Crowdsourcing Interaction Logs to Understand Text Reuse from the Web

Martin Potthast **Matthias Hagen** **Michael Völske** **Benno Stein**

Bauhaus-Universität Weimar
99421 Weimar, Germany
`<first name>.<last name>@uni-weimar.de`

Abstract

We report on the construction of the Webis text reuse corpus 2012 for advanced research on text reuse. The corpus compiles manually written documents obtained from a completely controlled, yet representative environment that emulates the web. Each of the 297 documents in the corpus is about one of the 150 topics used at the TREC Web Tracks 2009–2011, thus forming a strong connection with existing evaluation efforts. Writers, hired at the crowdsourcing platform oDesk, had to retrieve sources for a given topic and to reuse text from what they found. Part of the corpus are detailed interaction logs that consistently cover the search for sources as well as the creation of documents. This will allow for in-depth analyses of *how text is composed* if a writer is at liberty to reuse texts from a third party—a setting which has not been studied so far. In addition, the corpus provides an original resource for the evaluation of text reuse and plagiarism detectors, where currently only less realistic resources are employed.

1 Introduction

The web has become one of the most common sources for text reuse. When reusing text from the web, humans may follow a three step approach shown in Figure 1: searching for appropriate sources on a given topic, copying of text from selected sources, modification and paraphrasing of the copied text. A considerable body of research deals with the detection of text reuse, and, in particular, with the detection of cases of plagiarism (i.e., the reuse of text with the intent of disguising the fact that text has been reused). Similarly, a large number of commercial software systems is being developed whose purpose is the detection of plagiarism. Both the developers of these systems as well as researchers working on the subject matter frequently claim their approaches to be searching the entire web or, at least, to be scalable to web size. However, there is hardly any evidence to substantiate this claim—rather the opposite can be observed: commercial plagiarism detectors have not been found to reliably identify plagiarism from the web (Köhler and Weber-Wulff, 2010), and the evaluation of research prototypes even under laboratory conditions shows that there is still a long way to go (Potthast et al., 2010b). We explain the disappointing state of the art by the lack of realistic, large-scale evaluation resources.

With our work, we want to contribute to closing the gap. In this regard the paper in hand introduces the Webis text reuse corpus 2012 (Webis-TRC-12), which, for the first time, emulates the *entire process* of reusing text from the web, both at scale and in a controlled environment. The corpus comprises a number of features that set it apart from previous ones: (1) the topic of each document in the corpus is derived from a topic of the TREC Web Track, and the sources to copy from have been retrieved manually from the ClueWeb corpus. (2) The search for sources is logged, including click-through and browsing data. (3) A fine-grained edit history has been recorded for each document. (4) A total of 297 documents were written with an average length of about 5700 words, whereas diversity is ensured via crowdsourcing. Altogether, this corpus forms the current most realistic sample of writers reusing text. The corpus is publicly available.[1]

1.1 Related Work

As organizers of the annual PAN plagiarism detection competitions,[2] we have introduced the first standardized evaluation framework for that pur-

[1] http://www.webis.de/research/corpora

[2] http://pan.webis.de

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1212–1221,
Sofia, Bulgaria, August 4-9 2013.

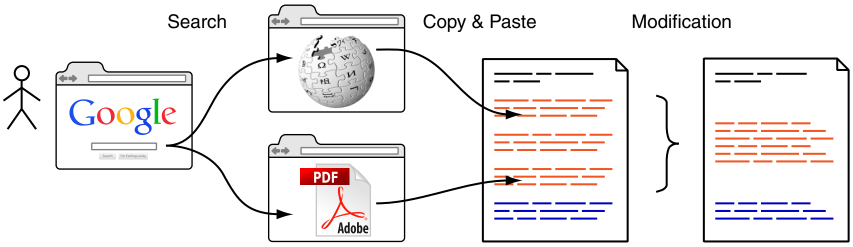

Figure 1: The basic steps of reusing text from the web (Potthast, 2011).

pose (Potthast et al., 2010b). Among others, it comprises a series of corpora that consist of automatically generated cases of plagiarism, provided in the form of the PAN plagiarism corpora 2009-2011. The corpora have been used to evaluate dozens of plagiarism detection approaches within the respective competitions in these years;[3] but even though they have been adopted by the community, a number of shortcomings render them less realistic:

1. All plagiarism cases were generated by randomly selecting text passages from documents and inserting them at random positions in a host document. This way, the reused passages do not match the topic of the host document.

2. The majority of the reused passages were modified in order to obfuscate the reuse. However, the applied modification strategies, again, are basically random: shuffling, replacing, inserting, or deleting words randomly. An effort was made to avoid non-readable text, yet none of it bears any semantics.

3. The corpus documents are parts of books from the Project Gutenberg. Many of these books are pretty old, whereas today the web is the predominant source for text reuse.

To overcome the second issue, about 4 000 passages were rewritten manually via crowdsourcing on Amazon's Mechanical Turk for the 2011 corpus. But, because of the first issue (random passage insertion), a topic drift analysis can spot a reused passage more easily than a search within the document set containing the original source (Potthast et al., 2011). From these observations it becomes clear that there are limits for the automatic construction of such kinds of corpora. The Webis text reuse corpus 2012 addresses all of the mentioned issues since it has been constructed manually.

Besides the PAN corpora, there are two other corpora that comprise "genuinely reused" text: the Clough09 corpus, and the Meter corpus. The former corpus consists of 57 answers to one of five computer science questions that were reused from a respective Wikipedia article (Clough and Stevenson, 2011). While the text was genuinely written by a number of volunteer students, the choice of topics is narrow, and text lengths range from 200 to 300 words, which is hardly more than 2-3 paragraphs. Also, the sources from which text was reused were given up front, so that there is no data about their retrieval. The Meter corpus annotates 445 cases of text reuse among 1 716 news articles (Clough et al., 2002). The cases of text reuse in this corpus are realistic for the news domain; however, they have not been created by the reuse process outlined in Figure 1. Note that in the news domain, text is often reused directly from a news wire without the need for retrieval. Our new corpus complements these two resources.

2 Corpus Construction

Two data sets form the basis for constructing our corpus, namely (1) a set of topics to write about and (2) a set of web pages to research about a given topic. With regard to the former, we resort to topics used at TREC, specifically to those used at the Web Tracks 2009–2011. With regard to the latter, we employ the ClueWeb corpus from 2009[4] (and not the "web in the wild"). The ClueWeb comprises more than one billion documents from ten languages and can be considered as a representative cross-section of the real web. It is a widely accepted resource among researchers and became one of the primary resources to evaluate the retrieval performance of search engines within several TREC tracks. Our corpus's strong connection to TREC will allow for unforeseen synergies. Based on these decisions our

[3] See (Potthast et al., 2009; Potthast et al., 2010a; Potthast et al., 2011) for overviews of approaches and evaluation results of each competition.

[4] http://lemurproject.org/clueweb09

corpus construction steps can be summarized as follows:

1. Rephrasing of the 150 topics used at the TREC Web Tracks 2009–2011 so that they explicitly invite people to write an essay.
2. Indexing of the ClueWeb corpus category A (the entire English portion with about 0.5 billion documents) using the BM25F retrieval model plus additional features.
3. Development of a search interface that allows for answering queries within milliseconds and that is designed along the lines of commercial search interfaces.
4. Development of a browsing API for the ClueWeb, which serves ClueWeb pages on demand and which rewrites links of delivered pages, now pointing to their corresponding ClueWeb pages on our servers (instead of to the originally crawled URL).
5. Recruiting 27 writers, 17 of whom with a professional writing background, hired at the crowdsourcing platform oDesk from a wide range of hourly rates for diversity.
6. Instructing the writers to write one essay at a time of at least 5000 words length (corresponding to an average student's homework assignment) about an open topic of their choice, using our search engine—hence browsing only ClueWeb pages.
7. Logging all writers' interactions with the search engine and the ClueWeb on a per-essay basis at our site.
8. Logging all writers' edits to their essays in a fine-grained edit log: a snapshot was taken whenever a writer stopped writing for more than 300ms.
9. Double-checking all of the essays for quality.

After having deployed the search engine and completed various usability tests, the actual corpus construction took nine months, from April 2012 through December 2012.

Obviously, the outlined experimental setup can serve different lines of research and is publicly available as well. The remainder of the section presents elements of our setup in greater detail.

2.1 Topic Preparation

Since the topics used at the TREC Web Tracks were not amenable for our purpose as is, we rephrased them so that they ask for writing an essay instead of searching for facts. Consider for example topic 001 of the TREC Web Track 2009:

> *Query.* obama family tree
>
> *Description.* Find information on President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc.
>
> *Sub-topic 1.* Find the TIME magazine photo essay "Barack Obama's Family Tree."
>
> *Sub-topic 2.* Where did Barack Obama's parents and grandparents come from?
>
> *Sub-topic 3.* Find biographical information on Barack Obama's mother.

This topic is rephrased as follows:

> *Obama's family*. Write about President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. Where did Barack Obama's parents and grandparents come from? Also include a brief biography of Obama's mother.

In the example, Sub-topic 1 is considered too specific for our purposes while the other sub-topics are retained. TREC Web Track topics divide into faceted and ambiguous topics. While topics of the first kind can be directly rephrased into essay topics, from topics of the second kind one of the available interpretations was chosen.

2.2 A Controlled Web Search Environment

To give the oDesk writers a familiar search experience while maintaining reproducibility at the same time, we developed a tailored search engine called ChatNoir (Potthast et al., 2012b).[5] Besides ours, the only other public search engine for the ClueWeb is Carnegie Mellon's Indri,[6] which, unfortunately, is far from our efficiency requirements. Moreover, its search interface does not follow the standard in terms of result page design, and it does not give access to interaction logs. Our search engine is on the order of milliseconds in terms of retrieval

[5] http://chatnoir.webis.de

[6] http://lemurproject.org/clueweb09.php/index.php#Services

time, its interface follows industry standards, and it features an API that allows for user tracking.

ChatNoir is based on the BM25F retrieval model (Robertson et al., 2004), uses the anchor text list provided by (Hiemstra and Hauff, 2010), the PageRanks provided by the Carnegie Mellon University alongside the ClueWeb corpus, and the Spam rank list provided by (Cormack et al., 2011). ChatNoir comes with a proximity feature with variable-width buckets as described by (Elsayed et al., 2011). Our choice of retrieval model and ranking features is intended to provide a reasonable baseline performance. However, it is neither near as mature as those of commercial search engines nor does it compete with the best-performing models from TREC. Yet, it is among the most widely accepted models in information retrieval, which underlines our goal of reproducibility.

In addition to its retrieval model, ChatNoir implements two search facets: text readability scoring and long text search. The first facet, similar to that provided by Google, scores the readability of a text found on a web page via the well-known Flesch-Kincaid grade level formula (Kincaid et al., 1975): it estimates the number of years of education required in order to understand a given text. This number is mapped onto the three categories "Simple" (up to 5 years), "Intermediate" (between 5 and 9 years) and "Expert" (at least 9 years). The "Long Text" search facet omits search results which do not contain at least one continuous paragraph of text that exceeds 300 words. The two facets can be combined with each other.

When clicking on a search result, ChatNoir does not link into the real web but redirects into the ClueWeb. Though the ClueWeb provides the original URLs from which the web pages have been obtained, many of these pages have gone or been updated since. We hence set up an API that serves web pages from the ClueWeb on demand: when accessing a web page, it is pre-processed before being shipped, removing automatic referrers and replacing all links to the real web with links to their counterpart inside the ClueWeb. This way, the ClueWeb can be browsed as if surfing the real web, whereas it becomes possible to track a user. The ClueWeb is stored in the HDFS of our 40 node Hadoop cluster, and web pages are fetched directly from there with latencies of about 200ms. ChatNoir's inverted index has been optimized to guarantee fast response times, and it is deployed alongside Hadoop on the same cluster.

Table 1: Demographics of the 12 Batch 2 writers.

Writer Demographics					
Age		*Gender*		*Native language(s)*	
Minimum	24	Female	67%	English	67%
Median	37	Male	33%	Filipino	25%
Maximum	65			Hindi	17%
Academic degree		*Country of origin*		*Second language(s)*	
Postgraduate	41%	UK	25%	English	33%
Undergraduate	25%	Philippines	25%	French	17%
None	17%	USA	17%	Afrikaans, Dutch, German, Spanish, Swedish each	8%
n/a	17%	India	17%		
		Australia	8%		
		South Africa	8%	None	8%
Years of writing		*Search engines used*		*Search frequency*	
Minimum	2	Google	92%	Daily	83%
Median	8	Bing	33%	Weekly	8%
Standard dev.	6	Yahoo	25%	n/a	8%
Maximum	20	Others	8%		

2.3 Two Batches of Writing

In order to not rely only on the retrieval model implemented in our controlled web search environment, we divided the task into two batches, so that two essays had to be written for each of the 150 topics, namely one in each batch. In Batch 1, our writers did not search for sources themselves, but they were provided up front with an average of 20 search results to choose from for each topic. These results were obtained from the TREC Web Track relevance judgments (so-called "qrels"): only documents that were found to be relevant or key documents for a given topic by manual inspection of the NIST assessors were provided to our writers. These documents result from the combined wisdom of all retrieval models of the TREC Web Tracks 2009–2011, and hence can be considered as optimum retrieval results produced by the state of the art in search engine technology. In Batch 2, in order to obtain realistic search interaction logs, our writers were instructed to search for source documents using ChatNoir.

2.4 Crowdsourcing Writers

Our ideal writer has experience in writing, is capable of writing about a diversity of topics, can complete a text in a timely manner, possesses decent English writing skills, and is well-versed in using the aforementioned technologies. After bootstrapping our setup with 10 volunteers recruited at our university, it became clear that, because of the workload involved, accomplishing our goals would not be possible with volunteers only. Therefore, we resorted to hiring (semi-)professional writers and made use of the crowdsourcing platform oDesk.[7] Crowdsourcing has quickly become one of the

[7] http://www.odesk.com

Table 2: Key figures of the Webis text reuse corpus 2012.

Corpus	**Distribution**				**Total**
characteristic	min	avg	max	stdev	
Writers	*(Batch* 1+2*)*				27
Essays (Topics)	*(Two essays per topic)*				297 (150)
Essays / Writer	1	2	66	15.9	
Queries	*(Batch* 2*)*				13 655
Queries / Essay	4	91.0	616	83.1	
Clicks	*(Batch* 2*)*				16 739
Clicks / Essay	12	111.6	443	80.3	
Clicks / Query	1	2.3	76	3.3	
Irrelevant	*(Batch* 2*)*				5 962
Irrelevant / Essay	1	39.8	182	28.7	
Irrelevant / Query	0	0.5	60	1.4	
Relevant	*(Batch* 2*)*				251
Relevant / Essay	0	1.7	7	1.5	
Relevant / Query	0	0.0	4	0.2	
Key	*(Batch* 2*)*				1 937
Key / Essay	1	12.9	46	7.5	
Key / Query	0	0.2	22	0.7	

Corpus	**Distribution**				**Total**
characteristic	min	avg	max	stdev	
Search Sessions	*(Batch* 2*)*				931
Sessions / Essay	1	12.3	149	18.9	
Days	*(Batch* 2*)*				201
Days / Essay	1	4.9	17	2.7	
Hours	*(Batch* 2*)*				2 068
Hours / Writer	3	129.3	679	167.3	
Hours / Essay	3	7.5	10	2.5	
Edits	*(Batch* 1+2*)*				633 334
Edits / Essay	45	2 132.4	6 975	1 444.9	
Edits / Day	5	2 959.5	8 653	1 762.5	
Words	*(Batch* 1+2*)*				1 704 354
Words / Essay	260	5 738.8	15 851	1 604.3	
Words / Writer	2 078	63 124.2	373 975	89 246.7	
Sources	*(Batch* 1+2*)*				4 582
Sources / Essay	0	15.4	69	10.0	
Sources / Writer	5	169.7	1 065	269.6	

cornerstones for constructing evaluation corpora, which is especially true for paid crowdsourcing. Compared to Amazon's Mechanical Turk (Barr and Cabrera, 2006), which is used more frequently than oDesk, there are virtually no workers at oDesk submitting fake results because of its advanced rating features for workers and employers. Moreover, oDesk tracks their workers by randomly taking screenshots, which are provided to employers in order to check whether the hours logged correspond to work-related activity. This allowed us to check whether our writers used our environment instead of other search engines and editors.

During Batch 2, we have conducted a survey among the twelve writers who worked for us at that time. Table 1 gives an overview of the demographics of these writers, based on a questionnaire and their resumes at oDesk. Most of them come from an English-speaking country, and almost all of them speak more than one language, which suggests a reasonably good education. Two thirds of the writers are female, and all of them have years of writing experience. Hourly wages were negotiated individually and range from 3 to 34 US dollars (dependent on skill and country of residence), with an average of about 12 US dollars. For ethical reasons, we payed at least the minimum wage of the respective countries involved. In total, we spent 20 468 US dollars to pay the writers—an amount that may be considered large compared to other scientific crowdsourcing efforts from the literature, but small in terms of the potential of crowdsourcing to make a difference in empirical science.

3 Corpus Analysis

This section presents selected results of a preliminary corpus analysis. We overview the data and shed some light onto the search and writing behavior of writers.

3.1 Corpus Statistics

Table 2 shows key figures of the collected interaction logs, including the absolute numbers of queries, relevance judgments, working times, number of edits, words, and retrieved sources, as well as their relation to essays, writers, and work time, where applicable. On average, each writer wrote 2 essays while the standard deviation is 15.9, since one very prolific writer managed to write 66 essays.

From a total of 13 655 queries submitted by the writers within Batch 2, each essay got an average of 91 queries. The average number of results clicked per query is 2.3. For comparison, we computed the average number of clicks per query in the AOL query log (Pass et al., 2006), which is 2.0. In this regard, the behavior of our writers on individual queries does not differ much from that of the average AOL user in 2006. Most of the clicks that we recorded are search result clicks, whereas 2 457 of them are browsing clicks on web page links. Among the browsing clicks, 11.3% are clicks on links that point to the same web page (i.e., anchor links using the hash part of a URL). The longest click trail contains 51 unique web pages, but most trails are very short. This is a surprising result, since we expected a larger proportion of browsing clicks, but it also shows that our writers relied heavily on the ChatNoir's ranking. Regarding search facets, we observed that our writers used them only for about 7% of their queries. In these cases, the writers used either the "Long Text" facet, which retrieves web pages containing at least one continuous passage of at least 300 words, or set the desired reading level to "Expert."

The query log of each writer in Batch 2 divides into 931 search sessions with an average of 12.3 sessions per topic. Here, a session is defined as a sequence of queries recorded for a given topic which is not divided by a break longer than 30 minutes. Despite other claims in the literature (Jones and Klinkner, 2008; Hagen et al., 2013) we argue that, in our case, sessions can be reliably identified by timeouts because we have a priori knowledge about which query belongs to which essay. Typically, completing an essay took 4.9 days, which includes to a long-lasting exploration of the topic at hand.

The 297 essays submitted within the two batches were written with a total of 633 334 edits. Each topic was edited 2 132 times on average, whereas the standard deviation gives an idea about how diverse the modifications of the reused text were. Writers were not specifically instructed to modify a text as much as possible—rather they were encouraged to paraphrase in order to foreclose the detection by an automatic text reuse detector. This way, our corpus captures each writer's idea of the necessary modification effort to accomplish this goal. The average lengths of the essays is 5 739 words, but there are also some short essays if hardly any useful information could be found on the respective topics. About 15 sources have been reused in each essay, whereas some writers reused text from as many as 69 unique documents.

3.2 Relevance Judgments

In the essays from Batch 2, writers reused texts from web pages they found during their search. This forms an interesting relevance signal which allows us to separate web pages relevant to a given topic from those which are irrelevant. Following the terminology of TREC, we consider web pages from which text is reused as *key documents* for the respective essay's topic, while web pages that are on a click trail leading to a key document are termed *relevant*. The unusually high number of key documents compared to relevant documents is explained by the fact that there are only few click trails of this kind, whereas most web pages have been retrieved directly. The remainder of web pages that were viewed but discarded by our writers are considered as irrelevant.

Table 3: Confusion matrix of TREC judgments versus writer judgments.

TREC	**Writer judgment**			
judgment	irrelevant	relevant	key	unjudged
spam (-2)	3	0	1	2 446
spam (-1)	64	4	18	16 657
irrelevant (0)	219	13	73	33 567
relevant (1)	114	8	91	10 676
relevant (2)	44	5	56	3 711
key (3)	12	0	8	526
unjudged	5 506	221	1 690	–

Each year, the NIST assessors employed for the TREC conference manually review hundreds of web pages that have been retrieved by experimental retrieval systems that are submitted to the various TREC tracks. This was also the case for the TREC Web Tracks from which the topics of our corpus are derived. We have compared the relevance judgments provided by TREC for these tracks with the implicit judgments from our writers. Table 3 contrasts the two judgment scales in the form of a confusion matrix. TREC uses a six-point Likert scale ranging from -2 (extreme Spam) to 3 (key document). For 733 of the documents visited by our writers, TREC relevance judgments can be found. From these, 456 documents (62%) have been considered irrelevant for the purposes of reuse by our writers, however, the TREC assessor disagree with this judgment in 170 cases. Regarding the documents considered as key documents for reuse by our writers, the TREC assessors disagree on 92 of the 247 documents. An explanation for the disagreement can be found in the differences between the TREC ad hoc search task and our text reuse task: the information nuggets (small chunks of text) that satisfy specific factual information needs from the original TREC topics are not the same as the information "ingots" (big chunks of text) that satisfy our writers' needs.

3.3 Research Behavior

To analyze the writers' search behavior during essay writing in Batch 2, we have recorded detailed search logs of their queries while they used our search engine. Figure 2 shows for each of the 150 essays of this batch a curve of the percentage of queries at times between a writer's first query and an essay's completion. We have normalized the time axis and excluded working breaks of more

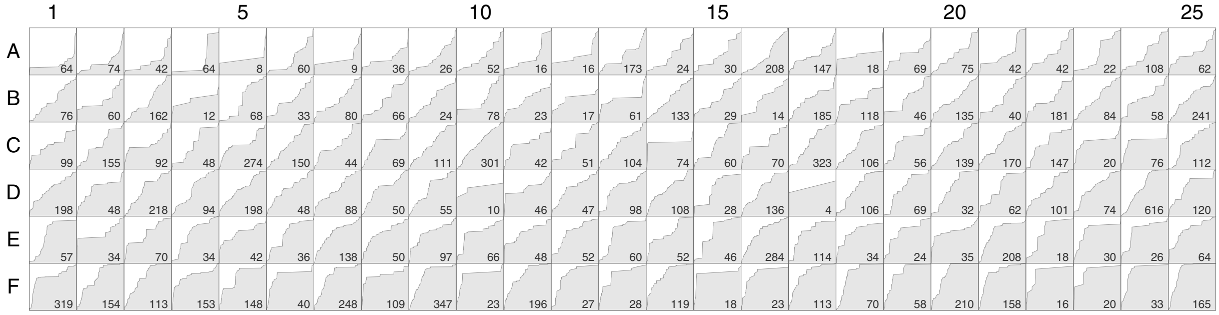

Figure 2: Spectrum of writer search behavior. Each grid cell corresponds to one of the 150 essays of Batch 2 and shows a curve of the percentage of submitted queries (y-axis) at times between the first query until the essay was finished (x-axis). The numbers denote the amount of queries submitted. The cells are sorted by area under the curve, from the smallest area in cell A1 to the largest area in cell F25.

than five minutes. The curves are organized so as to highlight the spectrum of different search behaviors we have observed: in row A, 70-90% of the queries are submitted toward the end of the writing task, whereas in row F almost all queries are submitted at the beginning. In between, however, sets of queries are often submitted in the form of "bursts," followed by extended periods of writing, which can be inferred from the steps in the curves (e.g., cell C12). Only in some cases (e.g., cell C10) a linear increase of queries over time can be observed for a non-trivial amount of queries, which indicates continuous switching between searching and writing. From these observations, it can be inferred that our writers sometimes conducted a "first fit" search and reused the first texts they found easily. However, as the essay progressed and the low hanging fruit in terms of search were used up, they had to search more intensively in order to complete their essay. More generally, this data gives an idea of how humans perform exploratory search in order to learn about a given topic. Our current research on this aspect focuses on the prediction of search mission types, since we observe that the search mission type does not simply depend on the writer or the perceived topic difficulty.

3.4 Visualizing Edit Histories

To analyze the writers' writing style, that is to say, how writers reuse texts and how the essay is completed in both batches, we have recorded the edit logs of their essays. Whenever a writer stopped writing for more than 300ms, a new edit was stored in a version control system at our site. The edit logs document the entire text evolution, from first the keystroke until an essay was completed. We have used the so-called history flow visualization to analyze the writing process (Viégas et al., 2004). Figure 3 shows four examples from the set of 297 essays. Based on these visualizations, a number of observations can be made. In general, we identify two distinct writing-style types to perform text reuse, namely to *build up* an essay during writing, or, to first gather material and then to *boil down* a text until the essay is completed. Later in this section, we will analyze this observation in greater detail. Within the plots, a number of events can be spotted that occurred during writing: in the top left plot, encircled as area A, the insertion of a new piece of text can be observed. Though marked as original text at first, the writer worked on this passage and then revealed that it was reused from another source. At area B in the top right plot, one can observe the reorganization of two passages as they exchange places from one edit to another. Area C in the bottom right plot shows that the writer, shortly before completing this essay, reorganized substantial parts. Area D in the same plot shows how the writer went about boiling down the text by incorporating contents from different passages that have been collected beforehand and, then, from one edit to another, discarded most of the rest. The saw-tooth shaped pattern in area E in the bottom left plot reveals that, even though the writer of this essay adopts a build-up style, she still pastes passages from her sources into the text one at a time, and then individually boils down each. Our visualizations also include information about the text positions where writers have been working at a given point in time; these positions are shown as blue dots in the plots. In this regard distinct writing patterns are discernible of writers who go through a text linearly versus those who do not. Future work will include an analysis of these writing patterns.

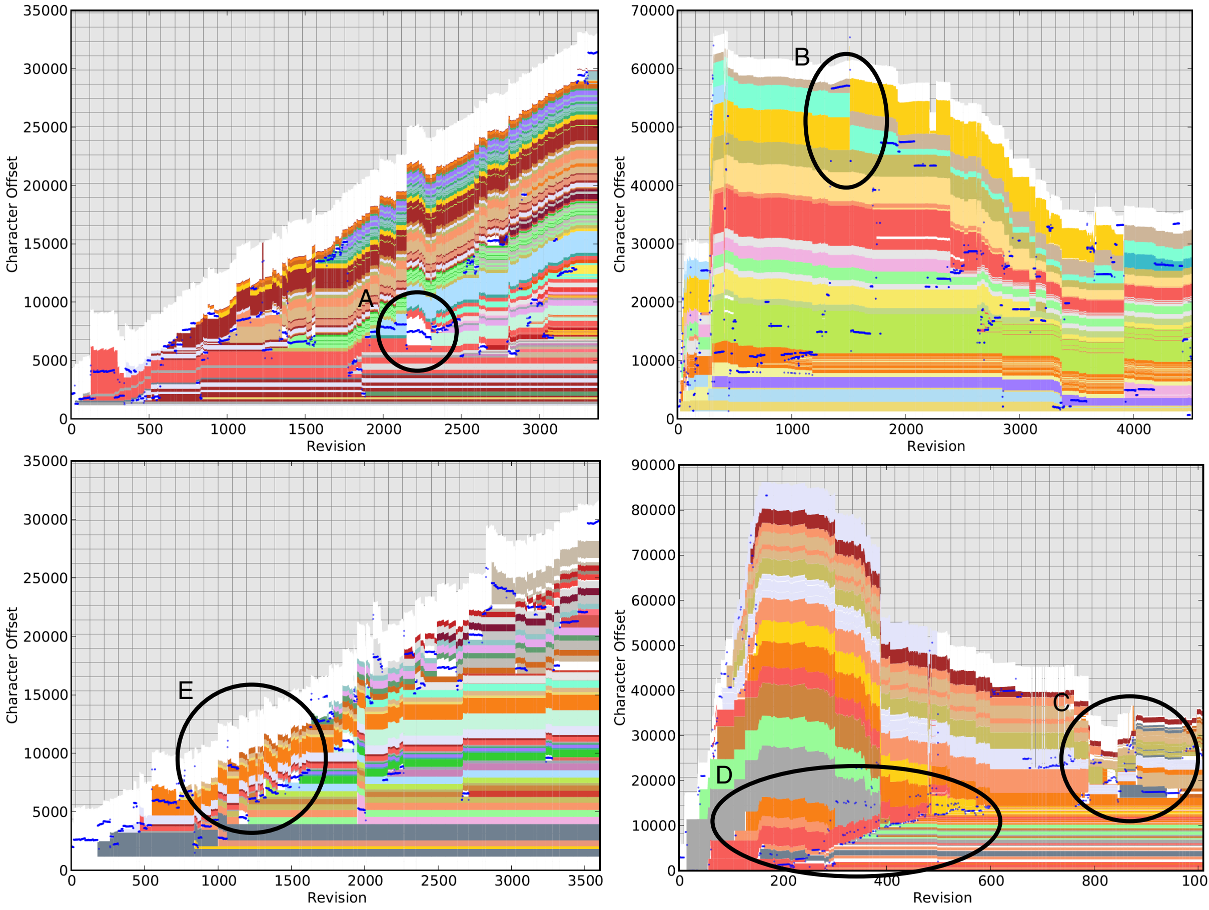

Figure 3: Types of text reuse: build-up reuse (left) versus boil-down reuse (right). Each plot shows the text length at text edit between first keystroke and essay completion; edits have been recorded during writing whenever a writer stopped for more than 300ms. Colors encode different source documents. Original text is white; blue dots indicate the text position of the writer's last edit.

3.5 Build-up Reuse versus Boil-down Reuse

Based on the edit history visualizations, we have manually classified the 297 essays of both batches into two categories, corresponding to the two styles build-up reuse and boil-down reuse. We found that 40% are instances of build-up reuse, 45% are instances of boil-down reuse, and 13% fall in between, excluding 2% of the essays as outliers due to errors or for being too short. The in-between cases show that a writer actually started one way and then switched to the respective other style of reuse so that the resulting essays could not be attributed to a single category. An important question that arises out of this observation is whether different writers habitually exert different reuse styles or whether they apply them at random. To obtain a better overview, we envision the applied reuse style of an essay by the skyline curve of its edit history visualization (i.e., by the curve that plots the length of an essay after each edit). Aggregating these curves on a per-writer basis reveals distinct patterns. For eight of our writers Figure 4 shows this characteristic. The plots are ordered by the shape of the averaged curve, starting from a linear increase (left) to a compound of steep increase to a certain length after which the curve levels out (right). The former shape corresponds to writers who typically apply build-up reuse, while the latter can be attributed to writers who typically apply boil-down reuse.

Table 4: Contingency table: writers over reuse style.

Reuse	**Writer ID**										
Style	A02	A05	A06	A07	A10	A17	A18	A19	A20	A21	A24
build-up	4	27	11	4	9	13	12	4	9	18	2
boil-down	52	5	0	14	2	13	11	3	0	0	24
mixed	10	3	0	1	1	7	6	0	0	3	1

When comparing the plots we notice a very interesting effect: it appears that writers who conduct boil-down reuse vary more wildly in their behavior. The reuse style of some writers, however, falls in between the two extremes. Besides the visual analysis, Table 4 shows the distribution of reuse styles

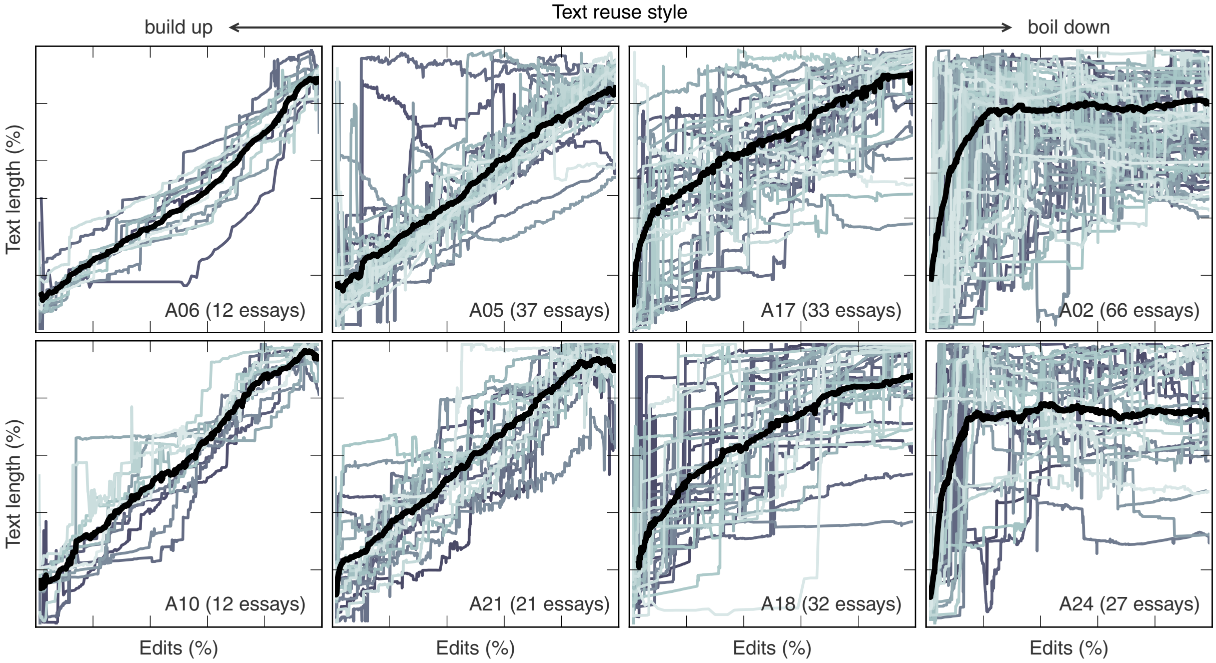

Figure 4: Text reuse styles ranging from build-up reuse (left) to boil-down reuse (right). A gray curve shows the normalized length of an essay over the edits that went into it during writing. Curves are grouped by writers. The black curve marks the average of all other curves in a plot.

for the eleven writers who contributed at least five essays. Most writers use one style for about 80% of their essays, whereas two writers (A17, A18) are exactly on par between the two styles. Based on Pearson's chi-squared test, one can safely reject the null hypothesis that writers and text reuse styles are independent: $\chi^2 = 139.0$ with $p = 7.8 \cdot 10^{-20}$. Since our sample of authors and essays is sparse, Pearson's chi-squared test may not be perfectly suited which is why we have also applied Fisher's exact test, which computes probability $p = 0.0005$ that the null hypothesis is true.

4 Summary and Outlook

This paper details the construction of the Webis text reuse corpus 2012 (Webis-TRC-12), a new corpus for text reuse research that has been created entirely manually on a large scale. We have recorded consistent interaction logs of human writers with a search engine as well as with the used text processor; these logs serve the purpose of studying how texts from the web are being reused for essay writing. Our setup is entirely reproducible: we have built a static web search environment consisting of a search engine along with a means to browse a large corpus of web pages as if it were the "real" web. Yet, in terms of scale, this environment is representative of the real web. Besides our corpus also this infrastructure is available to other researchers. The corpus itself goes beyond existing resources in that it allows for a much more fine-grained analysis of text reuse, and in that it significantly improves the realism of the data underlying evaluations of automatic tools to detect text reuse and plagiarism.

Our analysis gives an overview of selected aspects of the new corpus. This includes corpus statistics about important variables, but also exploratory studies of search behaviors and strategies for reusing text. We present new insights about how text is composed, revealing two types of writers: those who build up a text as they go, and those who first collect a lot of material which then is boiled down until the essay is finished.

Parts of our corpus have been successfully employed to evaluate plagiarism detectors in the PAN plagiarism detection competition 2012 (Potthast et al., 2012a). Future work will include analyses that may help to understand the state of mind of writers when reusing text as well as of plagiarists. We also expect insights with regard to the development of algorithms for detection purposes and for linguists studying the process of writing.

Acknowledgements

We thank our writers at oDesk and all volunteers for their contribution. We also thank Jan Graßegger and Martin Tippmann who kept the search engine up and running during corpus construction.

References

Jeff Barr and Luis Felipe Cabrera. 2006. AI gets a brain. *Queue*, 4(4):24–29.

Paul Clough and Mark Stevenson. 2011. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45:5–24.

Paul Clough, Robert Gaizauskas, Scott S. L. Piao, and Yorick Wilks. 2002. METER: MEasuring TExt Reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, PA, USA, July 6–12, 2002*, pages 152–159.

Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. 2011. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 14(5):441–465.

Tamer Elsayed, Jimmy J. Lin, and Donald Metzler. 2011. When close enough is good enough: approximate positional indexes for efficient ranked retrieval. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011), Glasgow, United Kingdom, October 24–28, 2011*, pages 1993–1996.

Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. 2013. From Search Session Detection to Search Mission Detection. In *Proceedings of the 10th International Conference Open Research Areas in Information Retrieval (OAIR 2013), Lisbon, Portugal, May 22–24, 2013*, to appear.

Djoerd Hiemstra and Claudia Hauff. 2010. MIREX: MapReduce information retrieval experiments. Technical Report TR-CTIT-10-15, University of Twente.

Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008), Napa Valley, California, USA, October 26–30, 2008*, pages 699–708.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Naval Air Station Memphis, Millington, TN.

Katrin Köhler and Debora Weber-Wulff. 2010. Plagiarism detection test 2010. `http://plagiat.htw-berlin.de/wp-content/uploads/PlagiarismDetectionTest2010-final.pdf`.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems (Infoscale 2006), Hong Kong, May 30–June 1, 2006*, paper 1.

Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. 2009. Overview of the 1st international competition on plagiarism detection. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009)*, pages 1–9.

Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. 2010a. Overview of the 2nd international competition on plagiarism detection. In *Working Notes Papers of the CLEF 2010 Evaluation Labs*.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010b. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, August 23–27, 2010*, pages 997–1005.

Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Overview of the 3rd international competition on plagiarism detection. In *Working Notes Papers of the CLEF 2011 Evaluation Labs*.

Martin Potthast, Tim Gollub, Matthias Hagen, Jan Graßegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, Alberto Barrón-Cedeño, Parth Gupta, Paolo Rosso, and Benno Stein. 2012a. Overview of the 4th international competition on plagiarism detection. In *Working Notes Papers of the CLEF 2012 Evaluation Labs*.

Martin Potthast, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. 2012b. ChatNoir: a search engine for the ClueWeb09 corpus. In *Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2012), Portland, OR, USA, August 12–16, 2012*, page 1004.

Martin Potthast. 2011. *Technologies for Reusing Text from the Web*. Dissertation, Bauhaus-Universität Weimar.

Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th ACM Conference on Information and Knowledge Management (CIKM 2004), Washington, DC, USA, November 8–13, 2004*, pages 42–49.

Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the 2004 Conference on Human Factors in Computing Systems (CHI 2004), Vienna, Austria, April 24–29, 2004*, pages 575–582.