Optimising and Automating the Choice of Search Strings when Investigating Possible Plagiarism

Fintan Culwin & Mike Child
London South Bank University
London SE1 0AA
fintan/childm@lsbu.ac.uk

Abstract

This paper describes how to optimise the use of Internet search engines when investigating a document for possible non-original content. Services such as Turnitin do not guarantee to identify all non-original content, leading tutors to have to conduct manual searches when suspicion of non-originality remains. Previous studies have suggested that the investigator should manually select memorable phrases from the paper and submit them to a general search engine. The studies in this paper demonstrate that selecting phrases at random is just as effective. Several corpora of documents were obtained from a number of different academic areas, and several phrases were obtained from each. Strings, of increasing length starting with a single word, from these phrases were submitted to specialised and general search engines and the number of hits recorded. A common finding of these searches was that, in almost all cases, strings of six words were sufficiently distinct to uniquely identify the document that the string was taken from. One consequence of this is that totally automated tools are possible for this search-engine based non-originality detection technique.

Index terms

Plagiarism, academic integrity, non-originality analysis, internet search engines

Introduction

When the lexicographers admit the verb ‘to google’ into formal dictionaries it will have at least two meanings. The most general meaning will refer to using a search engine to locate information on the World Wide Web, as in ‘I googled for information about xxxxx but found nothing of any use.’. A second, more specific, meaning relates to tutors using search engines to locate the text of a student submission that they are suspicious of. As in, ‘I googled phrases from xxxxxx’s submission but could not find where they got it from.’.

Specialised systems for checking students’ submissions, such as Turnitin or MyDropBox, are routinely used by many institutions; however these cannot be guaranteed to be totally exhaustive. Studies have shown (Slatterwhite & Gerein 2001, Culwin 2009) that greater, but not exhaustive, coverage can be obtained by using general search engines to check submissions which raise suspicion but are not shown to be compromised by the routine screening.

The process of googling a student submission can be very time consuming. It involves choosing phrases from the student’s submission, entering these as search terms into the search engine, downloading the top hits suggested by the search engine and comparing the downloaded documents against the submission in an attempt to find evidence of undue similarity. This entire process is repeated until a match is found or until the tutor can devote no more time to it. There are tools that automate parts of this process (Lancaster & Culwin 2004), but it essentially remains manual and laborious.

One operational problem with the manual search process is to know how to construct the most effective search terms from a student’s submission. A study by Kaner and Fielder (2008) suggests “. . . [skimming] each paper, looking for one or more memorable phrases to conduct a manual, full text search . . . “. The study used a test corpus of thirteen papers from the IEEE Xplore electronic library. The authors reported that they used no more than three phrases to find one that uniquely

identified the paper. The phrases ranged from three to eight words, with nine of the thirteen phrases being six or seven words long.

Studies by Olsson (Olsson 2004, 2008) investigated the concept of the ‘maximum string of coincidence’. That is, how long does a string have to be before it is sufficiently unique for its appearance in two or more documents to be beyond any reasonable doubt. Initially he postulated that the value would be in the order of 40 words, but discovered that it might be as low as 40 characters.

An informal study by Coulthard and Johnson (2007) considered the evidential value of single identical strings. Sequential Google searches were conducted, extending the length of the search string by one word each time until a unique string was obtained or no matches were found. The authors only investigated two phrases, relating to a disputed police investigation, and speculated that the rarity scores were comparable to those of DNA evidence routinely used in courts.

Although the strings used in the Coulthard and Olsen studies were deliberately selected from forensic linguistic investigations, they did not seem to be particularly unusual. The construction of a sentence can be thought of as a combination of two processes known as idiomatic and open choice, also known as statistical (Sinclair 1991). Idiomatic construction involves the use of an established phrase e.g. “hyper text markup language” or “to whom it may concern”.

Statistical sentence construction is where a choice is made every time a word is added to the sentence. At any point there are a number of words in the available vocabulary which could be chosen. Although some words might be statistically more likely than others, the probability of any particular word being chosen is less than one. Accordingly the a-priori probability of a particular phrase containing the exact words that it does, becomes smaller as the phrase gets longer.

An example that dramatically contrasts idiomatic and statistical sentence construction involves the sentence ‘*That would be an ecumenical matter.’*. This phrase became an idiom after it featured in an episode of the TV situation comedy Father Ted. This phrase is reported by Google as having about 718,000 hits. A

sentence closely related in construction and meaning might be '*That would be a parochial matter.'* which is reported by Google as having no hits[1].

Once	more	into	the	breach	dear	friends
once	again	assail	the	breach	beloved	brothers
yet	further	assault	that	break	cherished	comrades
{}	more	attack		crack	dear	friends
		charge		gap	precious	men
		into		rift	{}	soldiers
		unto		rupture		
		rush				

Fig. 1. Statistical sentence construction.

Fig. 1 further illustrates this concept by considering the alternative choices that could have been made by Shakespeare when constructing one of his more famous sentences. Although few are as eloquent as the idiomatic original '*Once more unto the breach dear friends*.', all are possible; e.g. "*Yet again attack that gap beloved brothers*." or "*Further into the rupture cherished soldiers*.". Choosing each word at random gives a total of 12,600 possible sentences and, assuming that all choices are equally likely, gives a probability for any single version of .00008.

In practice, sentences are constructed using both idiomatic and statistical processes and an idiomatic phrase can be thought of as a single statistical choice. When a tutor chooses a memorable phrase from a student submission they would presumably avoid anything they recognised as being idiomatic. A quality of a sentence known as markedness identifies non-idiomatic phrasing and is related to the extent to which the word choices taken are unexpected.

Markedness might be one of the characteristics that raises a tutor's suspicion in the first place and one which causes a phrase to be recognised as memorable. By way of trivial example the idiomatic phrase '*you and I'* might be rearranged by a student who did not recognise it as an idiom, in an attempt to disguise it, into '*I and you'*. This would be immediately regarded as odd by someone fluent in the English language.

[1] At least until this paper is discovered and indexed by Google!

One way of testing for markedness is to search a corpus for the idiomatic and marked phrase and compare the relative frequency of each. A Google search of '*you and I'* indicates about 56,700,000 occurrences compared to just 523,000 for '*I and you'.* Alternatively the phrases '*dear friends'*, "*beloved brothers'* and '*cherished soldiers'* is 4,730,000, 48,400 and 281 respectively. These frequencies reflecting the markedness of the two randomly constructed sentences above.

One way of automating the googling of phrases from a student's submission would be to choose random phrases and use these as search terms. However, the existence of idiomatic phrases in the submission might make this process ineffective. The studies reported in this paper were designed to investigate this possibility and, if shown effective, indicate that automation of the process is a possibility.

The Methodology

By way of introduction to the investigations which follow, the phrase "*It is possible by chance alone*" will be considered. Starting with the word "It" and then the two word phrase "It is", all five sub-phrases and the six word phrase itself were submitted as quoted strings to Google. The results are shown in Table 1.

Phrase	Hits
"It"	850,000,000
"It is"	265,000,000
"It is possible"	72,000,000
"It is possible by"	235,000
"It is possible by chance"	142,000
"It is possible by chance alone"	3

Table 1. Successive Google searches using progressively longer sub-phrasess.

A visual examination of the three hits given for the full phrase showed that they were three distinct documents, not copies or partial copies of a single document. In essence this is the basis of the methodology. Random six word phrases were chosen at random, from documents which had been chosen at random from a corpus. These phrases were then submitted to a specialist and a general (Google)

search engine and the number of hits recorded. For the final six word search if the number of hits reported was manageable, less than 10 documents, they were then examined to see if they contained the document being sought. On some occasions the six word search yielded no documents but a shorter search gave a more manageable number, in which case the hits from the shorter search were examined.

An initial informal investigation indicated that there was no essential difference between phrases of six or seven words, so on the basis of parsimony six was chosen. The study consisted of five distinct investigations including corpora from the Institute of Electrical and Electronic Engineers (IEEE) digital library, the Association of Computer manufacturers (ACM) digital library, the Academic OneFile resource, the International Index of Performing Arts (IIPA) collection, and Wikipedia. Each of these is described in turn in the sections below.

The IEEE investigation

The first part of the investigation consisted of a repeat of the Kaner and Fiedler investigation, but using random six word phrases instead of selected phrases. The same 13 IEEE papers were taken and three six word phrases were taken from each, using a scripted computer program. The phrases were selected randomly ignoring the first and last hundred words of the text version of the documents in order to avoid selecting phrases from the abstract and keywords or from the references. If the phrase contained a number expressed in digits, or a proper noun, or an acronym, or other uncharacteristic content, it was rejected. If the phrase straddled two sentences then either the last six words of the first sentence or the first six words of the second sentence were taken, depending upon where the sentence break occurred within the phrase.

These phrases were then used as quoted search terms within the open text search box on the advanced search page of the IEEE Xplore digital library. All searches were conducted on the same day, continually within a period of about 2 hours and all controls on the search page were left at their default settings. The results of this part of the investigation, known as the IEEE Xplore investigation, are shown in Table 2 and also in Figure 2.

N° Words	Average Hits	Best Hits
1	8272298	689398
2	118577	139912
3	70791	46832
4	24169	77
5	26	46
6	0.9	1

Table 2: Results of the IEEE Xplore investigation

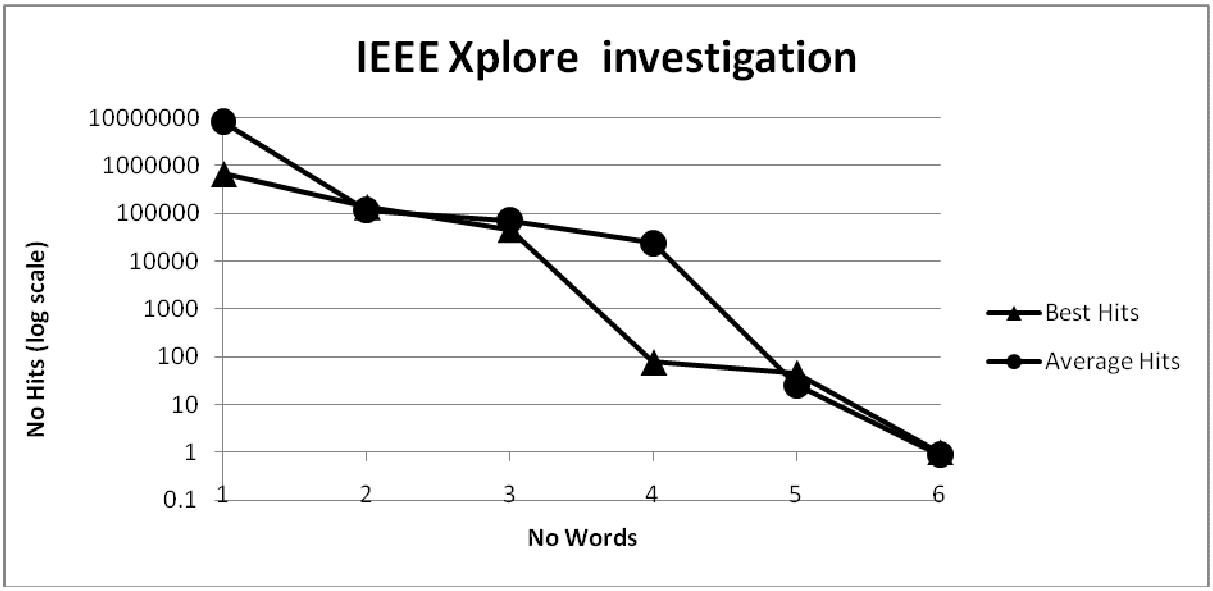

Figure 2. Results of the IEEE Xplore investigation

The average hits values are the average number of hits from all three searches for all 13 documents. The best hits values are the average of either the first search that returned a single hit or the search that returned the lowest (non-zero) number of hits, for all 13 papers. In this investigation all sets of three searches had at least a single phrase that resulted in a unique hit which identified the document being searched for.

The same corpus of 13 papers and the same 39 six word phrases were then used within a Google search. The results of this investigation, known as the IEEE Google investigation, are shown in Table 3 and Figure 3.

N° Words	Average Hits	Best Hits
1	6162053179	5879369231
2	245185844	61114838
3	19809199	22326794

4	6827089	1503267
5	10687	608
6	2452	2

Table 3: Results of the IEEE Google investigation

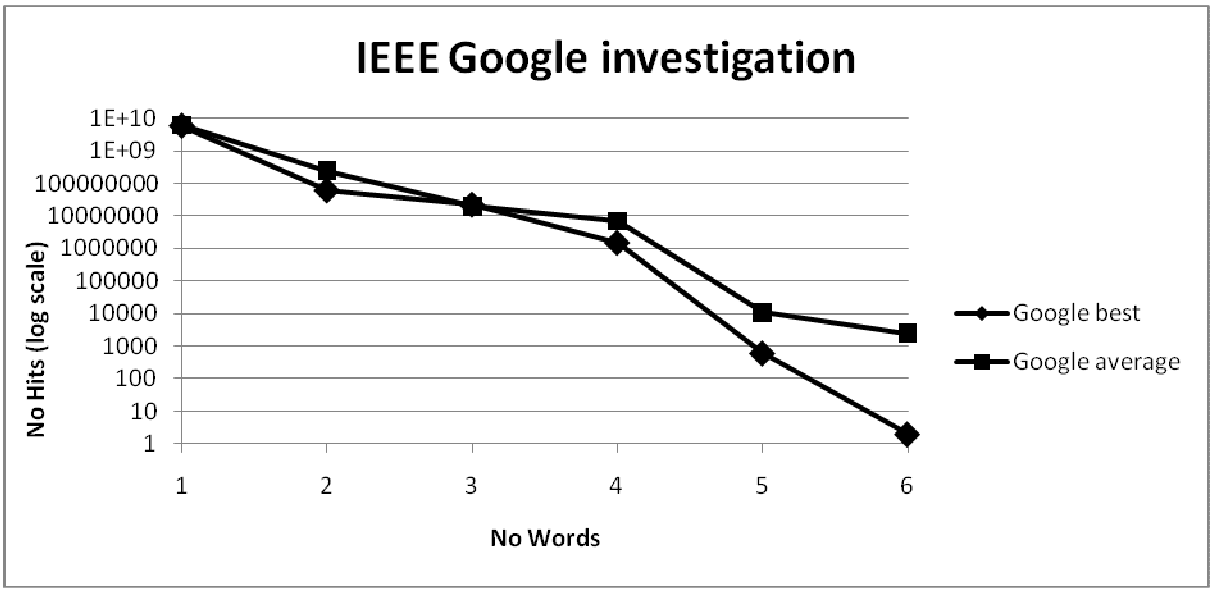

Figure 3: Results of the IEEE Google investigation

In this investigation the distinction between the average hits and best hits values is much more pronounced and its utility is more obvious. Whereas the average number of hits for all 39 six word phrases is 2452 hits, the average number of (non zero) best hits across the 13 documents is only 2.

Of the 13 documents searched in this investigation only one was not located, giving zero hits for two of the six word phrases and two false hits for the third. Of the seven six word searches which gave ten or fewer hits, only two did not report the sought for document as the first hit in the list. Operationally it would appear that Google is able to index documents within the IEEE Xplore digital library, but will only show a summary page, including an abstract, unless the user has a subscription to the library.

The ACM investigation

The second part of this investigation repeated the essence of the first part of the investigation but used the Association of Computer Manufacturers (ACM) digital library. The Kaner and Fiedler paper did not describe how the 13 document corpus

was obtained. An examination of the publication titles, and the journal or conference that they were published in, and keywords used, suggest that it was not a random sample.

The organisation of the ACM digital library appears to allocate every document it contains a sequential identifying number. This number is a part of the URL used to retrieve the document. The number of documents contained in the library is stated allowing a random number generator to be used to obtain documents at random. A corpus of 13 documents was assembled by repeatedly selecting papers at random, accepting only those that were published in or after 1996. This restriction being based upon the earliest paper in the Kaner and Fiedler corpus.

The investigations were conducted as closely as possible to the IEEE investigations, as described above, and the results of the investigation are given in Table 4 and Figure 4.

	ACM Search		Google Search	
No. Words	Average Hits	Best Hits	Average Hits	Best Hits
1	136128	116145	7495885384	9571225385
2	13410	10381	196868014	91215601
3	2776	87	17148544	1653760
4	1185	16	3835809	318971
5	84	1.6	12976	54
6	13	0.9	2140	1.8

Table 4. Results of the ACM investigations

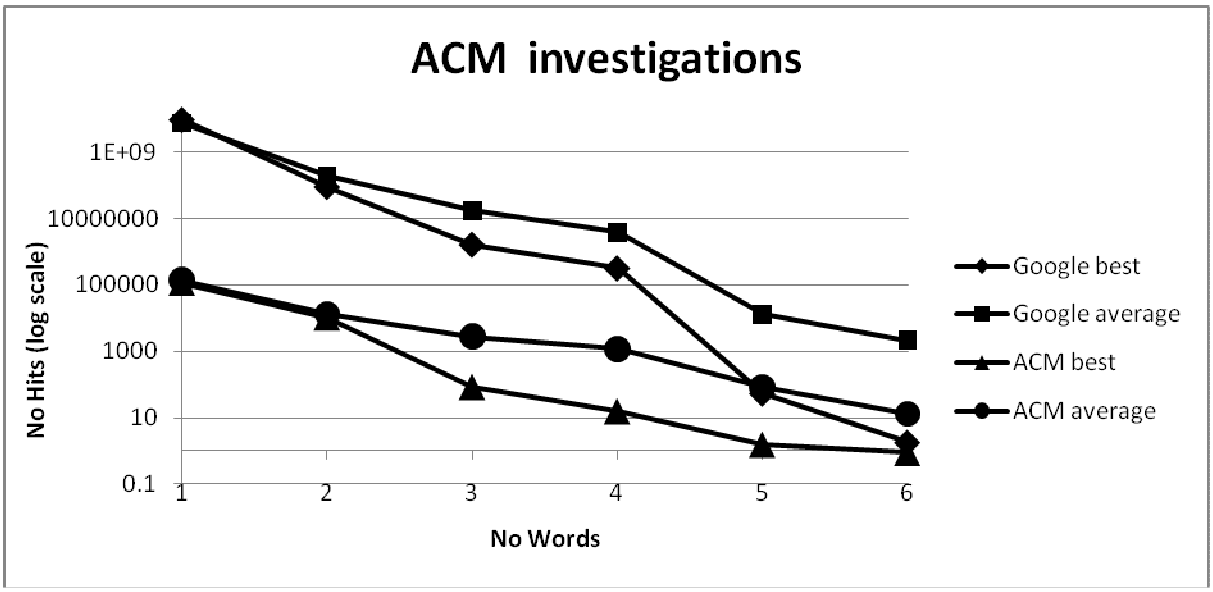

Figure 4. Results of the ACM investigations

The results of these investigations are largely comparable with those of the IEEE investigations. Of the 13 documents 12 were located using this technique and one was missed. The three searches for this missed paper gave 0, 342 and 0 hits; the target document not being in the top ten hits reported. A repetition of the search for this document, on a different day, using different 6 word fragments gave 0, 0 and 0 hits for the three phrases. This suggests that the style of writing in this document is particularly idiosyncratic.

Google was able to locate 11 of the 13 documents. Of the two documents which were missed, one was the same document that was missed in the ACM search. Once again these 2 documents were revisited on a different day with different sets of six word phrases. The document that was missed on both previous searches was located immediately but the other document was missed again.

The Academic One File & IIPA investigations

The two investigations described above used papers taken from the engineering and computing academic domains. It might be that the results obtained were peculiar to those genres. Accordingly two further comparable investigations were conducted. The first used the Academic OneFile resource which contains articles from 11,000 titles in the fields of current events, general sciences, social sciences, and humanities. The corpus was limited to documents written in English with more than 1,000 words which yielded approximately 86,500 documents from which 10, post 1996, were selected at random. The results of this investigation are shown in Table 5 and Figure 5.

	Academic OneFile Search		Google Search	
No. Words	Average Hits	Best Hits	Average Hits	Best Hits
1	546884	392033	4263089713	485860940
2	217590	15786	119569129	1331727
3	50365	263	57014385	249885
4	804	1.3	5431754	2243
5	171	1.2	4788443	9.4
6	94	1.1	315159	1.7

Table 5. Results of the Academic OneFile investigation

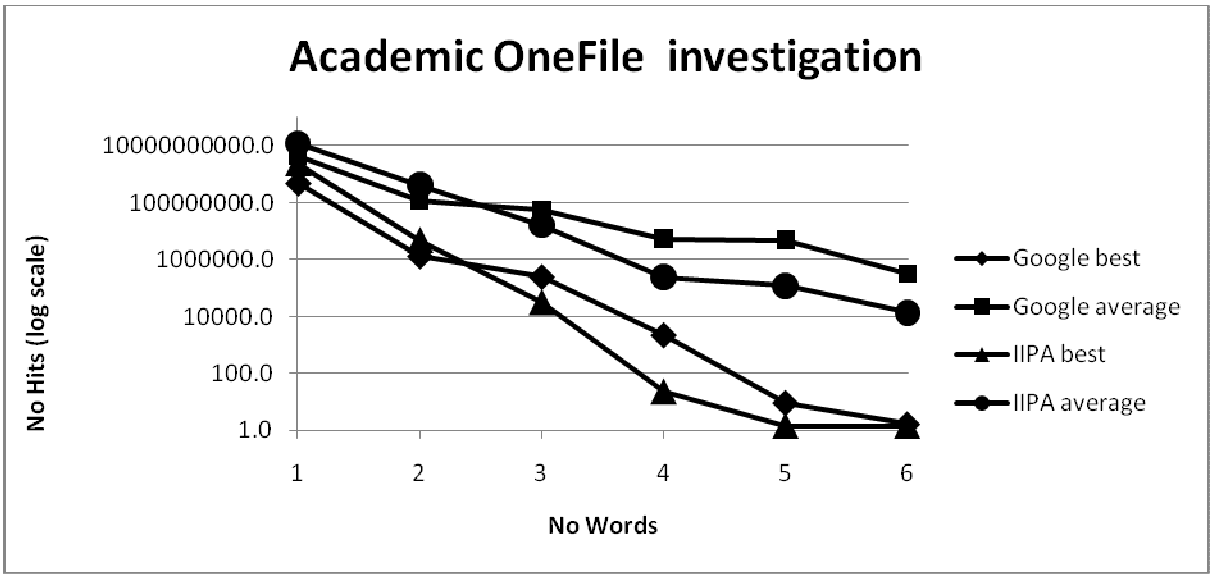

Figure 5: Academic OneFile investigation

The Academic OneFile search engine found the article as a single hit in at least one of the three phrases every time. Google found the article as a single hit in at least one of the three phrases on 7 occasions, as 1 of 2 hits on one occasion, as 1 of 6 hits on one occasion, and failed to locate the document at all once.

The second investigation in this part of the study used the International Index of Performing Arts (IIPA) which contains articles from theatre, dance and film. The resource contains 263 English language full text journals. Ten journals were chosen at random and from each journal a random, post 1996, issue and then a random article was chosen. The results of this investigation are shown in Table 5 and Figure 5.

	IIPA Search		Google Search	
No. Words	Average Hits	Best Hits	Average Hits	Best Hits
1	48893	18102.1	2832865586	590486760
2	5520	348	215823944	11604732
3	14	1.4	426293	28744
4	1	1	96586	20528
5	1	1	4047	1.3
6	1	1	854	1.2

Table 5. Results of the IIPA investigation

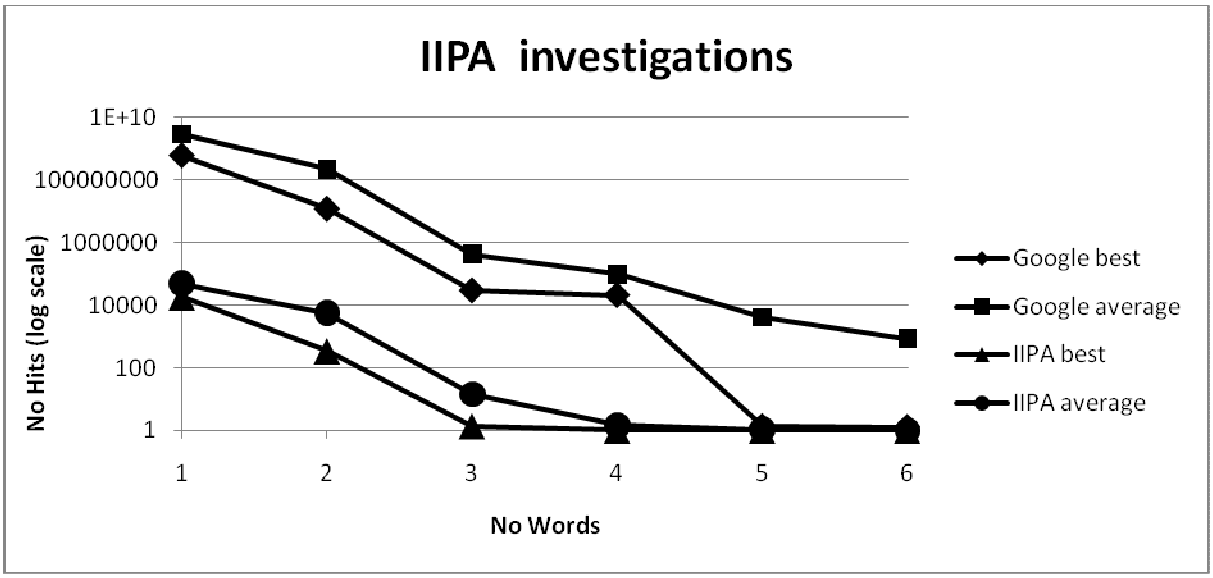

Figure 5: IIPA investigation

The IIPA search engine found the article as a single hit in at least one of the three phrases every time. Google found the article as a single hit in at least one of the three phrases on 6 occasions, as 1 of 2 hits on two occasions, and failed to locate the document at all twice.

The Wikipedia Investigation

A final investigation was conducted in an attempt to generalise the conclusions from academic to more general resources. The on-line encyclopaedia Wikipedia has a 'Random article' feature, which was used to generate a corpus of 13 random documents. Many Wikipedia articles are very short, less than 200 words, and many others consist mainly of lists rather than text (for example lists of albums and song titles or lists of sporting fixtures). Accordingly the corpus consisted of the first 13 documents which were longer than approximately 200 words and which were largely textual.

The investigation revealed another characteristic of Wikipedia content in that it is reproduced in many other locations. Wikipedia publishes its content under the GNU free documentation license, commonly known as 'copyleft'. This licence allows the content to be freely reused, or further developed, provided that the original source is acknowledged. Some organisations have taken advantage of this facility to

reproduce Wikipedia content. The effect of which for this investigation is that a search might give several hits which upon examination turned out to be the same content. Accordingly multiple hits such as these were recorded as a single hit.

The results of this investigation are shown in Table 6 and Figure 6. The results of this investigation are again largely comparable with the previous investigations. All of the documents were uniquely located by one of the three associated searches.

N° Words	Average Hits	Best Hits
1	3805667179	5977414615
2	194599242	7040076
3	216877	173501
4	16250	25520
5	6214	13
6	67	1

Table 6: Results of the Wikipedia investigation

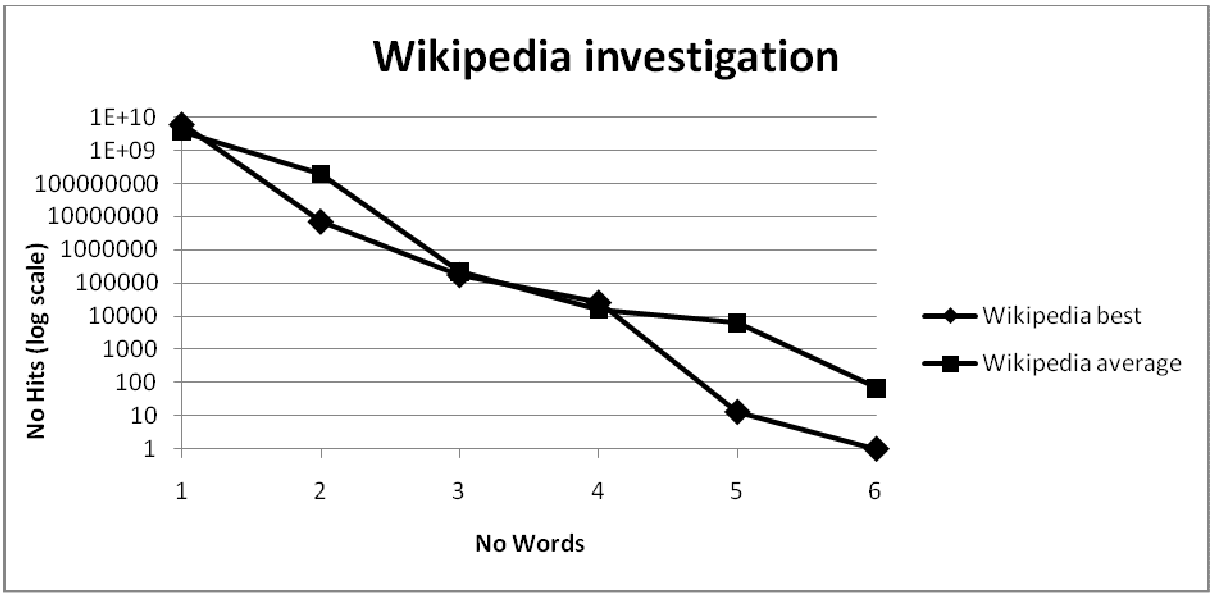

Figure 6: Results of the Wikipedia investigation

Discussion

The five different investigations involved a total of 59 different documents randomly chosen from a number of different genres. Of these 6 documents, or approximately 10%, were not located using three random six word phrase Google searches. The meaning of locating in this context being that the document (or a copy of it) was

identified in the top 10 hits returned by Google. Some of the documents missed in the first search were located on a second attempt when alternative phrases were chosen. Accordingly this technique has been shown to be 90%+ effective at locating documents via Google. This part of the discussion has restricted consideration to the Google, as opposed to a specialist, search engine as this would be how an investigation of a suspected document would proceed. However the size of the sample is rather small, although randomly chosen from the corpora, and a repetition of this study using a larger sample taken from a larger number of corpora would be useful.

In all there were 177 six word Google searches and the average number of hits per search was in the order of 55,000. This might suggest that an overall conclusion regarding the efficacy of the technique is compromised. However a more detailed examination of the data suggests a technique to discount searches which have inadvertently used idioms in the randomly chosen search phrases. Of the 177 searches, 40 yielded no hits, although 18 of these searches had already given less than 10 hits with a phrase of fewer than 6 words. A further 74 searches yielded a single hit, 22 searches yielded 2 hits and there were 17 searches were in the 3 to 10 hits range. If 10 hits is regarded as a manageable number of documents to investigate then the technique is approximately 75% effective on this measure.

There were only 24 searches (14%) that gave more than 10 hits and 12 (7%) which gave more than a thousand. The three largest numbers of hits were 9,360,000, 185, 000 and 94,000; which are clearly idiomatic and which produced the thoroughly misleading overall average cited above. Accordingly search phrases that yield more than 10 Google hits can operationally be regarded as idiomatic and automatically excluded from consideration.

This study has relied upon using the Internet as a general corpus in order to establish the uniqueness of a phrase. The suitability of the general Internet for this purpose might be questioned. However, Olsson (2008) suggests that as the ratios of common stop words such as: *the*, *of*, *to*, *and*, etc. as reported by Google are very similar to those in more general corpora, using the general Internet in this way is valid and reliable.

The study used a number of different search engines. Although all investigations made use of the Google search engine, four of the studies made use of a search engine located within a digital library. None of these four search engines were branded as making use of Google search technology, and although they might all four be implemented using the same underlying search technology or detailed algorithm, this seems most unlikely. Accordingly any conclusion regarding the uniqueness of six word phrases would seem to be either a characteristic of the English language in general or of search engines in general; but not a result which is particular to Google.

Some anomalies were noted during the investigations, but not pursued further. A well know anomaly relates to the outcomes of Google searches. Although precise technical details are not made public, it is known that the corpus used by Google is continually expanding and hence the results reported may not be stable. A search for a particular phrase made at different times may yield different results. Accordingly care was taken to ensure that all searches in an investigation, using any search engine, were made as quickly as possible and in all cases were made on the same day.

On several occasions as the length of the search string increased the number of hits reported also increased. For example on one investigation the phrase ‘likely to pervert’ yielded 2120 hits whilst the longer phrase ‘likely to pervert the’ yielded 1,194,000 hits. Logically the 1,194,000 documents containing the four word phrase should also contain the enclosed three word phrase. A repeat of these two searches on a different day yielded 332,000 and 188,000 hits. This latter result suggests that the anomalous result is not caused by anything inherent within the Google search algorithms. A more likely explanation would relate to load balancing. Google has a large number of servers located at a number of different locations around the planet. Any search may be directed to any server and the results reported by different servers may vary, even if the searches are submitted at the same time. For a search which yields a large number of hits Google estimates, rather than counts, the number of hits. The resources made available to produce this estimate are

dependent upon how busy the server is and hence may vary, even on the same server.

Conclusions

The overall results of the investigations indicate that choosing six word phrases at random is at least as effective as manually selecting memorable phrases. Although an individual search may use an idiomatic phrase and yield hundreds or thousands of hits, in most cases using three searches will ensure that a 'sufficiently unique' phrase is located. A sufficiently unique phrase is one that will yield zero, one or a very small number of hits. This is a suitably small number of hits for them to be compared with the document being investigated either manually or automatically. Although this technique is imprecise and is not guaranteed to produce an accurate result, this is also the case for the specialised systems such as Turnitin which this technique is intended to complement.

This conclusion also suggests that matches of less than about five words are of little or no evidential use for academic misconduct investigations. It would help improve the signal to noise ratio when looking at non-originality reports if there was a control to prevent matches of less than ***n*** words being shown, with a suitable default value of ***n*** being five or six.

This conclusion raises the question of the evidential value of single strings in an academic misconduct investigation. The occurrence of a string of as little as six words in a student submission, whose frequency of occurrence is shown by an Internet search to be unique or nearly unique, can be assumed to be, beyond reasonable doubt, copied. Although it would be most unreasonable for an institution to penalise a student for a single transgression such as this, this technique and this study does provide an evidential basis to dismiss a student's defence that the phrase was in common usage. There is the possibility that any particular phrase is in idiomatic usage within a cohort of students, but a search for the phrase within the corpus of student submissions would establish or deny this.

There is a major operational weakness of this technique in that it is only appropriate for detecting non-original content that has been used verbatim. Any changes to the original text will degrade its effectiveness and in the extreme case changing every sixth word will render it totally ineffective. However, many students do not make many or any changes to the text that they illicitly reuse and tutors could restrict the investigation to those sections of the text whose markedness makes them suspicious. One final operational consideration from this study is that should a search not identify a suspicious document then a second search, possibly on a different day, may be successful.

The results also confirm that a totally automated system can be built. The existing OrCheck tool (Lancaster and Culwin, 2004) requires the user to select phrases from the document being investigated and to copy and paste them into the search boxes. Due to technical changes in the software services provided by Google, the tool had become obsolete and could no longer be used. The tool has now been re-engineered, as OrCheck2, to make use of the changed Google services. It also has an auto-phrase feature which will select three random six word phrases from the document and submit them automatically. Although the tool is still under re-development, the early indications are that the process is at least as effective as manually choosing phrases.

Academic Integrity Statement

This work is unfunded and this is the first publication of this topic, hence there is no reuse of any previous publications.

References

Coulthard M. & Johnson A. (2007) An Introduction to Forensic Linguistics, Routledge UK.

*Culwin F. (*2009), *The Efficacy of Turnitin and Google*, Proc. 10th Annual HEA/ICS conference, Canterbury UK August 2009.

Kaner, C. & Fiedler, R.L., (2008) *A Cautionary Note on Checking Software Engineering Papers for Plagiarism* IEEE Transaction on Education, 51(2) pp 184-8.

Lancaster T. & Culwin F., (2004), *Using Freely Available Tools to Produce A Partially Automated Plagiarism Detection Process*, Proc. ASCILITE 2004, Perth Australia Dec 2004.

Olsson J. (2004) Forensic Linguistics (first edition), Continuum UK.

Olsson J. (2008) Forensic Linguistics (second edition), Continuum UK.

Sinclair J. (1991) Corpus, Concordance, Collocation, Oxford: Oxford University Press UK 1991.

Slatterwhite R. & Gerein M., 'Downloading detectives: searching for on-line plagiarism, online 2001, www.coloradocollege.edu/library/course/downloading_detectives_paper.htm