International Education Journal, 2005, 6(4), 421-429.
ISSN 1443-1475 © 2005 *Shannon Research Press*.
http://iej.cjb.net

Finding the true incidence rate of plagiarism[1]

Julie Price
School of Education, University of Southampton, UK *j.price@soton.ac.uk*

Robert Price
Information Systems Services, University of Southampton, UK *r.l.price@soton.ac.uk*

This paper reports on research that explores the use of detection software in the fight against plagiarism. The aim of the research was to determine if the true incidence rate of plagiarism could be found for a cohort of Higher Education students. The paper outlines the problems and issues when attempting this. In addition, this report highlights the views of students when such a service is being used. The findings suggest that the use of such detection services is not without problems and raises the issue that such services may have less value in detection and more value as a learning and teaching tool.

Plagiarism, detection software, higher education

INTRODUCTION

Concerns over plagiarism were once again brought to the fore in the United Kingdom in a Prime Minister's Special Report broadcast on BBC Radio 4, which outlined the results of a survey of British Universities regarding plagiarism (details in *The Guardian*, 2003). Approximately 50 plagiarism cases per university could be calculated from the figures broadcast, and "a third said they were having to deal with many more such cases [of plagiarism] compared with a few years ago" (The Guardian, 2003). It was clear that following the 2003 survey there was an overwhelming sense that plagiarism is on the increase.

The incidence rate of plagiarism varies widely in the literature. Some studies have calculated rates based on questionnaires asking students whether they have plagiarised in the past (such studies often investigate the wider topic of cheating of which plagiarism is a part). For example, Franklyn-Stokes and Newstead (1995) reported that "Behaviours such as: copying each other's work, plagiarism, and altering and inventing research data were admitted to by more than 60 per cent of the students" in their sample. Other tutors have closely examined work handed in by a cohort of students to determine the incidence of plagiarism. For example, Austen-Baker (2003) reported in the *Times Higher Educational Supplement* that, of the 60 scripts viewed, "Only 6 were wholly free of plagiarism and about 4 were significantly plagiarised or the result of collaboration". Jones (2003) suggests that "it is estimated that up to 10 per cent of degree level work is now affected by so called *mouse-click plagiarism*". These findings highlight the growing concern over the influence of the internet, which has expanded massively in the last few years.

The expansion of the internet has undoubtedly resulted in a vast resource base that is readily accessible to students (Gresham, 2002; Park, 2003). It makes life easy for the student to plagiarise and difficult for the tutor to catch the guilty. The ease of co-called 'cutting and pasting' from

[1] This paper was originally presented at the *Joint Information Systems Committee Plagiarism Conference at Newcastle, England* 28-30th June 2004 (Price (previously Lakomy) and Price, 2004).

sources on the web is clear and tutors can no longer be expected to know everything that has been written on a topic in intimate detail. Even without the use of essay banks or papermills, intentional plagiarism is easy.

The fight against plagiarism, in this new era, is using the plagiarist's tools – the web. Services such as the United Kingdom's Joint Information Systems Committee (JISC) Plagiarism Detection Service trawl the web and its own database of previously uploaded work to match text within an essay to that published on web pages or in the database. If matched text is found, and a reference or acknowledgement is conspicuous by its absence, the plagiarist has been caught.

In investigating the incidence of plagiarism and evaluating the use of such services, there is, however, the dilemma of whether or not to tell students before they submit a piece of work, that such a service will be used. Not telling them may be deemed unethical, but telling them will alert them and probably lead to a change in their behaviour, deterring some who might have been tempted to commit plagiarism. On the one hand, such a change in behaviour may be exactly what is desired – a deterrent to reduce the incidence of plagiarism. On the other hand, it disguises the true incidence of plagiarism when such a service is not in operation. Hence, it may be possible that true incidence rates for plagiarism will never be known unless essays are checked **after** they have been handed in, with no prior warning to students. This, however, creates an ethical dilemma, particularly if plagiarism is found.

The purpose of the present study was to try and determine a more accurate rate for plagiarism than has been established in many previous studies.

METHOD

The University of Southampton subscribed to a free trial offer of the JISC Plagiarism Detection Service. The service was offered to several Schools and Departments within the University. Information Systems Services (Southampton University's computing department) ran a number of demonstration sessions and tutors who showed an interest were invited to trial and evaluate the service.

Staff teaching on one particular course were keen to trial the service. They selected one unit for an undergraduate third year cohort where the service could be used as part of the assessment for that unit. The major piece of coursework for students on this unit was a traditional style essay on a related topic.

Students were not told the service would be used until they had written their draft essay **but** they knowingly submitted the draft to the service to receive feedback before a final (amended) submission was handed-in for marking. The timing of the submission and feedback is shown in Figure 1. The draft essay produced by each student was to be peer reviewed as part of the unit assessment with students giving feedback to each other before their final submission. The detection service, therefore, was incorporated as part of that feedback process. This meant that students were able to respond to the service's plagiarism report by changing their essay and eliminating and resolving any problem text (for example, by referencing correctly) before handing in the final version.

Data Analysis

Essays were submitted by the students to the service and were analysed, taking the basic concepts outlined in Weinstein and Dobkin (2002) but adapted for the present study. Weinstein and Dobkin investigated internet plagiarism by analysing identified text that matched internet sites and categorised them into three clearly defined groups: (a) legitimate research, (b) small-scale plagiarism, and (c) large-scale plagiarism. Initially it was proposed for the present study that each

highlighted piece of text from the Plagiarism Detection Service report within an essay was to be placed into one of three categories:

- **Not plagiarism** – referencing and acknowledgement present and correct according to our guidelines and expectations.
- **Minor plagiarism** – plagiarism in the strictest sense but deemed to be more a case of poor academic practice, for example, quote marks are missing for copied text but the author or source is acknowledged.
- **Plagiarism outright** – highlighted text with no evidence of reference or acknowledgement or quote marks if needed.

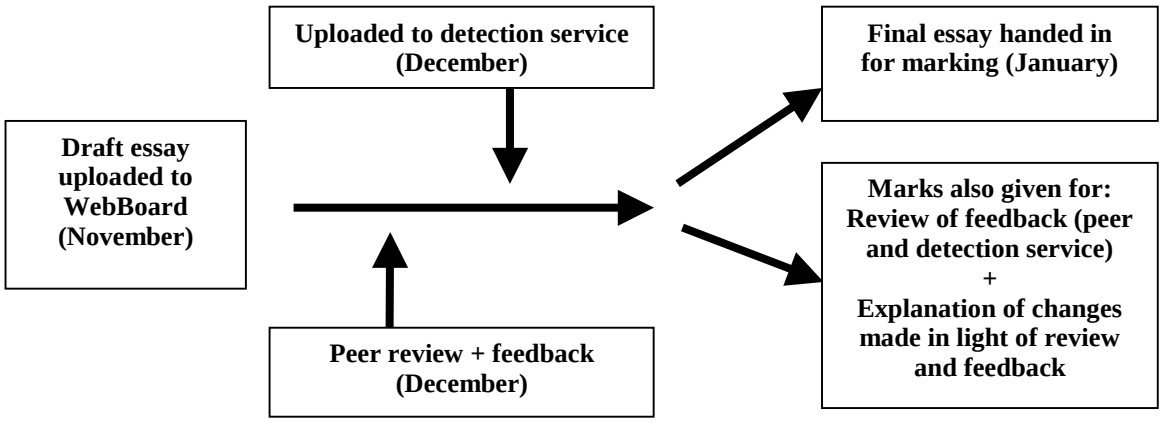

Figure 1. Schematic of assessment and timeline

From this, rates for internet plagiarism were to be calculated and compared to reports in other literature. In addition, the Year Three group was asked to comment on the service and their experience using it.

Difficulty with the Data Analysis[2]

In theory, the data analysis should have been straight forward but in reality it was not. The following are examples of the difficulties, issues and interesting examples found when trying to analyse the reports returned by the Plagiarism Detection Service.

Example 1

Note that this is a simple example to illustrate what happened with more complexity in other instances.

One student wrote: "Diabetes is one of the leading causes of death and disability in the United States with type 2 diabetes accounting for 90-95% of all diabetic cases" (Author name cited in Author name, 2000, p. 1345).

The text *Diabetes is one of the leading causes of death and disability in the United States* was highlighted by the JISC Plagiarism Detection Service in one colour, and the similarity was attributed to http://www.aoa … Diabetes.html.

[2] Please note that in order to maintain the anonymity of individuals and web sites highlighted in the following section, names, dates and full internet addresses have been shortened where necessary.

The text *accounting for 90-95% of all diabetic cases* was highlighted in a different colour, and the similarity was attributed to http://www.fit … diabetes.pdf by the JISC Plagiarism Detection Service.

This does not show plagiarism on the part of the student, but potentially there are three different possible sources for this quote and only two being highlighted by the Plagiarism Detection Service. It should be noted that this simple example more likely illustrates 'common knowledge' where attribution is not required. However, the current study did throw up similar examples to this where common knowledge was not likely to be a defensible argument.

Example 2

One student wrote: By the latter half of the century the Pima Indian lifestyle had become … and excessive food consumption (Author name, 2003, p. 101)

Text in the essay was highlighted by the Plagiarism Detection Service and the similarity attributed to the database when, in fact, the student had not used quote marks but had acknowledged a legitimate, though not web-based, source. The way the student has written this part of the essay, without quote marks, implies that he or she had paraphrased the information from the attributed source. However, checking the side-by-side version of the report and the database essay highlighted as similar, there was a strong similarity between both highlighted texts within each essay. The question arises, therefore, as to whether either of the students truly paraphrased from the original source (in this case it was actually a very poor attempt at paraphrasing). Alternatively, it could have been that one student had copied from the other – since the highlighted section was relatively short, direct copying from one student to another was not likely to be the case.

Example 3

The similarity with the database, which was highlighted in some instances, was flawed. For example, a few students uploaded their reference section in addition to the main text of the essay despite the fact that they had been asked not to. As several students used some of the same journal articles, once one reference list had been uploaded to the service then any others may have shown similarity in the report if they too uploaded their reference section and used the same reference source(s). The advice, therefore, is for tutors to insist that students only upload the main body of the text and not reference sections or titles within the essay upload box.

For data analysis in such cases, flawed highlighted text was not counted at the in-depth analysis stage.

Example 4

One student wrote, without quote marks: It is now important to move from demanding more data to learning how to apply what we already know to every day society (Author name, 2000, p. 670).

The JISC plagiarism detection service highlighted *move from demanding more data to learning how to apply what we already know*, and attributed it to http://www.ann … 010.html. Checking the side-by-side versions, the web site had the following text: We must move from demanding more data to learning how to apply what we already know.

The web site version had no acknowledgement of a source or quote marks, but it is questionable whether it is really the sort of sentence several authors would come up with and, therefore, might not need acknowledgement. This certainly should have had quote marks in the student's essay.

Example 5

One student wrote: It also supports Author name (2000, p. 669) who states that “one of the most powerful predictors of the development of diabetes in genetically susceptible persons is weight gain in adulthood.”

The detection service highlighted a similarity for the quoted text to that found on the web site http://www.ann … 010.html. The web site had the exact same wording: It was not in quote marks and was attributed to a completely different group of authors to that acknowledged by the student. As with the first example above, there are several different sources for the same quote. Again this exact example has been used as an illustration and may fall into the area of common knowledge but, as with Example 1, similar cases were found that could not be defended on the grounds of common knowledge.

Example 6

In some instances the identified web site could not be accessed to check against the student’s work. This made it difficult, at times, to determine the category to place it in. In order to conduct the data analysis the tutor had to best guess the category in these cases and this is clearly a potential risk of error within the data analysis.

Example 7

One student wrote (underlined text shows the text highlighted by the detection service’s report): In regards to exercise prescription for the prevention of osteoporosis there is a wide variation that exists thus the exercise intervention is often poorly defined (thus not reproducible) or is not applicable to clinical practice (e.g., "walking 50 minutes on a treadmill at 70 per cent VO2 max"). Coupled with this the types of exercise and skeletal sites measured vary widely across studies therefore, making it difficult to find a certain exercise that could delay the symptoms of the disease. Despite these shortcomings, most studies show at least a trend toward improvement in such measures as falls, strength, and balance, as well as Bone Mass Development.

In this case and others like it, the section has a large but broken up part of it highlighted by the Plagiarism Detection Service. Although the highlighted text comes from the same source, the individual bits are found spread around the identified source, for example, some from the beginning and some from the end of the web page. Deciding if this should be considered as one or several instances of similarity was a potential area for inconsistency. To avoid this, one tutor completed all the analysis and attempted to be consistent in handling the reports, counting each instance if it had been taken from several different places within a source.

Given the above difficulties, analysis of the data was slower than anticipated and it was decided to perform preliminary analysis on the whole group and then in-depth analysis on a systematically selected sample of 15 from the 57 essays uploaded in the class (that is 26 per cent of the group). The three essays attaining the highest similarity with colour coding of yellow (25 – 49 per cent similarity) were analysed in-depth plus every fourth essay that was uploaded until a total of 15 had been analysed.

Analysis was completed in the following two ways:

- reviewing all sources and instances highlighted, including those from the Plagiarism Detection Service’s database, but which were not flawed (see discussion above); and
- reviewing all sources and instances highlighted, excluding those from the Plagiarism Detection Service’s database.

RESULTS

When reviewing the similarity index for all 57 students, the initial results shown in Figure 2 suggest no **major** plagiarism problems within the majority of the group, although the three in the higher index (25 – 49%) may have given rise for concern from these initial figures. The mean similarity index was found to be 9 per cent (±8%). Only further review would determine if their work was of concern.

The average number of matched sources (including database sources) per essay was 3.8 ± 3.5 (mean ± standard deviation) with a mode of three and the average number of sources (excluding database sources) per essay was 3.3 ± 3.4) with a mode of one. This finding suggests that most students were not heavily reliant on internet sources, or at least those internet sites identified by the detection service.

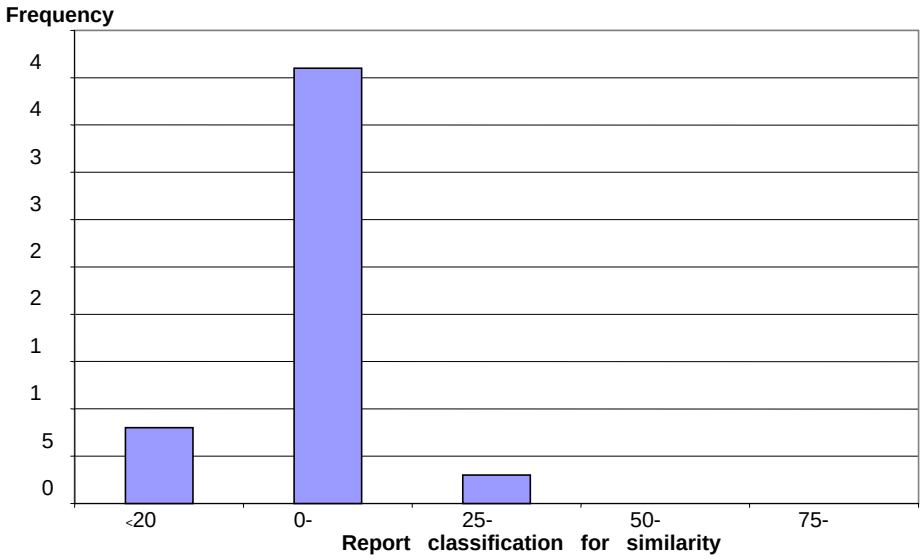

Figure 2. Frequency of students by similarity index

Table 1 and Table 2 show the in-depth data analysis for 15 (26%) of the group. It should be noted that the average number of sources identified was slightly higher for the sample than that for the group as a whole. It can be seen that students tended to use a source more than once within an essay by the fact that the average number of text highlights was greater than the average number of sources. Excluding the database sources did not change the plagiarism rates greatly although it did raise them very slightly.

Table 1. Data analysis on identified sources

Database matches	**Average number** sources identified per essay	showing minor plagiarism	showing plagiarism	**Plagiarism rate** not including minor plagiarism	including major and minor plagiarism
Including	4.8 ± 2.9	1.0 ± 1.6	1.8 ± 2.8	0.30	0.50
Excluding	4.1 ± 2.9	0.9 ± 1.4	1.7 ± 2.5	0.32	0.56

Table 2. Data analysis on instances of highlighted text

Database matches	**Average per essay** total text highlights	minor plagiarism	plagiarism	**Plagiarism rate** not including minor plagiarism	including major and minor plagiarism
Including	10.3 ± 10.0	2.9 ± 3.6	3.0 ± 5.0	0.19	0.46
Excluding	9.0 ± 8.8	2.5 ± 3.0	2.7 ± 4.4	0.22	0.49

Plagiarism rates were below one third when only considering plagiarism (but excluding minor plagiarism) although, the rate rose to around 50 per cent of identified sources and instances for all types of plagiarism (both 'major' and 'minor'). Plagiarism rates differed depending on whether identified sources or instances of highlighted text were considered. It should be remembered here that the rates are those of the highlighted text and are not rates for the essay as a whole.

Student Comments on the Service

Many students recognised the potential of such a service for prevention and as a deterrent. Several liked the idea of using it before handing in a final submission to check if the essay has any inadvertent problem text within it. The following are some examples of what students wrote about the service.

Example 1: "This program definitely has potential, but only highlighted quotes in my essay that had been correctly referenced. I believe JISC would reduce the risk of accidental plagiarism through simple human error."

Example 2: "… those individuals that are worried about plagiarism, and know their referencing ability is not always accurate would find the service beneficial to put their mind at ease."

Example 3: "Its ability to search all resources on the Internet was fantastic and in return gave me peace of mind."

Example 4: "It would keep students on their toes."

Example 5: One enthusiastic student even went so far as to say: "It helps to point out poor referencing and bad note taking styles. I hope it will become part of every essay procedure."

For some students, the report then prompted them to correct their work and add in a reference for highlighted text. In this context it is a useful teaching tool. The following examples highlight what students reported about using the feedback of the report.

Example 1: "When I got the report back it only highlighted one sentence…. I did go away and research this sentence and get a reference, even though I had written it off the top of my head."

Example 2: "Helped me to indicate an area where I need to add a reference."

Some astute students picked up on the limitations of the service. The following are examples of what students had to say on this aspect.

Example 1: "Some highlighted sentences that JISC gave me were not from the source where they believed it to come from. JISC also failed to pick up on sentences and quotes that came from many other sources that I had used and referenced correctly…."

Example 2: "… the essay obviously contains some work that was not my own i.e. other people's work which I had referenced, it seems like the database from which JISC compares work is not very large! My reference section contained over 20 references and the database did not pick up on a single one of them."

DISCUSSION

The present study resulted from an evaluation of the JISC Plagiarism Detection Service by staff at the University of Southampton. Data analysis was problematic and while it was hoped to find a true incidence rate for plagiarism by the fact that students uploaded their draft essays when they

had had no prior knowledge of the service, the finding of this study was that a true incidence rate may, in reality, never be found.

Analysis of data showed that the average number of internet sources identified by the service per essay was relatively low (approximately 3 to 4 sources per essay). This is likely to have been due to the fact that the students were all in their final year and tutors had stressed the importance of using journal articles for their work at this level. In addition, search strategies are taught in the first year and are then revisited at various stages throughout the degree course. During such sessions students are warned of the limitations of web sites as sources for information in as much that anyone is free to publish on the web and not all information is reliable. It would appear that students, therefore, are using the web sparingly by the time they reach the third year. It should also be remembered, however, that, at the time this work was undertaken, the service was not able to access some web sites, for example, through gateways, and so actual use of the internet for sources is likely to be higher than indicated by these results. Since most of the gateway sources would be for e-journals and other such reliable sites, the actual use of such sites would not generally be cause for concern unless there had been plagiarism of them. It must be acknowledged, of course, that not having these sites identified means that possible plagiarism of those sites could not be viewed and analysed.

In calculating plagiarism rates, it is clear that the rate is heavily dependent on what is analysed, and whether it is according to sources or instances. If further analysis were to be done at the level of counting the number of highlighted words of plagiarised text compared to the number of highlighted words not plagiarised, this may well result in yet another different statistic. The flaws and difficulties highlighted previously in analysing the reports make it extremely difficult to attain a consistent and comparable rate. Comparison of rates between cohorts and institutions can only be made if the same method and classifications are used throughout. For example Weinstein and Dobkin (2002) defined ‘small scale plagiarism’ as “material with either no attempt at citation or improper citation that composes less than 10% of the overall paper.” In the present study there was no percentage of the overall paper defined: sources and instances were either counted as not plagiarism, minor plagiarism, or plagiarism according the citation conventions expected by the tutors. Is it possible, therefore, that a true incidence rate for plagiarism will never be found. It not only depends on whether or not students have been warned of the use of the service (see introduction for argument of how this might affect incidence rates) but it also depends on the analysis undertaken. Until there is a common standard for analysis and a system that can trawl through **all** published material, then academia will never know the true incidence rate for plagiarism.

The question arises, therefore, how such a service is best used. The student comments would suggest that it may be best used as a teaching tool rather than as a policing service. Some students were quick to recognise the limitations of the report in so far as what the service was able to identify. If tutors wanted to use it as a detection tool, then they would need to hide its limitations, as they currently stand, from the students. This might mean having to deny them access to the returned report under normal circumstances, the exception being if it had been decided to take them to a plagiarism panel. The problem of ethics rears its head once again!

At the end of the day, one important question to answer is how many of the students would have been taken to a plagiarism panel if these essays had been the final submitted version? Of the 15 students viewed for in-depth analysis, all except one had some highlighted text. Of those 14 with highlighted text, only four showed no signs of any minor or major plagiarism. On the face of it, this looks alarming, but in viewing the sample group there was only one student who was likely to have been taken forward to a plagiarism panel. Interestingly, this was the student who had the highest similarity index at 33 per cent, had a high number of internet sources identified (12), had

37 instances of highlighted text of which 11 showed minor instances of plagiarism and 17 showed plagiarism.

CONCLUSION

In using the JISC Plagiarism Detection Service and trying to determine a true incidence rate for plagiarism, more questions have been raised than answered. It is clear that the originality report produced by such a service is only the starting point and that further detailed analysis by the tutor is required. This is acknowledged by JISC who state that "The report is, however, non-judgemental …. Academic judgement is still required to determine whether plagiarism has, in fact, occurred"(JISC Plagiarism Advisory Service pamphlet, 2003). The difficulty for the academic fraternity is deciding if a case for plagiarism should be taken forward to a panel. This must be based on quantifiable evidence and clearly defined criteria against which the work can be accurately and consistently measured. In the present study, plagiarised material was counted in terms of sources and instances, but where should the line be drawn for a student to be accused of plagiarism and then taken on to a panel? Is it if 10 per cent of sources are plagiarised, 20 per cent of the words, 50 per cent of the instances, or only if **major** plagiarism appears, whatever the definition of that may be? Institutions need to set policy giving clear definitions and statements regarding plagiarism in quantitative terms. The JISC Plagiarism Detection Service clearly has potential for identifying the amount of text to be further analysed and can help provide quantifiable evidence following further analysis. However, current limitations, and difficulties with analysis, added to the lack of defined policy within some institutions, means that at the moment it would probably be better used as a deterrent and in a teaching and learning capacity, to improve referencing techniques and develop good academic practice.

REFERENCES

Austin-Baker, R. (2003). Pondering the plagiarism plague 1. *The Times Higher Educational Supplement,* 1st Aug.

Franklyn-Stokes, A. and Newstead, S. E. (1995). Undergraduate cheating: Who does what and why? *Studies in Higher Education,* 20 (2), 159-172.

Gresham, J. (2002). Cyber-plagiarism: Technological and cultural background and suggested responses. *Catholic Library World,* 73 (1), 16-19.

JISC (Feb 2003). Plagiarism Advisory Service. (pamphlet), Northumbria University.

Jones, S. (2003). Education: On the prowl for copycats, *The Sunday Times,* 15th June.

Park, C. (2003). In other (people's) words: plagiarism by university students – literature and lessons, *Assessment and Evaluation in Higher Education,* 28 (5), 471-488.

Prime Ministers Special Report, BBC Radio 4 (2003). Details of the survey in The Guardian Rise Pages, *The Guardian (19 July, 2003)* [Online] http://web2.infotrac.galegroup.com/itw/infomark/379/643/70961337w2/purl=rc1_SP00_0_CJ105581670&dyn=9!xrn_2_0_CJ105581670?sw_aep=unisoton [Accessed: 31/8/05].

Weinstein, J. W. and Dobkin, C. E. (2002). Plagiarism in U.S. Higher Education: Estimating Internet Plagiarism Rates and Testing a Means of Deterrence, [Online]: http://64.233.183.104/search?q=cache:6Df0V-ze_N0J:webdisk.berkeley.edu/~Weinstein/Weinstein-JobMarketPaper.PDF+Weinstein+Dobkin+Plagiarism+Internet&hl=en [Accessed: 31/8/05].