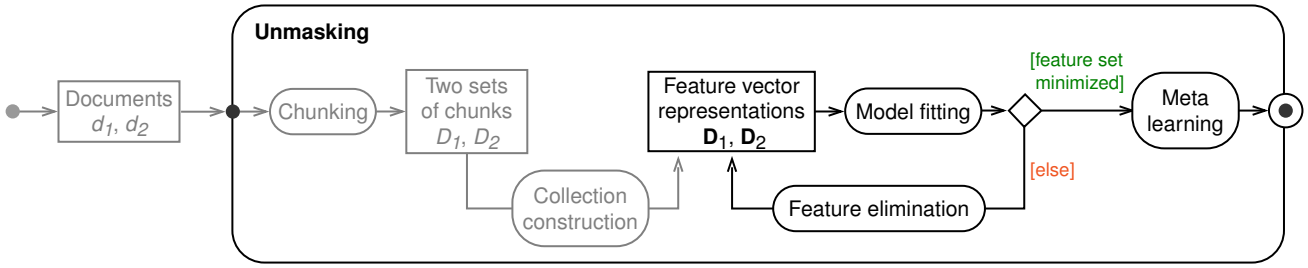

Figure 3: UML activity diagram of the unmasking technology from [6]. Input are two sufficiently large documents, d_1, d_2, from which two collections $\mathbf{D}_1$ and $\mathbf{D}_2$ are constructed. Basic idea is a meta learning analysis, which quantifies the separability of $\mathbf{D}_1$ and $\mathbf{D}_2$ when the feature representation of the chunks in $\mathbf{D}_1 \cup \mathbf{D}_2$ is increasingly impaired.

impaired. For two documents from the same author the reconstruction error develops differently compared to two documents from two different authors. In their paper Koppel and Schler present also a meta learning procedure to automatically identify the same-author curves, given a large set of unmasking experiments.

3.2 Rationale of the Hybrid Approach

Authorship veri cation and intrinsic plagiarism analysis represent two sides of the same coin. This subsection discusses the similarities and differences and gives the rationale of our hybrid approach.

In an authorship veri cation problem the interesting document d_2 with the unsettled authorship is explicitly given, and, d_2 is large enough to be analyzed with unmasking. In an intrinsic plagiarism analysis problem the sections in d for which the authorship is unsettled are unknown. In principle, unmasking could be applied to the decomposition $s_1, \ldots, s_n$ of d, taking each s_i in the role of d^- and the remaining $d \setminus \{s_i\}$ in the role of d^+. However, in most cases a single section s_i is too small to be analyzed with unmasking, and our style marker analysis serves the purpose to construct a d^- of maximum length.

In this sense the style marker analysis is a heuristic lter (or generator) function that identi es both potentially plagiarized and suf ciently long auxiliary documents d^-. The underlying search space is the set of all subsets of a document d. Let k, $k < n$, denote the minimum number of sections that must be chosen from a decomposition $s_1, \ldots, s_n$ of d in order to obtain an auxiliary documents d^- of suf cient length. With θ as the plagiarized portion of d, $k' = \lceil \theta \cdot n \rceil$ de nes an upper bound for the number of sections that can be plagiarized at all. Hence, a brute-force analysis of d had to investigate r auxiliary documents, with

$$r = \binom{n}{k} + \ldots + \binom{n}{k'}, \quad k < k'$$

An umasking analysis of r document pairs will not be tractable in most cases, which shows the nesse of the hybrid approach: the preceding style marker analysis enables us to concentrate on a very small number of auxiliary documents d^-.

A further important difference between authorship veri cation and intrinsic plagiarism analysis relates to impurity. In an authorship veri cation problem a model of the target class can be learned from the examples $d_{1_1}, \ldots, d_{1_n}$, each of which belonging de nitely to the target class. In an intrinsic plagiarism analysis problem a model of the target class has to be learned from the examples $s_1, \ldots, s_n$ (= document sections), from which only the a-priori unknown portion $1 - \theta$ belongs to the target class.

Note that, from a statistical viewpoint, the reliability of the unmasking analysis depends not only on the length of an auxiliary document d^- but also on its purity , i. e., the precision of the retrieved plagiarized sections. Like before, without a style marker analysis this problem had to be addressed by a complete but intractable brute-force search.

Related Questions. Koppel and Schler evaluate their method with twenty-one 19^{th} century English books written by ten authors, and they obtain convincing results. However, against the background of intrinsic plagiarism detection several questions arise with respect to the exibility of the unmasking approach:

1. Does unmasking work for technical and scienti c texts or is it primarily suited for novels?
2. What are minimum section lengths in the chunking step?
3. Are the initial feature set and the number of eliminated features in the impairing step independent of document lengths and section lengths?
4. Within the model tting step a model for the target class is learned. Within an intrinsic plagiarism analysis problem the model tting for the target class relies on d^+; likewise the model tting for the unknown (outlier or target) class relies on d^-. How large is the impact of precision and recall that was achieved by the style marker analysis on the model tting step?

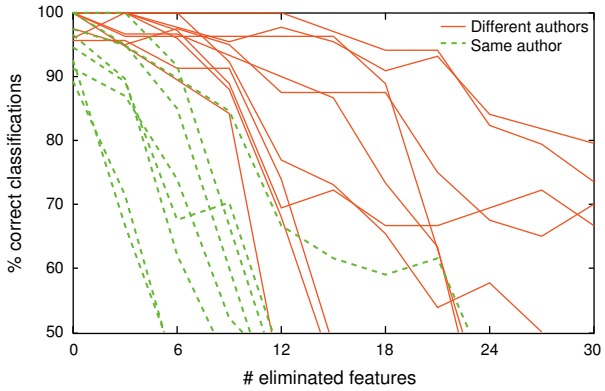

Figure 4: Authorship verification with unmasking for short documents of 4-8 pages. Each line corresponds to a comparison of two papers, where each solid red (dashed green) line results from the analysis of papers from two different authors (the same author).

We analyzed these questions within our experiments; one result is shown in Figure 4. Here, short scienti c computer science texts formed the analysis base; the average section length in the chunking step was 500 words.

4. ANALYSIS

The analysis presented here relates to documents written in German. In the next subsection an analysis of the intrinsic approach according to the outlier detection method of Type (1a) on arti cial plagiarism cases is presented, and its results are further re ned using a meta learning approach. The next but one subsection reports on a real-world plagiarism case.

4.1 Arti cial Data

We compiled a corpus of 50 scienti c documents from several domains that were downloaded from German universities. Each of these documents (written in German by a single author,) was cut down to 12-15 pages. We plagiarized the documents by hand with up to ve sections from other authors. A resulting document with k plagiarized passages served as a template document from which 2^k instance documents were generated, depending on which of the k plagiarized passages were actually included in the instance. The resulting instance documents are plagiarized at a portion $\theta \in [0.05; 0.5]$.

The rst experiment with this corpus analyzes the power of the unmasking technology, illustrated in Figure 5: Each of the red lines shows a learning curve of the plagiarized sections, d^-, against the remaining document, d^+. Likewise, a dashed green line shows a learning curve of randomly drawn sections from d^+ against the rest from d^+.

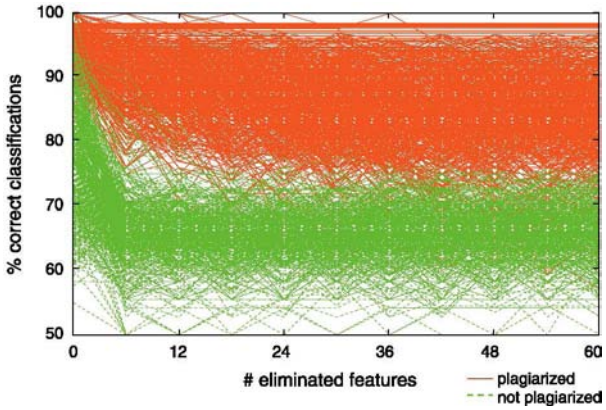

Figure 5: Unmasking applied to artificial data. Each red line shows a learning curve when separating the plagiarized parts of a document (= d^-) from the non-plagiarized part (= d^+). A dashed green line shows a learning curve when two different non-plagiarized parts of the same document are to be distinguished. The document lengths varied between 10-20 pages.

In a second experiment the intrinsic analysis as described in Section 2 was analyzed. For this purpose a classi er based on 20 part-of-speech features and 9 style markers was trained, including simple markers like average sentence length, average syllables per word, average stopword number, as well as specially crafted indexes like the Wiener Sachtextformel index, Amdahl's index, Honore's R, the Smog index, average German word frequency class, and the Kullback-Leibler divergence of the POS feature distribution. Altogether, the instance documents gave 16,000 vectors for non-plagiarized sections, and 1500 vectors for the plagiarized passages.

The classi er, based on a discriminant analysis, performed acceptably well: the precision and recall values for the non-plagiarized sections were between 80-90%, depending on the portion θ of plagiarized passages. The recall of the plagiarized sections was about 70%, having a precision of 55% given that an a priori probability of 50% for the plagiarized and non-plagiarized sections is assumed. Note that these results for an imbalanced set of feature vectors correspond to a realistic setting in which only a fraction θ of a suspicious document is plagiarized.

4.2 A Real-World Case

Given was a plagiarized postdoctoral thesis from the 1980s. The thesis was scanned, converted to plain text using OCR technology, and decomposed into 138 natural sections. The classi er that was outlined in the previous section was applied to generate a d^-, resulting in 13 suspicious sections. Three of these sections are known to be plagiarized from other textbooks from the 1980s, while the remaining 10 suspicious sections may or may not be plagiarized. Two more passages that are known to be partly plagiarized have not been detected by the classi er; an analysis has shown that the reason for missing these surrounding sections lies in the decomposition, which was too coarse for this purpose.

Figure 6 shows two learning curves for the plagiarism case. The red curve shows the classi cation rate when the 13 suspicious sections from d^- are learned against the rest of the thesis, d^+. The green dashed curve shows the classi cation rate when the original parts from d^+ are trained against 13 randomly drawn sections from d^+. The allegedly plagiarized parts can be distinguished from the original parts even when dropping the most important features. According to [6] this is a strong indication for different authors.

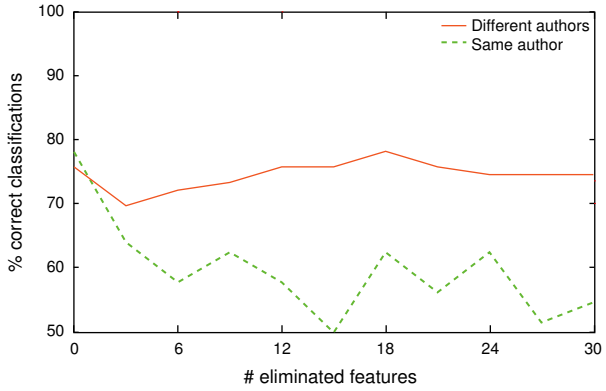

Figure 6: Analysis of a possibly plagiarized habilitation. The red line shows the learning curve when separating the 13 suspicious sections (= d^-) from the rest of the thesis (= d^+). The dashed green line shows the learning curve when 13 randomly drawn sections from d^+ are are to be distinguished from the rest of d^+.

5. REFERENCES

[1] J. Chall and E. Dale. *Readability Revisited: The new Dale-Chall Readability Formula*. Brookline Books, 1995.

[2] E. Dale and J. Chall. A formula for predicting readability. *Educ. Res. Bull.*, 27, 1948.

[3] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221 233, 1948.

[4] A. Honore. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172 177, 1979.

[5] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom. Derivation of new readability formulas (automated readability index, fog count and esch reading ease formula) for navy enlisted personnel. Research Branch Report 8U75 Millington TN: Naval Technical Training US Naval Air Station, 1975.

[6] M. Koppel and J. Schler. Authorship veri cation as a one-class classi cation problem. In *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004. ACM Press.

[7] J. Mans eld. Textbook plagiarism in psy101 general psychology: incidence and prevention. In *Proceedings of the 18th Annual Conference on Undergraduate teaching of psychology: ideas and innovations*, SUNY Farmingdale, New York, USA, 2004.

[8] S. Meyer zu Eissen and B. Stein. Intrinsic plagiarism detection. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *Proceedings of the European Conference on Information Retrieval (ECIR 2006)*, volume 3936 of *Lecture Notes in Computer Science*, pages 565 569. Springer, 2006.

[9] S. Meyer zu Eissen, B. Stein, and M. Kulig. Plagiarism Detection without Reference Collections. In R. Decker and H. Lenz, editors, *Advances in Data Analysis*, pages 359 366. Springer, 2007.

[10] D. Tax. *One-Class Classi cation*. PhD thesis, Technische Universiteit Delft, 2001.

[11] G. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.