South African Computer Journal
Number 23
July 1999

Suid-Afrikaanse Rekenaar-tydskrif
Nommer 23
Julie 1999

Computer Science and Information Systems

Rekenaarwetenskap en Inligtingstelsels

The South African Computer Journal
An official publication of the Computer Society of South Africa and the South African Institute of Computer Scientists

Die Suid-Afrikaanse Rekenaartydskrif
'n Amptelike publikasie van die Rekenaarvereniging van Suid-Afrika en die Suid-Afrikaanse Instituut vir Rekenaarwetenskaplikes

World-Wide Web: `http://www.cs.up.ac.za/sacj/`

Subscriptions

	Annual	Single copy
Southern Africa	R80.00	R40.00
Elsewhere	US$40.00	US$20.00

An additional US$15 per year is charged for airmail outside Southern Africa

to be sent to:

Computer Society of South Africa
Box 1714, Halfway House, 1685
Phone: +27 (11) 315-1319 Fax: +27 (11) 315-2276

Guest Editorial
Computer Science and Information Systems: The Future?

Philip Machanick

Department of Computer Science, University of the Witwatersrand, South Africa
philip@cs.wits.ac.za

1 Introduction

As president of the South African Institute for Computer Scientists and Information Technologists (SAICSIT), I have visited a number of campuses and companies, in an attempt at arriving at a general assessment of the state of our subjects in South Africa.

An issue which I consistently pick up is that while everyone seems to think that computer-related skills are extremely important and in short supply, our academic departments are also extremely under-resourced.

At the last Southern African Computer Lecturers Association (SACLA) conference (28-29 June, Golden Gate), I had the opportunity to discuss the problems other academics see. This editorial lists some of the problems reported at SACLA, and proposes a way forward.

2 Problems

At SACLA, I led a discussion of problems seen in our academic departments.

There was wide agreement that both Computer Science (CS) and Information Systems (IS) departments were under pressure to increase student numbers (massification), and were seen as cash cows to prop up less popular subjects. It was broadly agreed that staffing was a critical issue: too few posts for the workload, salaries way out of line with industry (half or less, as compared to the US, where an academic salary may be 80% of an industry salary). Recent graduates often make more than professors which makes it hard to persuade our students to become academics (even to do higher degrees). Attracting a recent PhD with a sense of adventure is may be possible, but attracting experienced people used to earning a salary in a strong currency is hard. IS jobs are worse than CS, as the skills required are more like those in business. Support staff salaries are an even harder issue: their skills relate even more directly to job descriptions in industry.

A problem in addressing our concerns is that we are so overworked that we don't have time for "politics": academics with no students have time on their hands, but we don't. More industry support not only with directly addressing problems but with taking on university administrations would be useful, but they too have major problems and don't have free time.

3 Solutions?

Solutions are harder to identify than problems.

The SACLA session ended with a proposal that we conduct surveys of our institutions and businesses, to find out what the problems are, as a starting point for going to university administrations, government and business.

Another idea was to attempt to find common cause with business in taking on problems they have in common with academia, including the skills shortage, the insufficient capacity of our education system, and dealing with employment equity.

One of our biggest difficulties is to free up time to deal with issues such as resource allocation within our universities. The "competition" is frequently other academics with time on their hands, since they have too few students, and therefore are in a position to spend time looking after their interests.

What is needed now is some thought about how to pull ourselves out of the mess we are in. In particular, we need strategies to exploit our strengths: our high demand among students, the high demand for the skills we produce and the ubiquitous applicability of computer technology.

Given the wide use of computers, it would seem obvious that our areas should be strongly supported by a range of role players, yet the fact that so many different groups are interested in computer technology in one way or another has tended to fragment efforts to enhance our industry and academic institutions.

Clearly, from conversations I have held, some departments are in much better shape than others. Even so, some kind of collective effort is likely to achieve more results than if we allow ourselves to be pushed around as individuals. Addressing the fragmentation of efforts seems a worthy goal in itself, to reduce duplication and contradictory goals.

I appeal to anyone who has constructive ideas on how to take our subjects forward to contact me. Let us work on building ourselves up. The economy depends on us, much more than on most other academic disciplines. It's time we made that point, and made it strongly.

SAICSIT'99

South African Institute of Computer Scientists and Information Technologists

Annual Research Conference 17-19 November 1999

Prepare for the New Millennium

Is there life after y2k?

Mount Amanzi Lodge, Hartebeespoort

near Johannesburg and Pretoria

keynote speaker: Barbara Simons, ACM President

and many other local and international speakers (academic and industry)

Call for Participation

papers in a many areas of Computer Science and Information Systems are expected

Price Waterhouse Coopers prizes: Best Paper R10000 • Best Student Paper R5000

please check the conference web site for accepted papers:

http://www.cs.wits.ac.za/~philip/SAICSIT/SAICSIT-99/

To Register

go to the conference registration web page:

http://www.cs.wits.ac.za/~philip/SAICSIT/SAICSIT-99/reservation.html

or contact **SAICSIT'99 Secretary** *for details:*

Department of Computer Science, Senate House 1137
University of the Witwatersrand
Jorissen Street
Wits, 2050
South Africa

phone (011)716-3309 fax 339-3513 (international: replace 011 by 27-11)

saicsit99-info@cs.wits.ac.za

Dates and Publication Details

early booking deadline: 14 September • on-site registration starts: 17 November 1999

workshops, tutorials 17 November 1999 • paper sessions: 18-19 November 1999

papers will appear in a special issue of South African Computer Journal

sponsors

NRF PRICEWATERHOUSECOOPERS

Research Article

Discriminators for Authorship Attribution

Hans Paijmans

Tilburg University, Tilburg, Holland,
paai@kub.nl

Abstract

This paper describes some experiments with the automated attribution of authorship. Lexical cohesion in combination with machine learning techniques are used as a method to compare texts of different authors. A methodology is described to create "stylistic fingerprints".
Keywords: *authorship attribution, machine learning, information retrieval*
Computing Review Categories: *H.2.4, H.3, I.2.6, I.2.7*

1 Introduction

The attempts at automated authorship attribution described here are a secondary result of a line of research that is aimed at the identification of information-rich passages in texts: the so-called "gravity wells of meaning" (Paijmans [30], [31]). The hypothesis underlying these "gravity wells" is that passages in texts not only differ in content or topicality, but also in the degree to which that content is emphasized: the "gravity" of the passage. Identification of such passages, then, should lead to the construction of information-rich document surrogates that in turn may serve as nuclei for information retrieval activities.

Following the example of earlier research by Hearst and Plaunt [18], who used lexical cohesion as a discriminator for topical differences we included in our experiments a number of features that quantify various measures of text cohesion (see also Morris and Hirst, [28]). While it is not yet clear whether such features can be used to measure the *gravity* of passages as meant above, they presented themselves as potential factors in the recognition of *style*. We decided therefore to apply the tools that we had collected to the problem of authorship attribution, keeping in mind that, as Burrows [8] observes, *the manner in which stylistic differentiae are interlocked enables them to register even on defective instruments.*

2 Authorship attribution

2.1 Some notes on terminology

First a few notes on terminology: it should mostly be clear from the context when we mean the author of a disputed text or the author in the sense of a scholar or scientist whom we cite for some reason or another. When doubt can arise, we will sometimes use the term 'writer' when we refer to the author of a disputed text. We will also use the term *target author* when we try to establish the authorship of a known author and the term *target text* for a corpus of positively attributed texts of that author. In the same way, we will use *control authors* and *control texts* as the complement of target author and target texts, i.e. authors that are positively identified as *not* being the target author and texts that are positively *not* written by the target author.

Methods by which attribution of texts to authors may be attempted, should be considered in the broader perspective of the analysis of style in general. In this context we may define 'style' as the various ways in which an author can allow himself freedom of expression inside the more or less fixed structure of rules and conventions that are necessary for transmitting a written message. Also, we assume that most of these are measurable; that is, we will not concern ourselves with variations that cannot be objectively identified and measured.

The three dimensions in which texts can differ, then, are those of 'genre', 'content' and style. We briefly consider each of these in turn.

2.2 Genre classification

Examples of attempts at text classification are the work of Pieper [32] for German texts and Biber [5] for English texts. We mention these authors because they both use statistical methods to identify the types or genres in the respective languages. However, they approach their typologies from opposite directions. The earlier work, by Pieper, first forms hypotheses on a number of genres in the German language, called *clines* to emphasize their gradual transition from one class into another. She then tries to identify them by a multivariate analysis of linguistic characteristics, such as the ratio of nouns or finite verbs. Biber, on the other hand, starts out by defining dimensions of difference in terms of linguistic characteristics and uses objective, statistical methods (factor analysis) to create groups of texts that are maximally different on all dimensions. He introduces the word *register* for such a group [6].

Where this research is aimed at the creation of classifi-

	Unstemmed	Stemmed
Scheme	Correct-%	Correct-%
Human-average	89	not perf
TFIDF	84	85
PPMC	64	65
GZIP	47	52
COMPRESS	23	26
...		
C4.5/10c/b/IR	68	not perf
...		

Table 1: Classification results on stemmed and unstemmed articles (from Littin, 1995).

cation systems in which complete texts may be positioned, the next step is to identify properties of text that may lead to a classification system of *parts* of the text. Again several researchers from many different disciplines have applied themselves to this task, e.g. Kieras [21], vanDijk [11] and others.

2.3 Content analysis

A rather different classification of texts is that accordig to content: aptly called "content analysis" (Krippendorf [23]). Note that the word 'content' in this context not only refers to what the text is about, but also to emotional, rhetoric or other categories. For instance, the German sociologist Ertel [13] classified texts according to dogmatism by counting words like "always", "whenever" or "never", which indicate a dogmatic state of mind in the writer, or "often", "sometimes" and "occasionally" as indicators of a more tentative state of mind.

A different approach of content classification is found in the field of Information Retrieval. Littin [24] describes an application of machine learning to text categorization. A number of schemes, including human judgment, tf.idf weighting (explained in detail below), Quinlans C4.5 (a supervised machine learning algorithm) and even the standard Unix compression utilities gzip and compress are used to classify 1600 articles from ten Usenet newsgroups (tf.idf and C4.5 are explained later). As expected, humans performed best, categorizing 89% of the articles. $tf.idf$ came in a very good second (84%). The best C4.5 variation scored 65% and the gzip compression utility scored a surprising 47%. Table 1 shows part of the results.

2.4 Comparing authors

In this section a short survey is given of work pertinent to the problems of author recognition. Two main approaches may be distinguished to the attribution of texts to authors. The first is almost as old as literary criticism itself and is based on literary or historical evidence, i.e. other evidence than that furnished by quantitative properties of the texts. Of course these literary and historic properties are of central importance for the scholar and the connaisseur, but they often are far from unambiguous. Therefore, additional proof is sought in the quantitative or statistical study of the texts under consideration; an activity that is sometimes called "stylometry". *The stylometrist looks for a unit of counting which accurately translates the "style" of the text, where we may define "style" as a set of measurable patterns which may be unique to an author* (Holmes, 1994 [19]). More detailed surveys of the state of the art may be found here and in Forsyth & Holmes, [15], [14].

The beginning of this discipline is to be found in the last century, in the work of A. de Morgan and T.C. Mendenhall. These scholars concentrated on features like sentence length and word frequency and such measures are still used in modern stylometry. Nevertheless the methodology of considering an author's writings as random samples of his/her own fixed frequency distribution of word-lengths is nowadays considered unreliable, when works of various literary genres or different eras are compared.

After the second world war the statistical approach received a tremendous boost by the development of the modern computer, not only because of the greatly improved methods to compute statistics, but in more recent years also because of the ready availability of huge machine-readable corpora and sophisticated tools for analysing texts such as stemmers, taggers, and weighted indexing systems.

In 1962 Ellegard [12] already used the frequency of function words and synonym pairs, but perhaps the most influential and certainly the most cited work is the study by Wallace and Mosteller of 1964 [29] on the *Federalist papers* (see also Francis [16]), where they proposed to attibute texts on grounds of synonym preferences of the potential authors. As synonym-pairs were few in number in the papers under consideration, in the end they selected certain function words and compared the frequencies with which these were used by the two authors.

A somewhat different approach was adopted by scholars, such as Tallentire, Baker and several others. They also worked on the assumption that every writer favours some words more than others, and that this preference can be detected in differences in the frequency profile of the word types used. The type-token ratio presented itself as a potential measure (Tallentire [36], Baker [4]), this ratio has the drawback that it is not stable over samples of different sizes as the number of tokens in increasingly large samples will show a growth-rate that is different from that of the type-dictionary. As it is generally possible to use samples with a fixed size this will rarely be a problem.

Syntactic categories are more difficult to identify than the lexical features on which most of the research mentioned above is based. Yngve, 1961 [38] proposed to use the depth of nesting of syntactic categories as a measure, but his suggestion has not been followed by later researchers. However, more recently Dutch researchers have looked into the discriminatory potential of syntactic rewrite rules for authorship attribution, with promising results (Baayen, 1995 [3]).

A rather different approach was adopted by Matthews and Merriam, using a neural network (see also [25], [26], [27], Tweedie/Singh/Holmes:1996a).

Hence a brief and not exhaustive inventory of features that have been tried as discriminators is as follows:

- general statistics, such as average word and sentence

length

- synonym preferences
- distribution of part of speech categories like nouns, verbs, or articles
- conditional clauses and phrases
- type-token ratio
- depth of nesting in sentences
- syntactic categories

Most experiments mentioned so far have been conducted in situations where the group of target authors was very small, typically one or two. As already noted by Forsyth and Holmes [15], there are very few records of attempts to apply different methods on the same corpus.

3 Stylistic fingerprints

The *modus operandi* of attributing authorship on the basis of the texts he (or she) has written, may be contrasted with that of identifying an individual on the basis of his fingerprints. We can hope for the emergence of a single pattern, lexical or otherwise, that uniquely binds every author to his texts, analoguous to the use of fingerprints in forensic proceedings. However, it is highly improbable that such a pattern can be found. A writer can deliberately change his style in an attempt to remain anonymous, to mimic a different writer, or for any other reason, but the changing of one's fingerprints is less lightly undertaken. Most authors on the subject prefer the use of patterns that are, as much as possible, beyond the conscious control of the writer. The problem with this assumption is that many features that define the style of an author certainly are under conscious control. We therefore prefer to postulate a 'cooperative attitude' of the writer in that he does not wilfully disguise his style.

On the other hand there is no law of nature, other than that of probability, that says that every man has to have unique fingerprints: it is just that the number of potential combinations of all features in a fingerprint are so great that even with five billion living individuals, the possibility that two individuals have exactly the same fingerprint may be discarded. Perhaps we may also accept after all the concept of 'stylistic fingerprints' in the sense of a combination of several standardized features, i.e. combinations of measurable textual features that identify the author of the text beyond reasonable doubt.

We want to emphasize the phrase *standardized features*. If the texts of only a few authors have to be attributed, it is feasible to search all attributed texts for some heuristic feature that may be used to attribute the unknown texts. In the case of the Federalist papers, for example, initially so-called *marker words* were selected, such as "while" and "whilst", that differentiated between the two authors. Such features are too narrow to differentiate between several authors. For "stylistic fingerprints" we would want to use features that do apply to all authors under consideration and that are, moreover, easy to measure; in short: standardized features. Therefore we will assume the following conditions:

- Most important is the availability of sufficient data. This means that we presuppose at least three texts: a text to be classified, a text or group of texts that is positively identified as being written by the author under consideration (the *target author* and the *target group*) and a text or group of texts that certainly are not written by that author (the *control group*, written by the *control authors*). For instance, in the case of the Federalist papers an as yet unattributed paper may be hypothesized to be by Hamilton. In that case Hamilton is the target author; the papers that are positively attributed to be by Hamilton are called the target group and the papers positively attributed to Madison (and to Jay) form the control group.

- Next we may assume a cooperative attitude. As already indicated we assume that the writers under consideration did not take measures to disguise their style to prevent detection or, if they did, that these measures can be recognized and separated from the features used for classification. This goes for both the target author and the control authors. In fact, this assumption would have to be relaxed in cases as those of the Federalist papers, where all authors wrote under the same "nom de plume" and may consciously have tried to mimick each others stylistic idiosyncrasies.

- Ceteris Paribus: it is also important that the texts to be compared resemble each other in as many respects as possible. If the target author was a 16th century playwright and the text to be attributed is a play too, we should select 16th century plays in the control group. There may be circumstances when this is not possible, e.g. when all positively identified texts of the target author are poems and the disputed text is a letter.

- Standardized features: the features that we use to compare the texts should be chosen such that they apply to all texts that could possibly come under consideration, both in the target group and in the control group. For authorship recognition that has general validity, we should avoid the use of ad hoc features and concentrate on general features.

So when do 'stylistic fingerprints' in fact become a feasible goal? First, the set of texts written by the authors between which to differentiate is large enough to make it a non-trivial subset of all texts written in that language and also that a text is easily classified as to its membership of this subset. The second condition would be that the discriminating features are easily recognized and quantified in the texts themselves.

4 Lexical cohesion

Looking back on the effort that has already gone into author attribution, it is perhaps amazing that no attempts have been reported at using so-called *lexical cohesion* as a measure for discriminating between authors. It is a measure that depends on the identification of recurring word tokens and, to a lesser degree, of sentences; tasks that are typically very easily performed in automated text processing.

Also, the avoidance or repetition of words that have occurred earlier in the text is a stylistic act that is performed almost consciously: most of us will recognize the repetition of the same word within too short a distance as an unaesthetic figure of speech unless the author has very good reasons to do so, for example for rhetoric emphasis.

Occurrences of a non-function word type therefore have a tendency to cluster because they are relevant to the local focus of a text, but an opposing tendency also exists in that this clustering cannot be too tight, because that would sin against an aesthetic, indeed a stylistic, principle. The repeated use of function words may be governed by their role as a placeholder for a non-function word (anaphora) to avoid unaesthetic repetitions, but it can of course also be influenced by a host of other factors. We will comment later on the differences between function words and non-function words, when used in the context of author attribution.

The subject of lexical cohesion has been addressed by several authors in the field of linguistics and computational linguistics such as Morris and Hirst[28]. In principle, the recurrence of concepts is used as a measure for the coherence of a discourse. Five classes of lexical cohesion are defined by Morris and Hirst:

1. Reiteration with identity of reference.

   1 Mary bit into an *apple*.
   2 Unfortunately the *apple* was not ripe.

2. Reiteration without identity of reference.

   1 Mary bit into an *apple*.
   2 She likes *them* very much.

3. Reiteration by means of super- or subordinate terms.

   1 Mary bit into an *apple*.
   2 She likes *fruit* very much.

4. Systematic semantic relation

   1 Mary bit into a *red* apple.
   2 She likes *green* ones too.

5. Nonsystematic semantic relation.

   1 Mary went into the *orchard*.
   2 She took an *apple*.

The first three classes depend on reiteration of the concepts involved; not necessarily of the same lexical term, but also of anaphora or direct thesaural relations such as broader terms, narrower terms and synonyms. Classes four and five depend on other relations than the repetition of (a reference to) the concept. The relations exemplified by class four, the systematic semantic relations, still may be solved relatively easily by a thesaurus or other knowledge representations; the references in class five, the nonsystematic relations, are often very difficult to solve by a formal system.

According to Morris and Hirst, lexical cohesion fulfills two roles: (**a**) that of word interpretation in context and (**b**) that of cohesion and discourse structure. They give the example of how the (narrower) meanings of the words *drink* or *wave* are defined by the context $\{gin, alcohol, sober\}$ and $\{hair, curl, comb\}$. The second function of lexical cohesion, then, is that of identifying units in discourse structure and connecting such units over gaps of several sentences, and it is this last function that is pressed in service to help in attributing texts to authors.

4.1 Text Tiling

So far we have mentioned two different approaches to the identification of such units: that described in Morris and Hirst and extended into a computational system by Kozima and Furugori [22], and secondly that applied by Hearst and Plaunt [18],[17]. There is an interesting difference between the two approaches: Morris and Hirst, and by extension Kozuma, concentrate on the semantics of the words by looking up possible related words in a thesaurus, whereas Hearst applies a measure composed from frequency-based weights of the words in a sentence to identify stretches of sentences that are connected by the occurence of identical word tokens; a technique she calls *text tiling*.

To achieve this she first computes weights for the words in the text in the following manner. First the text in the document is divided into blocks of a heuristically chosen length of 3-5 sentences. Then every word-block combination is weighted with the $tf.idf$ measure, which gives a greater weight to the word-block combination when there are more occurrences of the word in the block and fewer in the complete document (see Salton and McGill [35]). The algorithm then walks through the blocks, computing the similarity between each pair of blocks by application of e.g. the cosine formula. After the application of a smoothing algorithm to lessen the effect of local fluctuations the similarities are plotted and the valleys in the graph are pronounced to be the places where tile boundaries occur.

Hearst mentions the possibility of using her algorithm not on logical sentences but on text windows of varying sizes. This was taken up by Callan [9] in experiments in which he tried various ways of breaking up long texts for IR purposes. It was found that text windows of a fixed number of words performed better than passages that were based on textual discourse units (sentences or paragraphs). We decided to use this windowing technique as one of the parameters in our own experiments.

4.2 Measures for Lexical Cohesion

From the above, it will be gleaned that there are a number of different approaches to the computation of lexical cohesion. For our experiments we used a combination of both the chaining method of Morris and Hirst, with some refinements, and the so-called "text-tiling" of Hearst.

1. Following Morris and Hirst, we first wrote a program[1] that counted for each sentence the number of active word chains. An active word chain is the reoccurrence of a word token within a certain number of sentences or words; if two subsequent occurrences of the token are further apart than this threshold value, the chain is considered broken and a new chain starts when again two occurrences of that word within the threshold are detected. The obvious parameter was the size of the threshold itself; we added the possibility to include or exclude certain word categories, and whether the distances were measured in logical sentences (i.e. something that starts with a capital and ends with a dot) or in non-overlapping windows of a fixed number of words.

2. Hearst's method was changed in that we did not compute the similarity between consecutive blocks of text, but between consecutive sentences. The weights were computed for blocks of approx. 2000 words. Again we added variations by including or excluding word categories or by using different ways to divide the texts in sentences or in windows of a fixed number of words.

5 Machine Learning

The recognition of authors or even text genres depends on a great number of noisy and interacting features. Problems of this type have often been solved satisfactorily by machine learning techniques such as neural networks or instance based learning. Success was reported by Matthews and Merriam [27] in recognizing two authors, Fletcher and Shakespeare, in a number of plays attributed to Shakespeare and even in discriminating between passages of those authors in the same play.

Training consisted of presenting the frequencies of the function words *are, in, no, of* and *the* of the training sets to the input layer of a three-layered neural network (figure 1), putting the correct author on the output layer ((0,1) for Fletcher and (1,0) for Shakespeare. After being trained in this way the neural net was able to correctly attribute each of ten remaining plays. However, it must be noted that these words were suggested as a result of unrelated (i.e. not related to neural network) research (Horton [20]).

A different tack is the Instance Based Learning approach. This approach is based on the assumption that the simplest form of learning is memorisation. In computer terms, this means storage of the features of an object in a table, together with the identification of the object. If a new object is considered, this table is searched for either an object with the same features or for objects that most strongly resemble that object. But what does "resemble" mean in this context?

If the features of the objects in the table consist of a single real number, there is no problem; if object A has the feature value 10.3 and object B has 15.2, it is clear that object C with the value 14.1 *in this respect* is most like object B.

However, objects are generally defined by more features than one; these features are often of different classes that are difficult to compare and some or all of these features may influence each other in a number of ways. This is where statistics can play a role: after all, this discipline was developed to make sense of data. Statistical analysis generally is confirmatory: a pattern is hypothesised to exist in the data and the analysis confirms or denies its presence. Machine Learning, on the other hand, is a tool to explore the data and to report existing patterns in a way that is relatively easy to understand.

Our experiments were inspired by the availability of the Waikato Environment for Knowledge Analysis (WEKA [2]), essentially a user interface giving a standardized way to perform a number of machine learning schemes, such as C4.5 (Quinlan [33]), K* (Cleary and Trigg [10]) and the IBL variations (Aha and Kibler, [2]), among others. Preliminary experiments indicated that one of the so-called 'lazy learning' algorithms (IBL4) published by Aha [1] and the Kolmogorov (K*) algorithm published by Cleary and Trigg performed best on our data.

5.1 IBL4

The IB1, IB2, IB3 and IB4 algorithms are four variations on simple instance-based nearest-neighbour classifiers. IB1 just computes Euclidean distances between the new object and the objects already in the database and assigns to it the class of the nearest neighbour:

$$sim(x,y) = \frac{1}{\sqrt{\sum_{i \in P} Attr\_diff(x_i, y_i)}}$$

where P is the number of attributes and

$$Attr\_diff(x_i, y_i) = \begin{cases} (x_i - y_i)^2 & i \text{ is numerical} \\ x_i \neq y_i & otherwise \end{cases}$$

Every new instance becomes part of the *partial concept description* which in IB1 is the set of stored instances. As IB1 is therefore rather wasteful of storage space, an improvement was made in that only incorrectly classified instances were stored to become part of the database (IB2). This drastically reduces the storage requirements but is less noise-tolerant than IB1. Therefore IB3 was introduced, which also maintains a record of correct and incorrect classification attempts for each instance stored in the partial concept description. In this way the fitness of every instance as a classificator is determinded. This strengthens

[1] This program belongs to a suite of utilities for corpus linguistics and information retrieval written by the author. They are available at http://pi0959.kub.nl/Paai/Publiek as *Paais Text Utilities*

[2] http://www.cs.waikato.ac.nz/ ml

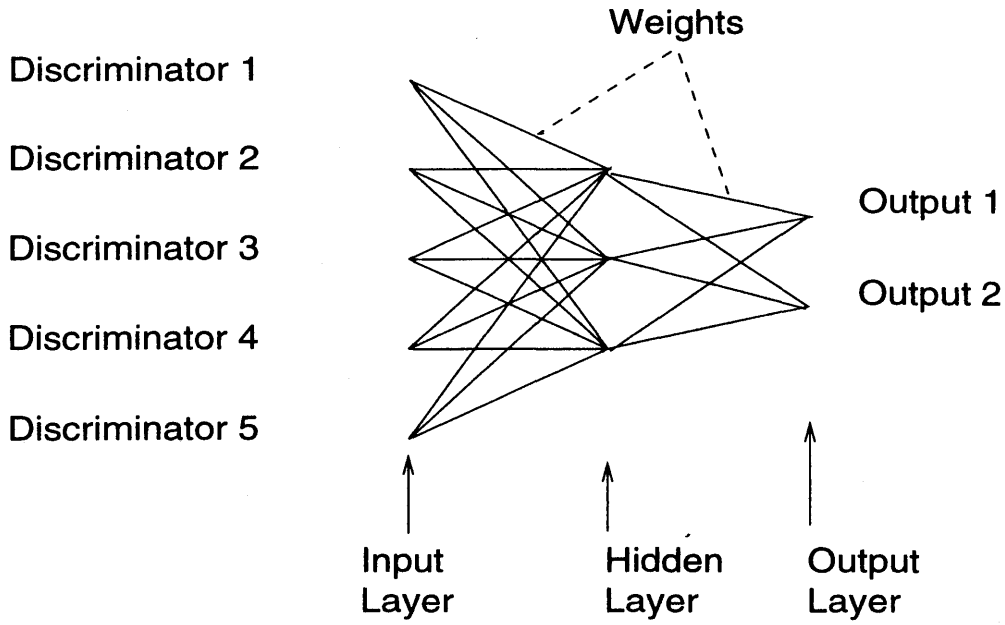

Figure 1: Topology for a stylometric neural network (Matthews and Merriam)

the noise tolerance and keeps down the storage requirements.

IB4 adds an important improvement on IB3. In the other algorithms it was assumed that the attributes carry equal weight in the predictions of a class. This is often not the case and when instances were described by many irrelevant attributes the older algorithms IB1 - IB3 had no way to detect the less relevant features. The similarity function in IB4 is defined as

$$sim(x,y,t,P) = \frac{1}{\sqrt{\sum_{i \in P} w_t * Attr\_diff(x_i,y_i)}}$$

where w_t is attribute i's weight when predictions are requested for target concept t and P is the set of attributes. When two instances are compared the weight may differ depending on the target concept. As Aha says: "For example: the similarity of a tiger and a cat is higher if the task is to predict whether they are animals than whether they are potential pets."

5.2 Kolmogorov*

The other learning algorithm that performed well on our data besides IB4 is the K* classifier. This algorithm assumes that, if two instances resemble each other, then there is a high probability of one instance transforming into the other by some accumulation of small mutations. By assigning probabilities to these mutations, a measure can be computed to calculate a distance between one instance and the other. K* incorporates all possible transformation paths in its similarity function and takes as the distance measure the sum of the probabilities of all possible transformation programs, rather than the shortest path.

Positive properties of the K* classifier (which it shares with IBL4) are the fact that attributes of different types, such as reals and symbolic values, can be dealt with within the same framework. The computation of distances between instances that have more than one attribute is straightforward.

The ultimale goal of using a classification system is that the system is trained on a dataset with known classes and that subsequently the class of new, unknown cases is decided on using the results of the training set. By contrast, when we want to study the performance of an algorithm or of selected features, we perform the second test run on data the class of which is already known. The percentage of correct predictions is then used as an indicator of the performance of the algorithm or the suitability of the features. In Weiss and Kulikowsky [37] several procedures are described to test the validity of such assumptions, of which the so-called "tenfold cross-validation" is considered to be the most stringent test of the performance. According to this procedure the data is divided into ten equal parts and that every partition is tested against the nine other partitions. The figures we will quote in the tables are averages computed over the results of tenfold cross-validation, unless explicitly stated otherwise.

6 Methodology and Experiments

Our main concern here is to establish the performance of lexical cohesion measures as authorship indicators whenever a text fragment has to be attributed to either of two target authors (i.e. authors of whom a sizable text is already available). A secondary goal was to establish the potential of lexical cohesion to uniquely identify an author between *all* authors of a particular genre or group.

We use two different ways of computing lexical cohesion. The first, straightforward procedure is that of count-

ing re-occurring words in sentences and measure their distances in number of sentences or words (chains). The second is that applied by Hearst: sentences are considered as vectors of word weights and lexical cohesion is computed as the sentence-sentence similarity. We decided to create feature databases with variations on both features in the hope that they would reinforce each other. Although we subsequently established that the lexical cohesion as expressed in the number of active chains per sentence carries most of the weight, we decided to keep both features in the database.

6.1 Preparation of the texts

We used three small corpora, C-I, C-II and C-III. The first consists of the J-category of the LOB corpus, including thirty fragments of scientific writings. A drawback of this corpus is that only two fragments are by the same author and that the fragments are very short (2000 words). The second corpus contains three books of Jane Austen and one, Wuthering Heights, by Emily Brontë. These writings were collected from the Internet. From each book a 'text' was selected consisting of ten fragments of approx. 2000 words each from the beginning of the book, except for Wuthering Heights, where we extracted two of such 'texts' (the second of which immediatly followed the first). The third corpus and *piece de resistance* was formed by thirty from the eighty-odd federalist papers[3].

First the texts were normalized and enriched by attaching the word category to each word in the text. To obtain the word categories we used the Brill-tagger [7], except for C-I as the LOB corpus already has tags attached.

The attachment of the word category has two purposes. First it reduces the problem of homographs in those cases where the same token could be reduced to two or more word categories. Second, and more importantly, it allows us to introduce the word category as a variable in our experiments.

6.2 Creation of the databases

For each experiment we created for every text a number of databases in which for every sentence the following information was collected:

1. number of active chains,
2. sentence-sentence similarity using the $atc-$weights (see below),
3. mean of the $atc-$weights of the words in that sentence,
4. for each of these three attributes, the difference between two subsequent values.

In the sections below we will give a more detailed description of the manner in which this information was collected.

[3]Details on corpora and authors are to be found in the appendix.

6.2.1 General details

The programs that extracted the information from the texts were designed with a number of general options, that applied to all programs, and options, which were used to control the parameters of specific programs. General options included:

- limiting the processing of the data to the function words or to the non-function words,
- minimum wordlength,
- the option of using all word categories in the text or only one or two categories (nouns and/or adjectives),
- whether sentences were used or windows consisting of a fixed number of words,
- the option of using n-grams instead of word tokens.

These options concern mainly various parts that can be filtered out from the text before the actual processing starts or how the text is divided into parts. The individual programs also had options to influence the processing proper:

- for word chains: the maximum length of a chain before it is considered 'broken';
- for word weights: the particular way in which the word weight was obtained (atc);
- for vector comparisons: the exact similarity measure (Jaccard, Dice or cosine).

6.2.2 Details of the atc-weight

For the tagged files we computed the tf/df or $tf.idf$ weight of each word-fragment combination. The tf or term frequency is the number of occurrences of a certain word in the fragment, and the df or document frequency is the number of fragments in which that word occurs. A popular variation is the so-called atc-weight, that was also used in the Hearst experiments. It calculates the $tf.idf$ in three steps. The first step creates the value new_tf for the term-frequency (tf) as

$$new\_tf = 0.5 + 0.5 * \frac{tf}{max\_tf}$$

where max_tf is the frequency of the term with the highest frequency in the fragment. Then the weight new_wt is calculated as

$$new\_wt = new\_tf * log\frac{N}{D_t}$$

where as before N is the number of fragments and D_t the document frequency of term t. Finally the cosine normalization is applied by

$$new\_wt = \frac{new\_wt}{\sqrt{\sum_{i=1}^{T} new\_wt_i^2}}$$

Austen	Brontë	classified as
346	189	a Austen
61	396	b Brontë
		Correct: 74%
Sense	Pride	classified as
319	166	s Sense & sensibility
248	263	p Pride & prejudice
		Correct: 58.4 %

Table 2: Confusion tables from K* (class E)

where T is the length of the document vector, i.e. the number of unique terms in the database.

For a detailed discussion of these and similar techniques see, for example, Salton and McGill [35] and Salton, [34].

6.2.3 Preparing the data

The results were organized in databases consisting of the features of every sentence for a text fragment with both the author and the fragment as potential classes.

The next step consisted of running two of the ML algorithms, IB4 and K*, that were included with WEKA, on those databases.

Before we applied the machine learning algorithms directly, we first tried to gauge the performance of both algorithms in some more detail.

This was done by concatenating two databases into a new database with randomized order, splitting the new database in two parts, training the algorithm on one part and then making it classify the second, unseen part. The performance is measured in percentages of correctly classified cases and if we have two classes, a random attribution would cause 50% of the cases to be classified correctly. This extreme would mean that the ML algorithm performed badly or that the cases were very similar or both. On the other hand, a score of 100 percent would mean that the algorithm worked very well and that the cases differed strongly between the classes.

Therefore, the precision with which the ML algorithm was able to classify the unseen part of this database was taken as a measure for the *dissimilarity* between the two original databases: a high performance in assigning the sentences (cases) to the correct texts (classes) meant that the two texts were different from each other; bad classification performance indicated a high similarity between the texts.

The hypothesis to be tested was that texts of *different* authors displayed big differences, i.e. good scores in the classification by the ML algorithms, whereas texts by the *same* author, even from different works, should be difficult for the ML program to classify and therefore approach the 50% mark.

Trying K* and IB4: it was found that our datasets were best classified by IB4, although the other algorithm also performs satisfactorily.

As an example, we provide in table 2 two confusion

	sm	fw	wnd	ng
A	1	.	.	.
B	5	.	.	.
C	5	.	20	.
D	5	1	.	.
E	5	1	20	.
F	5	2	.	.
G	5	2	20	.
H	5	.	.	3

Table 3: Variations

tables from the output of the K* algorithm. In the upper part of the table, class *a* refers to sentences from a text by Austen; class *b* to sentences from a text by Brontë. The first column displays the number of lines from Austen that are recognized as belonging to the Austen text and those wrongly assigned to Brontë, the second column gives the attributions for the Brontë lines. In this particular test almost 75% of the sentences is classified correctly, with (in this particular case) an as yet unexplained bias towards wrongly recognizing lines from Brontë as coming from Austen.

In the lower part we see the result when classification is attempted over two fragments from different books, but from the same author (Jane Austen). The number of correctly classified instances is now only 58%.

6.3 The search space

With all possibilities and variations the experiment space had grown rather large and we did not have the opportunity to exhaustively test all possible variations to find the optimal combination of databases and algorithms. In Table 3 we have aligned the variations that we tried. The first column refers to the identifiers given to the datasets. The second column, *sm*, gives the smooth-factor, i.e. the number of sentences we averaged over to smooth local fluctuations. The third column, *fw*, indicates whether the list with function-words was omitted (1) or whether the processing was limited to the function words in the text (2). The column *wnd* indicates the window size when windows of a fixed number of words were used in stead of grammatical sentences. If the last column, *ng* is filled, it refers to the length of the n-grams, if used. A second collection of files, identified by the capitals I-O was similar to the files B-H, but with a smoothing factor of ten. This group is not shown in the table.

In Table 4 the results of the experiments are displayed. Columns 1 and 5 indicate the character associated with the experiment as defined in Table 3. The columns *same* show the classification results for two fragments from the same author (Austen). The columns *other* for two fragments from different authors (Austen and Brontë). Experiments E and L (with features that were computed over text windows, using a list of function words to be ignored) display great differences in classification accuracy, and so, to a lesser degree, do the experiment pairs C-J, D-K, F-M and G-N. The asterix attached indicates differences on the 99% level as computed by the T-test.

It was found that tri-grams performs badly; the same

exp	same	other	diff	exp	same	other	diff
	no smoothing						
A	57.26	57.23	-0.02				
	smooth n=5				smooth n=10		
B	64.89	66.91	2.01	I	74.26	75.04	0.78
C*	60.95	69.11	8.16	J*	69.56	76.00	6.44
D*	66.86	72.74	5.88	K*	75.31	81.06	5.75
E*	65.27	79.86	14.58	L*	73.83	86.44	12.6
F*	66.82	71.11	4.29	M*	74.51	77.64	3.14
G*	65.52	71.87	6.35	N*	74.07	79.27	5.20
H	63.36	63.80	0.44	O	68.50	69.59	1.09

Table 4: Differences between the experimental databases classified by IB4. Asterisks indicate the .99 confidence level (T-test).

	pride	sense	nabby	wuther	awuth
pride	.	61.7	57.9	75.8	71.7
sense	59.1	.	59.4	70.1	67.8
nabby	54.4	59.9	.	73.2	70.7
awuth	75.2	72.2	74.5	.	55.6
wuther	73.9	67.4	71.9	56.8	.

Table 5: Cross-table of three fragments by Austen and two by Brontë, showing averages from a ten-fold classification test. Method: E

.

is true to a lesser degree for the features measured over the logical sentences (A, B and I).

6.3.1 Austen versus Brontë

As we have noted, the effectiveness of lexical cohesion as an author recognizer depended on the accuracy with which the ML algorithm was able to classify the sentences of different texts. It should be significantly more difficult to classify sentences from two texts by the same author than those of two texts by different authors. In table 5 we see a cross-tab of five texts, three from different books by Austen and two collections of fragments from Wuthering Heights by Brontë. The upper left and lower right segments display the percentages of correctly classified sentences of texts by the same author; upper right and lower left for different authors. Again it is clear that the program performs far better on texts by different authors than on texts by the same author.

Finally we applied the ML algorithm Ḱ directly, training on two fragments of Austen and Brontë respectively and then leaving it to the algorithm to classify the test fragment (see table 6).

6.3.2 The LOB corpus

As already indicated, we applied this method to three corpora. The tables displayed above all were taken from experiments on the Austen-Brontë corpus (C-II). In the next corpus to consider, the LOB texts, the situation is rather different in that not two texts were compared but thirty and that, moreover, the texts were much shorter (2000 versus 20,000 words per fragment). Also, only two texts (23 and 24) were by the same author. We first did a tenfold cross-validation, comparing text 23 with all other texts, including itself. As text 23 and 24 were the two texts by the same author one would expect, if lexical cohesion was a sound author discriminator, that text 23 would score lowest, followed by text 24, with a sizable gap between 24 and all other texts.

As a matter of fact this was not always the case. Over the experiment classes C-G and J-N, text 24 scored consistently low, but in every run one or two other texts would score even lower, so that 24 never would come out lowest. But when we take the averages over all experiment classes (see Figure 2) text 24 still does show up as closest to text 23. It is possible that the differences between the authors would have been more pronounced if the available fragments had been longer, but we never expected that lexical cohesion in itself would suffice to discriminate between any two authors.

	pride(A)	wuth1(B)	pride(A)	wuth(B)
	C		I	
sense(A)	0.64	0.36	0.65	0.35
nabby(A)	0.64	0.36	0.66	0.34
wuth2(B)	0.39	0.61	0.34	0.66
	D		K	
sense(A)	0.42	0.58	0.41	0.59
nabby(A)	0.51	0.49	0.52	0.48
wuth2(B)	0.28	0.72	0.23	0.77
	E		L	
sense(A)	0.51	0.49	0.58	0.42
nabby(A)	0.62	0.38	0.68	0.32
wuth2(B)	0.26	0.74	0.30	0.70
	F		M	
sense(A)	0.[illegible]	0.39	0.66	0.34
nabby(A)	0.[illegible]	0.40	0.68	0.32
wuth2(B)	0.41	0.59	0.41	0.59
	G		N	
sense(A)	0.57	0.43	0.57	0.43
nabby(A)	0.58	0.42	0.63	0.37
wuth2(B)	0.41	0.59	0.39	0.61

Table 6: Results of direct classification in K of two fragments by Austen and one by Brontë after training on Pride and Wuth1.

.

K*				
	ham	mad1	mad2	disp
ham	.	87.4	81.1	80.7
mad1	85.5	.	64.3	71.4
mad2	84.0	63.0	.	73.2
disp	80.9	72.0	73.0	.

IB4				
	ham	mad1	mad2	disp
ham	.	84.3	80.6	75.8
mad1	83.4	.	58.7	72.1
mad2	81.3	59.5	.	70.6
disp	76.9	70.9	73.3	.

Table 7: tenfold, KS, IB4 federalist.

6.3.3 The Federalist papers

The third group of papers that we used for our experiments were the Federalist papers. The initial experiments, conducted on the individual papers, showed disappointing re-

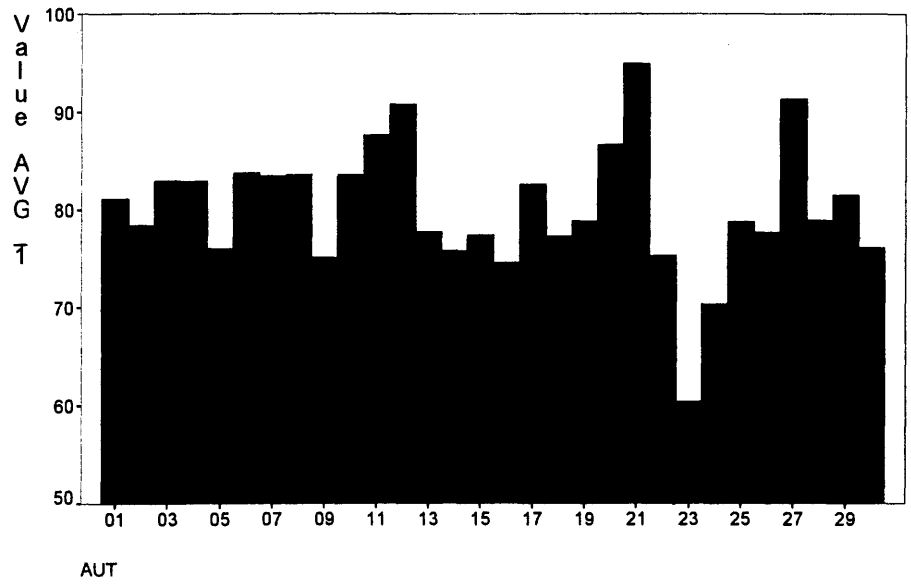

Figure 2: Average accuracy on comparing frag. 23 with all other fragments

sults: no real differences between Madisons and Hamiltons papers were visible. As these papers also were relatively short (anything between 80 and 175 lines) we decided to combine several papers of the two protagonists to larger texts each consisting of four or five original papers. We did the same with the disputed papers and proceeded to compare these groups in the same manner as used with the other two corpora.

Table 7 shows that according to our *modus operandi* and using the K* algorithm, the disputed papers lie between Hamilton and Madison: the Madison-Madison classifications score 64%, the Hamilton-Madison classifications between 87% and 81%, the Hamilton-disputed score is 80% and the Madison-disputed score is 72-73%. In other words: the disputed papers are almost as dissimilar from Hamilton as the Madison papers, but compared with the Madison groups they lie between Hamilton and Madison. The IB4 algorithm did not so well here; the disputed papers are shown to be different from Hamilton, but still nearer to Hamilton than to Madison.

Now Mosteller and Wallace conclude that the disputed papers are probably written by Madison. Our method so far indicates that they are probably *not* written by Hamilton. The discrepancy that still exists between the figures for Madison and those of the disputed group might be caused by the fact that the writer of the disputed papers consciously tried to change his natural style to conform as much as possible to the "group style" of the papers.

7 Conclusions

We have tried to show that lexical cohesion is a computationally cheap way of comparing the style of authors. As expected it does perhaps not suffice in itself to discriminate between *any* two authors, but it certainly is a candidate for inclusion in the set of standardized features necessary to obtain 'stylistic fingerprints'.

Three lines of further research suggest themselves at this point.

- First it could be useful to continue the line of experiments described here. As we have seen only a small part of the experiment space relating to lexical cohesion has been explored. For instance, the maximum chain length was rather arbitrarily fixed at six grammatical sentences, respectively artificial windows of twenty words. Also, we did not limit chains to selected word categories, such as nouns or verbs. By doing this, we could probably get a better combination of lexical cohesion features and the various procedures to obtain them from the original text files.

- Another matter is the quantity of texts that is needed to obtain enough data to train the algorithm on. The best results were obtained when the databases were typically a thousand records (sentences) or more (Austen-Brontë). The LOB-corpus and the Federalist papers generally have no more than two hundred records (sentences) per text. When we combined the papers of Hamilton and Madison in two big texts, the results did improve, but tot to the point that they could be compared to the results of Wallace and Mosteller.

- A third and rather promising line of research would consist of collecting other features of texts that also perform well as author discriminators, and combine them in a standard set, using the *modus operandi* as described above to recognize individual authors.

References

[1] D W Aha. 'A study of instance-based algorithms for supervised learning tasks: Mathematical, empirical and psychological evaluations'. Technical report, University of California, Irvine, (1990).

[2] D W Aha, D Kibler, and M K Albert. 'Instance-based learning algorithms'. *Machine Learning*, **7**:37–66, (1990).

[3] H Baayen, J van Halteren, and F J Tweedie. 'Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution'. *Literary and Linguistic computing*, **11 (3)**:121–131, (1996).

[4] J C P Baker. 'A test of authorship based on the rate at which new words enter an authors text'. *Journal of the association for Literary and Linguistic Computing*, **3**(1):36–39, (1988).

[5] D Biber. *A typology of english texts*. Linguistics, Vol. 27-1. p. 3-44. jan., 1989.

[6] D Biber. *Using Register-Diversified Corpora for General Language Studies*. Computational linguistics, Vol. 19, no. 2, pp. 219-241, June, 1993.

[7] E Brill. 'Some advances in transformation-based part of speech tagging'. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, (1994).

[8] J F Burrows. *Computers and the Study of Literature*, pp. 167–204. Oxford: Blackwell, 1992.

[9] J P Callan. 'Passage-level evidence in document retrieval'. In W B Croft and C J van Rijsbergen, eds., *SIGIR '94; Proceedings of the 17th annual international ACM-SIGIR conference on research and development in Information Retrieval*, pp. 302–310. Springer Verlag, (1994).

[10] J G Cleary and L Trigg. 'K*: an instance-based learner using an entropic distance measure'. In *Proceedings of the International Conference on Machine Learning*. Morgan Kaufmann, (1995).

[11] T V Dijk. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Lawrence Earlbaum Associates. Hillsdale, NJ, 1980.

[12] A Ellegard. *A Statistical Method for Determining Autorship: the Junius Letters, 1769-1772*. Gothenburg: the University of Gothenburg, 1962.

[13] S Ertel. 'Dogmatism: an approach to personality'. In A Deichsel and K Holzenscheck, eds., *Maschinelle Inhaltsanalyse, Materialien I*, pp. 34–44. Hamburg University, (1976).

[14] R S Forsyth and D I Holmes. 'The federalist revisited: New directions in authorship attribution'. *Literary and Linguistic computing*, **10(2)**:111–126, (1995).

[15] R S Forsyth and D I Holmes. 'Feature finding for text classification'. *Literary and Linguistic computing*, **11(4)**:163–174, (1996).

[16] I Francis. 'An exposition of a statistical approach to the federalist dispute'. In H P Vincent, ed., *The Computer and Literary Style*, pp. 38–78. Kent State University, (1966).

[17] M A Hearst. 'Cases as structured indexes for full-length documents'. In *Proceedings of the 1993 AAAI Spring Symposium on Case-based Reasoning and Information Retrieval, Stanford CA*. Stanford CA., (1993).

[18] M A Hearst and C Plaunt. 'Subtopic structuring for full-length document access'. In R Korfhage, E Rasmussen, and P Willet, eds., *SIGIR '93; Proceedings of the 16th annual international ACM-SIGIR conference on research and development in Information Retrieval*, pp. 59–68. New York, ACM press - 361 pp., (1993).

[19] D I Holmes. 'Authorship attribution'. *Computers and the Humanities*, **28**:87–106, (1994).

[20] T B Horton. *The effectiveness of stylometry of function words in discriminating between Shakespeare and Fletcher*. PhD thesis, University of Edinburg, 1987.

[21] D E Kieras. *Thematic processes in the comprehension of technical prose*, volume 1, chapter 4, pp. 89–108.

[22] I Kozima and T Furugori. 'Similarity between words computed by spreading activation on an english dictionary'. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 232–239, (1993).

[23] K Krippendorf. *Content analysis. An introduction to its methodology*. Sage publications, London, 1980.

[24] J N Littin. 'Applications of machine learning in information retrieval'. Technical report, Computer Science Department, University of Waikato, Hamilton, (1995).

[25] R Matthews and T Merriam. 'Neural computation in stylometry i: An application to the works of shakespeare and fletcher'. *Literary and Linguistic computing*, **8(4)**:203–209, (1993).

[26] R Matthews and T Merriam. 'Neural computation in stylometry i: An application to the works of shakespeare and marlowe'. *Literary and Linguistic computing*, **9(1)**:1–6, (1994).

[27] R A J Matthews and T V N Merriam. 'Distinguishing literary styles using neural networks'. In E Fiesler and R Beale, eds., *Handbook of Neural Computation*, chapter 8. IOP publishing and Oxford University Press, (1997).

[28] J Morris and G Hirst. 'Lexical cohesions computed by thesaural relations as an indicator of the structure of text'. *Computational linguistics*, **17(1)**:1991, 21–48, (1991).

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

[30] J J Paijmans. 'Relative weights of words in documents'. In L G M Noordman and W A M de Vroomen, eds., *Conference proceedings of STINFON*, pp. 195–208. StinfoN, (1994).

[31] J J Paijmans. 'Gravity wells of meaning: detecting information-rich passages in scientific texts'. *Journal of Documentation*, **53**(5):520–536, (1997).

[32] U Pieper. *Ueber die Aussagekraft statistischer Methoden fuer die linguistische Stilanalyse.* Gunter Narr Verlag Tuebingen,, 1979.

[33] J R Quinlan. *C4. 5: programs for Machine Learning.* Morgan Kaufman, 1993.

[34] G Salton. *Automatic text processing: the transformation, analysis and retrieval of information by computer.* Addison Wesley, - 530 pp., 1989.

[35] G Salton and M J McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill New York [etc.] - 448 pp., 1983.

[36] D R Tallentire. 'Confirming intuitions about style using concordances'. In A Jones and R F Churchouse, eds., *The Computer in Literary and Linguistic studies.* University of Wales press, (1976).

[37] S M Weiss and C A Kulikowski. *Computer Systems that Learn.* Morgan Kaufmann, 1991.

[38] V Yngve. 'A model and hypothesis for language structure'. In *Proceedings of the AMerican Philosophical Society*, volume 104, pp. 444–466, (1961).

8 Appendix

List of texts used in the comparisons:

8.1 Corpus I

The first thirty texts of the LOB-corpus, section J (scientific writings). Each section contains approx. 2000 words. Number 23 and 24 are from the same author (K. Lovell).

8.2 Corpus II

Five fragments from four books, downloaded from Internet, the Gutenberg project.

pride: Jane Austen, Pride and Prejudice, first 20,000 words.
sense: Jane Austen, Sense and Sensibility, first 20,000 words.
nabby: Jane Austen, Northanger Abbey, first 20,000 words.
wuth1: Emily Brontë, Wuthering Heights, first 20,000 words.
wuth2: Emily Brontë, Wuthering Heights, words 20,000 - 40,000.

8.3 Corpus III

Series of 78 essays, written by three American politicians in 1787. The precise authorship of a few of these essays is disputed. The Federalist Papers, numbers 40-70 as found on the CD-rom "Bookshelf Compendium", Medialine, Holland, 1996. This group includes nine papers attributed to Madison (40-48), five attributed to Hamilton (64-69), one attributed to Jay (64) and thirteen contested papers (49-63).

Received: 19/5/98, Accepted: 11/3/99

Notes for Contributors

The prime purpose of the journal is to publish original research papers in the fields of Computer Science and Information Systems, as well as shorter technical research notes. However, non-refereed review and exploratory articles of interest to the journal's readers will be considered for publication under sections marked as Communications of Viewpoints. While English is the preferred language of the journal, papers in Afrikaans will also be accepted. Typed manuscripts for review should be submitted in triplicate to the editor.

Form of Manuscript
Manuscripts for *review* should be prepared according to the following guidelines:

- Use wide margins and $1\frac{1}{2}$ or double spacing.
- The first page should include:
- the title (as brief as possible)
- the author's initials and surname
- the author's affiliation and address
- an abstract of less than 200 words
- an appropriate keyword list
- a list of relevant Computing Review Categories
- Tables and figures should be numbered and titled.
- References should be listed at the end of the text in alphabetic order of the (first) author's surname, and should be cited in the text according to the Harvard. References should also be according to the Harvard method.

Manuscripts accepted for publication should comply with guidelines as set out on the SACJ web page,

`http://www.cs.up.ac.za/sacj`

which gives a number of examples.

SACJ is produced using the LaTeX document preparation system, in particular LaTeX 2_ε. Previous versions were produced using a style file for a much older version of LaTeX, which is no longer supported. Please see the web site for further information on how to produce manuscripts which have been accepted for publication.

Authors of accepted publications will be required to sign a copyright transfer form.

Charges
Charges per final page will be levied on papers accepted for publication. They will be scaled to reflect typesetting, reproduction and other costs. Currently, the minimum rate is R30.00 per final page for contributions which require no further attention. The maximum is R120.00, prices inclusive of VAT.

These charges may be waived upon request of the author and the discretion of the editor.

Proofs
Proofs of accepted papers may be sent to the author to ensure that typesetting is correct, and not for addition of new material or major amendments to the text. Corrected proofs should be returned to the production editor within three days.

Letters and Communications
Letters to the editor are welcomed. They should be signed, and should be limited to about 500 words. Announcements and communications of interest to the readership will be considered for publication in a separate section of the journal. Communications may also reflect minor research contributions. However, such communications will not be refereed and will not be deemed as fully-fledged publications for state subsidy purposes.

Book Reviews
Contributions in this regard will be welcomed. Views and opinions expressed in such reviews should, however, be regarded as those of the reviewer alone.

Advertisement
Placement of advertisements at R1000.00 per full page per issue and R500.00 per half page per issue will be considered. These charges exclude specialised production costs, which will be borne by the advertiser. Enquiries should be directed to the editor.

South African Computer Journal

Number 23, July 1999
ISSN 1015-7999

Suid-Afrikaanse Rekenaar-tydskrif

Nommer 23, Julie 1999
ISSN 1015-7999

Contents