The Art of Finding and Documenting Plagiarism Without Commercial Plagiarism Detection Software

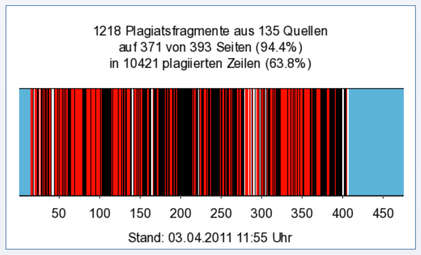
Debora Weber-Wulff

HTW Berlin, Germany

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

Introduction

Germany has been intensively discussing the topic of plagiarism in doctoral dissertations over the past two years. Starting with the dissertation of the defense minister Karl-Theodor zu Guttenberg in 2011, public documentation of plagiarism now extends to 46 additional works that encompass dissertations, a few habilitations, and a handbook on how to write professionally for law students.



Visualization of the zu Guttenberg thesis. Red means more than one plagiarism source on the page, black is one source.
CC-BY-SA, GuttenPlag Wiki

The collaborative and public documentations found on the GuttenPlag Wiki [1], the VroniPlag Wiki [2], and the Schavanplag blog [3] have led to eleven German dissertations being rescinded. The universities are still examining many of the other cases, sometimes taking more than a year to deliberate. Some dissertations have not been rescinded that have been extensively and publically documented, despite massive word-for-word text parallels that should be considered plagiarism (see in particular the cases Nk, Dd, and Jg on [2]).

Many people assume that commercial plagiarism detection software is used in discovering and documenting the plagiarism—however most of the work is done manually. This poster will attempt to explain how one can go about detecting plagiarism in a large thesis using primarily open source software tools.

The crowd

The three different platforms mentioned above all have different structures. The GuttenPlag Wiki had a core of activists somewhere between 50 and 100 strong. They were helped by an enormous crowd of transitory participants who contributed to the effort. They used an Internet Relay Chat (IRC) text chat system for real-time communication and documented the plagiarisms found in a commercial wiki run by the San Francisco-based company Wikia.

The VroniPlag Wiki has a smaller core of activists, between 10 and 20 persons, with occasional contributors. They, too, use a chat and a Wikia wiki for documenting purposes. They are still very active, having just published the 46th case in April 2013. The Schavanplag blog was put up by one individual activist from the VroniPlag Wiki group on a Wordpress blog and was occasionally assisted by persons from that group.

Dissertations to be examined by VroniPlag Wiki are suggested anonymously by the general public, or found by chance, or are cases in which a whistleblower knows possible sources for the thesis and informs the group.

Obtaining the thesis

German doctoral dissertations must be published in some form or another. They are filed with the German National Library [4] and are kept on deposit in the university library of the university at which the dissertation was submitted, as well as in other libraries throughout Germany. The union catalogue for the national library is an invaluable tool for finding both theses and possible sources.

Reading

Once the thesis has been obtained, the first step should be to read it. As one reads, it is important to be sensitive to the tone and style of the writing.

- Does it proceed smoothly from topic to topic, or does the style jump from journalistic copy to dry scientific writing with interspersed sentences that are not grammatically correct?
- Are verbs missing in a sentence or are there interesting misspellings?
- Is the choice of wording or style not consistent with the author?
- Does the format or font change erratically?
- Are there odd words underlined? These could be links from an online source.

Any of these could be a sign of possible plagiarism.

Google

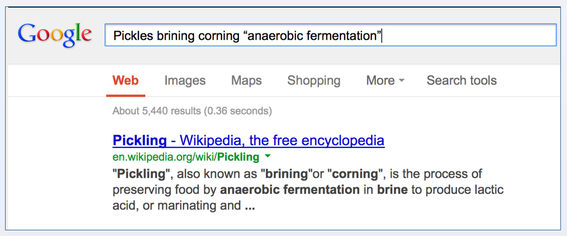
The first step should be to check for possible sources using a search machine such as Google. Good search terms include:

- a **sentence** from a suspicious paragraph,
- the **misspelled word** with other nearby words,
- a **well-formed phrase**, or just
- **three to five uncommon words** from the same paragraph.

Kosher Pickles

Pickling, also known as brining or corning is the process of preserving food by anaerobic fermentation in brine (a solution of salt in water) to produce lactic acid, or marinating and storing it in an acid solution, usually vinegar (acetic acid). The resulting food is called a pickle. This procedure gives the food a salty or sour taste. In South Asia edible oils are used as the pickling medium instead of vinegar. The term pickle is derived from the Dutch word pekel, meaning brine.

Student paper



Five words in Google finds a source

Just using a few words and not a longer phrase will also find a source even if the word order has been changed, words have been removed, or additional words inserted. After a general check on the Google page, it can be advantageous to check out Google Scholar or Google Books. The results from searches in these materials are not always included in the normal search. Especially Google Books can be quite fruitful, even if the result is only a snippet, a short excerpt from a book. If a match can be found, then that particular book can also be obtained from a library as a candidate source.

Other good candidates can be found by following the footnotes and references. Many authors do give pointers to their sources, they just don't bother to clearly delineate the beginning and end of the text taken. And sometimes the wording is far too close, or even identical. This is called a "pawn sacrifice" [5]. One can also attempt a systematic "brute force", digitizing and comparing all of the sources listed in the bibliography with the thesis. This is rather time-consuming, but can unfortunately often be fruitful.

Digitization

Once a text parallel has been found in a source, it can be useful to check the entire chapter surrounding that fragment, or even the entire book. Many libraries today have book scanners that will take pictures of book pages as they lie flat on a cradle. They will even separate the two pages into individual pages and some systems will detect and remove pictures of the fingers holding down a page. The resulting pictures can then be stored on a USB stick. With a bit of practice, one can easily scan up to 300 pages an hour.

The next step is to recognize the text from the pictures. This step is called **O**ptical **C**haracter **R**ecognition. There are many systems available, The Abbyy FineReader is not free, but it gives excellent results and produces a text file from the pictures. Once a text is available for the thesis and a candidate source, a comparison can be made.

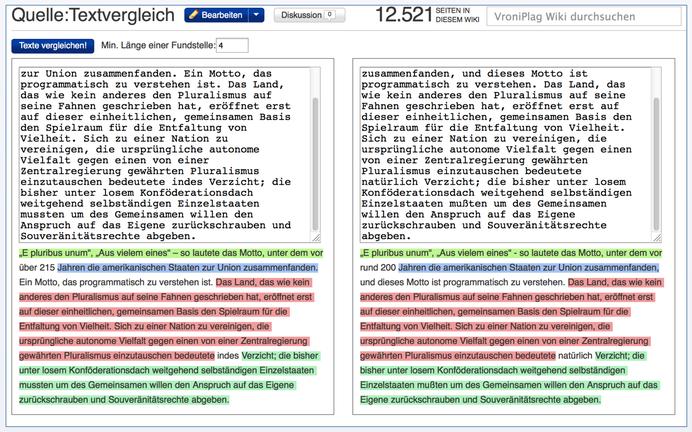
Comparison

There are many simple comparison tools that can be used to compare a text with a candidate source—or even with itself, in order to find duplicated matter. One such system is SIM, an open algorithm described by Gruene & Huntjens [6]. The VroniPlag Wiki group offers a version programmed in JavaScript that can be used free of charge [7]. An example of a simple comparison is shown in the next column.

Documentation

The most time-consuming part of dealing with plagiarism is the documentation. It is not sufficient to just state that a paper or a book is a plagiarism, it must be sufficiently documented so that others can examine the fragments in order to decide if they agree with the conclusions or not.

A side-by-side documentation with the work being examined on the left and the source on the right has proven to be effective. However, page numbers and line numbers must be included, in order to enable independent confirmation that the digital version is indeed a true copy of the printed work. The GuttenPlag Wiki and VroniPlag Wiki groups insist on each fragment being judged by more than one person before it is listed as being a text parallel.



K.-T. zu Guttenberg, *Verfassung und Verfassungsvertrag: Konstitutionelle Entwicklungsstufen in den USA und der EU*, Duncker & Humblot: Berlin, 2009, S. 15 vs. B. Zehnpfennig, "Das Experiment einer großräumigen Republik", *FAZ*, Nov. 27, 1997, http://www.faz.net/aktuell/politik/das-experiment-einer-grossraeumigen-republik-1590883.html

Why not use commercial plagiarism detection software?

This question is often asked, as there seems to be a general belief in the powers of software to solve practically any problem. The problem is that plagiarism is not just a simple copy & paste job that would be relatively easy to detect. There are many kinds of plagiarism that include rewriting a statement but ignoring the author of the statement.

The author has tested such software five times previously [8], and has found a number of issues that preclude their general use:

- **False positives**: Correct citations, trivial phrases, or even just wrong results report a plagiarism where in fact there is none.
- **False negatives**: Many systems report no plagiarism where in fact there is much – it may just be from a database that a search engine does not crawl, or the sample investigated was too small to find the true source.
- **Umlauts**: Many systems cannot deal with the character sets found in countries that use other languages than English.
- **Google Books**: Material that can easily be located using Google Books is often not found using these systems.
- **Link rot**: Systems report plagiarism from links that no longer exist online, or for which the URL has changed.
- **Copies kept**: Many systems keep copies of material submitted and do not make it clear that they are doing so.

Citations and Links

[1] GuttenPlag Wiki: http://de.guttenplag.wikia.com/wiki/GuttenPlag_Wiki (in German)

[2] VroniPlag Wiki: http://de.vroniplag.wikia.com/wiki/Home (in German)

[3] Schavanplag: http://schavanplag.wordpress.com/ (in German)

[4] German National Library: http://www.dnb.de/EN/Home/home_node.html

[5] Lahusen, B. (2006). Goldene Zeiten - Anmerkungen zu Hans-Peter Schwintowski, Juristische Methodenlehre, UTB basics Recht und Wirtschaft 2005. In: *Kritische Justiz*, 39(4) pp. 398-405

[6] Grune, D. & Huntjens, M. (1989), Het detecteren van kopieën bij informatica-practica, In: *Informatie*, 31(11), pp. 864-867 (English translation: http://dickgrune.com/Programs/similarity_tester/Paper.ps)

[7] Text comparison tool: http://de.vroniplag.wikia.com/wiki/Quelle:Textvergleich

[8] Tests results on the effectiveness of plagiarism detection software http://plagiat.htw-berlin.de/software-en/

Contact

Debora Weber-Wulff, HTW Berlin, FB4, Treskowallee 8, 10318 Berlin, Germany
Email: weberwu@htw-berlin.de

- **Portal Plagiat**: A collection of material about plagiarism, including the E-Learning unit "Fremde Federn Finden" http://plagiat.htw-berlin.de (in German)
- **Copy, Shake & Paste** – A blog about plagiarism and scientific misconduct http://copy-shake-paste.blogspot.de/ (in English)