Measuring the evolution of a revised document

Michael Wininger

University of Hartford | USA

Abstract: By analyzing two drafts of a single written piece, we open windows into that document's evolution that may not be knowable otherwise. However, existing document comparison tools generally do not facilitate scientific inquiry, as they are generally low-throughput and lacking in visual accessibility. Here, we introduce sequence homology analysis (SHA) as an alternative approach to measuring changes between two documents. SHA is a technique common to molecular biologists viz. studies of amino acid- and DNA sequences, and has been extensively validated. Whereas there is no known application of sequence homology analysis in writing researches, we give overview to its implementation and interpretation, and present a novel algorithm which incorporates SHA into the study of any number of documents in a semi-automated fashion. Additionally, we propose a method for visualization based on standard network analytic conventions. We illustrate SHA, the algorithm, and the network visualization via a publicly accessible dataset of historical significance: consecutive drafts of United States President Dwight D. Eisenhower's farewell speech (EFS). Additionally, we describe the parameterization of this routine, its potential for further automation, and its extension into other areas of writing research.

Keywords: Sequence homology, revision, Eisenhower, draft, plagiarism

journal of
WRITING RESEARCH

Wininger, M., (2014). Measuring the evolution of a revised document. *Journal of Writing Research, 6*(1), 1-28. http://dx.doi.org/10.17239/jowr-2014.06.01.1

Contact: Michael Wininger, University of Hartford | Department of Rehabilitation Sciences, 200 Bloomfield Avenue, West Hartford, CT 06117 - Department of Veterans Affairs, Cooperative Studies Program, 950 Campbell Avenue, West Haven, CT 06516 | USA – wininger@hartford.edu.

1. BACKGROUND

1.1 Overview of this work

Here we present a new approach to the analysis of written works with an interest in facilitating the measurement of the changes made to a document over serial drafts. This approach is based on sequence homology analysis (SHA), a technique well-known to researchers in molecular biology and computer science, where analysis of text-like genomic data has been commonplace for years. However, while the tools themselves have a broad support in those literatures, their application to prose texts is as yet unheard of. The scope of this work is to 1) introduce sequence homology analysis as a tool with application to lexical data, 2) describe a novel algorithm that incorporates SHA to facilitate the matching of text snippets across a set of documents, and 3) to illustrate a visualization paradigm based in network analysis. We will demonstrate these concepts via a set of drafts of a well-known, publicly accessible and interesting historical document, namely the Eisenhower farewell speech (EFS). Purposely we limit this work to a technical discussion and intentionally avoid the testing of specific hypotheses. However, it is intended that this document serve as a self-supporting work with sufficient operational detail to permit any investigator to begin testing their own hypotheses straight-away.

1.2 Relevance of revision to writing research

The evolution of new ideas is a complex process involving forecasting, appraisal, and revision (Lonergan, Scott, and Mumford, 2004; Alamargot & Lebrave, 2010). Revision, in particular, reflects betterment borne out of personal development, new knowledge gained by instruction and collaboration, and maturation of a topical understanding (Silveira, 1999; Groenendijk, Janssen, and Rijlaarsdam, 2001; Belda, Boni, Peris, and Terol, 2012). In this way, revision processes yield data from which we can measure the cultivation of a creative product; from two drafts of a single document, cognitive growth can be studied (Mumford, Medeiros, and Partlow, 2012). Moreso than other aspects of the creative process, revision is becoming an increasingly important activity for research: In addition to the continual progress made in the fields of pedagogical practices that frequently emphasize writing exercises (MacArthur, 2009; Wingate & Tribble, 2011), the explosion of community-editable documents (wikis) has changed our perspective on how thought and creative products evolve when availed to broad inputs (Albors, Ramos, and Hervas, 2008; Cress & Kimmerle, 2008; Pifarré & Fisher, 2011). Cross-document comparison is of emergent interest to a wide range of communities, including those in the fields of psychology, education, computer science, and engineering (Ji et al., 2013).

The data for conducting such research are freely available: students at every grade level continually revise drafts of their own course-related writings, and academicians routinely re-submit manuscripts pursuant to publication of scholarly enterprise; even professional writers may work through dozens of drafts before the final copy is delivered (Bisaillon, 2007). Many online wikis make the revision history publicly accessible, and governmental archives often warehouse drafts of official speeches following declassification. However, despite the copious interest in document revision by the research community, and the utter volume of available data, there is a paucity of analytical tools for visualizing or quantifying material changes to iterative drafts.

1.3 Digital revolution: Written works and the Human Genome

In drawing an analogy of rewriting of text to other natural processes of revision, we consider the billions of characters comprising genetic "texts," and how they are revised between generations (e.g. evolution or adaptation), or how two different drafts might be compared against one another in order to assess similarity (e.g. comparative biological research). For these pursuits, a suite of tools is already in steady use: Sequence Homology Analysis is routinely used to search for genetic matches in the interest of ascertaining familial relations (e.g. paternity), disease risk factors, and forensic identification. The maturation of SHA is partly a reflection of the success of the Human Genome Project (The White House, 2000), which created a full library's equivalent of genetic information for the dissemination to the global research community. A similar revolution is happening presently within the international community: the digitization of historical documents for repository in electronic format (e.g. National Archives, 2007). Whereas the first phase of the digitization of the United States' National Archives will endure for 10 years (projected completion in 2016), it is imperative to ask at this stage: how can we make the most efficient use of this incredible data resource? Here, we propose that a technique developed during the "digital revolution" of our genome could prove equally impactful in the transformation of the world's historical artifacts into accessible, digital archives.

1.4 Sequence homology analysis

In its most elemental form, sequence homology analysis is a binary assessment of the match between any two character sequences, e.g. two strands of DNA. Where DNA is comprised of four species (nucleotide molecules, NT, abbreviated A, C, G, and T), consider the hypothetical example of a search for the 4-NT sample ACGC within the 10-NT motif TATAAACGCC found in multiple phrased repeats in genomic sequences (Widlund et al., 1997). We lay out these two sequences in grid fashion, filling in any cell corresponding to a character match (Figure 1a).

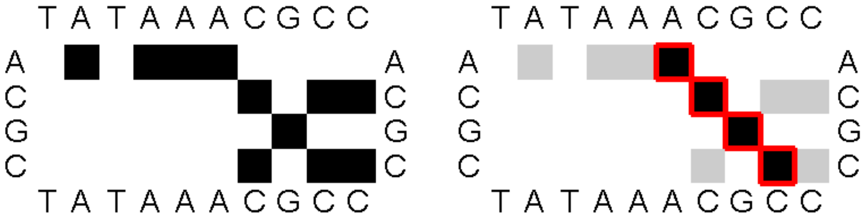

Figure 1. Sequence homology matrix for sequence ACGC in TATAAACGCC. Character matches are indicated by filled cells in the matrix (Left); runs of consecutive character matches are seen as down-and-right diagonal lines (Right). Some match cells shaded grey for clarity.

Whereas the sequences are laid out top-to-bottom and left-to-right, any run of consecutive matching characters will manifest as a diagonal line moving down and to the right within the matrix.

There are a variety of ways to "score" this matrix, i.e. to estimate the probability that sequence 2 (ACGC) is actually found in sequence 1 (TATAAACGCC) (Mount, 2008). In the present work, we will adopt the simplistic heuristic of the ratio of the sum of all character matches (black cells within the matrix) to the length of the shorter of the two sequences (shorter edge of the matrix). From the figure above, the score would be 11 ÷ 4 = 2.75, however we will condition this matrix somewhat before computing the homology score. Specifically, it is often helpful to filter the homology matrix for small character runs.

As evident from the above example, there are many singleton matches turned up by chance. It may be preferable to filter this matrix for runs below a given threshold, e.g. 1 or 2 characters. By discarding singletons and dyads, we greatly increase our likelihood of retaining only true positives (Figure 2).

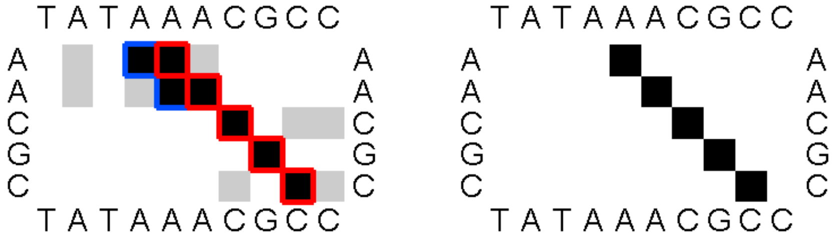

Figure 2. Increasing the length of sequence 2 by one character (to AACGC) yields two character runs (Left). Filtering for run length < 3 (i.e. removal of singletons and dyads) will reduce the likelihood of false positives. Some match cells shaded grey for clarity. Red boxes highlight true match, blue box highlights spurious false positive.

Caution is urged here as filtering will eliminate all sub-threshold runs, including those associated with a legitimate sub-sequence match (i.e. filtering may convert true positives into false negatives). Consider, for example that we look for a match to AAGCC. This would yield a 2-character match (indeed: two 2-character matches), and a 3-character match, in addition to some stray character matches (Figure 3a).

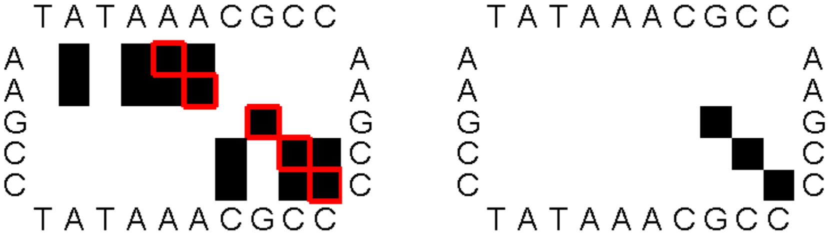

Figure 3. A slightly altered sequence 2 (such that it is an imperfect sub-sequence of sequence 1: AAGCC) yields two character runs (Left). Filtering for run length < 3 will increase the likelihood of false negatives. Red boxes highlight true match.

Here, filtering was successful in clearing the matrix of "noise," but also inadvertently eliminated cells associated with the true sequence match (Figure 3b).

Though illustrated here in the 4-letter genomic alphabet, the extension to the >100 ASCII character set comprising digital text is straight-forward. Clearly, filter length will have a complex and sometimes subtle impact on the homology score, although homology matrices for full-alphabetic texts (with dozens of unique characters) will be much less labile to the vagaries of filtering. This parameter should be treated as an equal study parameter, akin to the lens objective on a microscope where the type of phenomenon available for observation is a direct function of the magnification level. We discuss the impact of filtering this in greater detail in a later section.

2. METHOD

2.1 Dataset creation

All examples discussed here are shown in a comparison across four drafts of Dwight Eisenhower's farewell speech (EFS) in his departure from the United States Presidency in January 1961, i.e. the "military-industrial complex" speech. Digitized copies of the typewritten drafts were accessed from the online repository at the Eisenhower Presidential Library and Museum where the documents are posted in chronological order, except for an undated draft which was posted ahead of the three dated drafts (December 21, 1960, January 7, 1961, and January 17, 1961, i.e. the final version of the speech as it was delivered to a television audience). Drafts were accessed from the

National Archives web repository [http://www.eisenhower.archives.gov/research/online_documents/farewell_address.html]; drafts were downloaded as .PDF content and manually transcribed to digital text and reviewed for transcription errors both by visual inspection and with the aid of a computerized spell check.

We sought here to identify pairs of sentences between speech drafts that convey the same content, even if altered substantially between draft iterations. While it is possible to search for matches with greater (entire paragraphs, pages, or sections) or lesser (single words or characters) scope, comparing the drafts at the sentence level strikes a balance between granularity and interpretability. Single sentences were extracted using an automated "string-splitting" routine, via a custom script written for the Matlab programming environment: for each draft, sentences were identified as any text preceding a period and a space character: ". " (See Appendix A). Individual sentences were then stored in a separate work file, along with information related to the draft number, *D*, (*D*=1, 2, 3, or 4), and sentence order number, *S*, (an integer).

2.2 Sentence-pair identification

The desired end product of our data processing was to identify every sentence in Drafts 2, 3, and 4 as either a) a completely new addition to the speech, or b) a reiteration of a sentence found in a previous draft; each sentence would thus be labeled with a "tag" *T* (an integer). Sentences identified as pairing to a previously tagged sentence would be assigned to the same tag; new additions would be assigned a new tag.

As a matter *de rigueur*, all sentences in Draft 1 were automatically assigned a new tag corresponding to the sentence order within the draft, i.e. *T* = *S*. For all other sentences, tags were proposed on the basis of maximum probability of match, following SHA of all sentences prior. For each candidate sentence *c* possibly matching to the untagged sentence *s*, the homology matrix M was computed as follows: M_{ij}=1 if c_i=s_j, 0 otherwise (c.f. Figures 1-3).

This matrix was then filtered for singletons, yielding F, where F_{jj} = 1 if M_{ij}=1 and ($M_{i-1,j-1}$=1 or $M_{i+1,j+1}$ =1) and F_{ij} = 0 otherwise (naturally, this criteria is amended if i =1 or i = |*c*| and j = 1 and j = |*s*| (where |x| denotes the number of elements in sequence x). From this filtered matrix, a homology score *h* was computed as:

$$h = \frac{1}{\min(|c|,|s|)} \cdot \sum_{i,j} F_{ij} \qquad [1]$$

Homology scores were sorted in descending order, and a candidate match to *s* was proposed on the basis of maximum homology score across all candidates.

This routine was semi-automated in that matches were proposed but only accepted with User consent. The tag assignment process worked as follows: The User was prompted to accept or reject the candidate match. If the proposed match was accepted, the tag information *T* was carried over from the tagged match and assigned to the newly tagged sentence. If the proposed match was not accepted, User was offered the opportunity to Force a tag match (either to label the untagged sentence with an existing tag from a previous sentence, or to instantiate a brand new tag altogether). Alternatively, if the match was not accepted, but the User did not want to force a tag, then the next-best candidate in the bank of scored sequences was presented as an alternative, at which point the User was prompted again to agree or disagree with the match. The loop through candidates would continue until the User either a) accepts a suggested candidate, b) elects to force a match, or c) all candidates are exhausted, at which point a new tag is forced.

2.3 Sentence categorization

In addition to the draft number, *D*, the sentence number *S* within the draft, and the tag identifier *T*, a final piece of information was associated to each sentence: a category designation *C* (also an integer value). This category label reflects the User's interpretation of the general theme supported by the sentence and is a helpful design feature for the reason that the EFS is not a short composition and as a consequence, visualization may be difficult. In order to mitigate the burden of visualization across such a lengthy text, sentences were grouped into topics. It is stressed that these topic designations were useful only in enhancing the visual accessibility of the EFS, and that this is a *feature* of the analysis but not an integral component of it: the document analysis is fully supported without the specification of topic information. The topic categories used in the analysis of the EFS drafts are listed in the Table 1.

Table 1. Topics within the Eisenhower Farewell Speech drafts

1	Relationship with Congress	5	Military-Industrial Complex (future)	9	Goals of government
2	Lifetime of service	6	View to the future	10	Society and government
3	Military-Industrial Complex (past)	7	Technological revolution	11	Reflection on modern US History
4	Military-Industrial Complex (present)	8	Free society	12	Change of Administration

Category designation for any new sentence (including the sentences comprising Draft 1) were noted manually; for any sentence paired to a previously tagged sentence, the category information of that previous sentence was ported over to the newly tagged sentence automatically.

Thus, *D* and *S* are passive descriptors of the sentence extracted automatically in the data processing; the tag *T* allows sentences to be linked across multiple drafts (even if the contents of the sentence have only modest homology), and the sentence category *C* is a non-essential data point that facilitates grouping of sentences within topics for ease of visualization and interpretation.

All sentence properties (*D*, *S*, *T*, *C*) were stored in a file that could be opened mid-routine to remind the User of the existing tags and categories (and category descriptions). The workflow of this routine is summarized in the Figure 4.

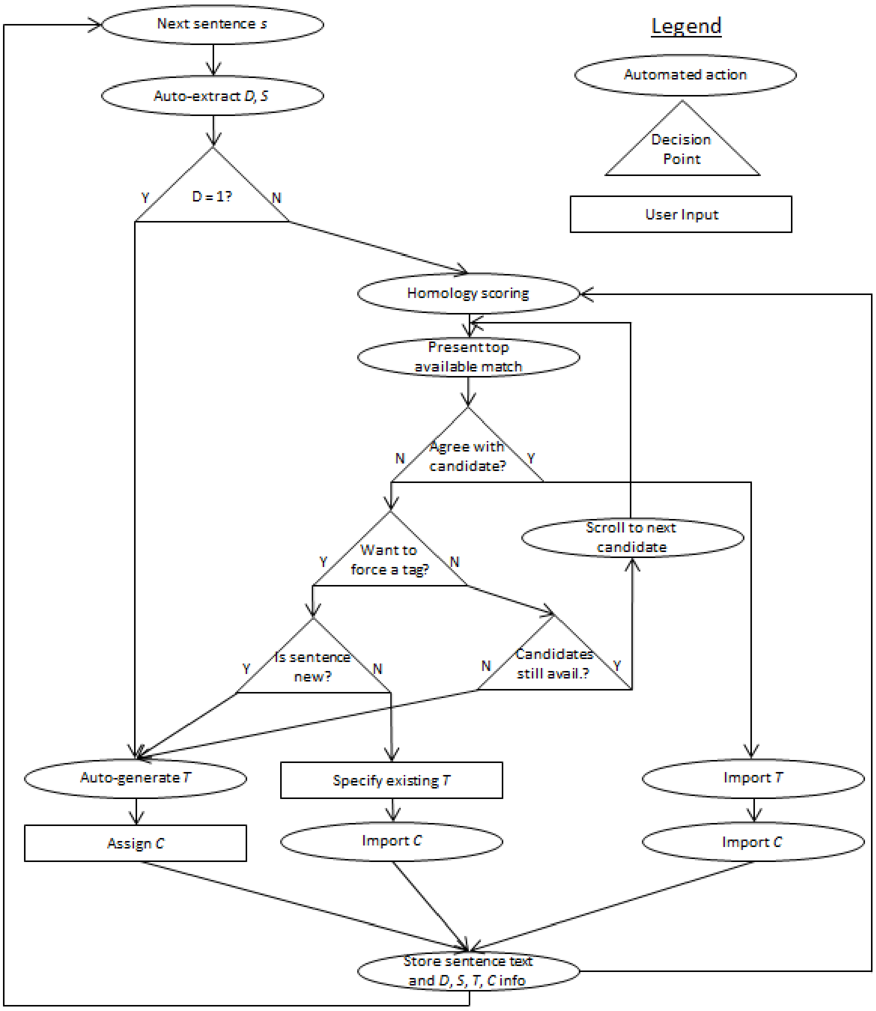

Figure 4. Workflow for semi-automated algorithm for matching sentences across drafts (“tagging”).

Homology scoring involves calculation of homology score *h* between candidate sentence *s* and all candidate sentences *c* encountered previously, selecting top candidate (the "available match") based on maximization of *h*.

2.4 Network building

In a writing research application, conversion of raw data into an interpretable, convenient graphical representation is an important facet of any research approach (Caporossi & Leblay 2011, Southavilay, Yacef, Reimann, and Calvo, 2013). Following tagging and categorization of all sentences in the EFS draft series, the data were packaged as a network of linked sentences, pursuant to an interpretable graphical representation. A network comprises two components: nodes (sentences) and edges (connection of two sentences via a tag). Our node properties are as follows: vertical and horizontal placement in space were given by draft number (top row = earliest draft; bottom row = latest draft) and sentence order (left-to-right), respectively; category was visualized as node color. On account of one draft being undated, it was considered the *de facto* first draft, as it is the first one listed in the download source (the Eisenhower Library website). Edge width was given by the similarity between tagged sentences: identical sentence pairs received a thick line; a heavily revised sentence pair was connected by a thin line as determined by homology score *h*. Nodes with no connection from the top are new sentences with no prior representation in an earlier draft; nodes with no connection out from the bottom are sentences that are eliminated from all future revisions.

While many other parameters of the sentences and their within-pair relationships could be retained (and the network map customized accordingly), descriptors were limited to this small set for the sake of clarity. These interaction and property files were then imported into the Cytoscape network visualization software (Shannon, Markiel, Ozier, Baliga, and Wange, et al., 2003), a freeware and open-source software platform used by researchers in many different fields. Node and Edge properties were customized via the Cytoscape VizMapper, as summarized in Table 2. Sample output from Cytoscape's visualization environment are highlighted in the Results section.

Table 2. Network visualization properties within the EFS draft network

Feature	Property	Mapping style	Notes
Node color	Category	Discrete	Set manually to maximize clarity
Node Location X	Sentence order within draft	Passthrough	Left-to-right arrangement
Node Location Y	Draft number	Passthrough	Top row = Draft 1
Edge width	Connection strength	Passthrough	Thick lines reflect maximum homology, thin lines reflect minimum homology

Node Label	Sentence number	Passthrough	Redundant with Node Location X, but facilitates accessibility

3. RESULTS

3.1 Descriptive Statistics

Among the four EFS drafts, there were 260 sentences: 32, 71, 79, and 78 respectively. While there are a wide variety of analyses that would reveal the extent to which a document is altered in its revision, we show a small sample of descriptive statistics here. Firstly, we note that sentences shared between Draft 2 and Draft 3 were -on average- less homologous than the sentences shared in the other Draft pairs (average homology 1.27 versus 1.56 and 1.54, following filtration for singletons, Table 3).

Table 3. Measure of revision intensity between drafts: average ± standard deviation of homology score (higher score: greater average homology); shown for scoring following filtering of 1-character matches (small filter) and 1- and 2-character matches (medium filter)

	Draft 1 – Draft 2	Draft 2 – Draft 3	Draft 3 – Draft 4
Small filter (≤1 char.)	1.56 ± 0.69	1.27 ± 0.69	1.54 ± 0.59
Medium filter (≤2 char.)	1.15 ± 0.28	0.58 ± 0.23	1.00 ± 0.35
Sentence insertions	43	49	11
Sentence deletions	4	41	12

Furthermore, in addition to the low homology in the revision between Draft 2 and Draft 3, there were more insertions/deletions: (49/41) than either of the other revisions. We also report this homology score following filtration for dyads as well as singletons, as we believe that this is a more accurate reflection of the veridical change though revision (See Discussion).

3.2 Draft ordering (graphical approach)

The 260 sentences comprising the four publicly available drafts of the EFS are shown in Figure 5.

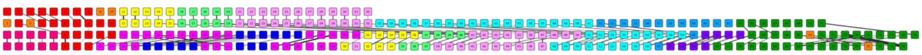

Figure 5. Broad network view of the four drafts of the EFS. Sentences shown as nodes, colored by topic, and connected to highly homologous sentences in other drafts.

3.3 Draft ordering (numerical approach)

The graphical depiction of the network suggests that the undated draft may precede the December 21 draft: they have a similar distribution of categories across the sentences, the December 21 appears to copy-paste most of the sentences of the undated draft before initiating a series of new topics, and there was a very small number of deletions versus a very high number of insertions, which suggests a "building" process. Circumstantially, it would appear that the undated draft was an early-stage draft upon which the December 21 was built.

But is there a quantitative basis to support this notion? We ask two questions: 1) how many sentences are shared between Draft 1 and each of the other drafts, and 2) among these shared sentences, what is the average homology?

One outstanding question related to this dataset concerns the proper ordering of the drafts with regard to the undated draft. We see that the undated draft (positioned as the top row in Figure 5), is substantially shorter than the other drafts, and has a very similar architecture to the first half of Draft 2. Though uncertain, it seems plausible -based on these tandem observations- that the undated draft could well precede Draft 2.

Table 4. Homology analysis of sentences shared between the undated draft (Draft 1) and each other drafts

	Draft 1 – Draft 2	Draft 1 – Draft 3	Draft 1 – Draft 4
N	27	12	11
Small filter (≤1 char.)	2.61 ± 1.05	1.41 ± 0.67	1.72 ±1.03
Medium filter (≤2 char.)	1.19 ± 0.21	0.48 ± 0.29	0.56 ± 0.31

From Table 4, it is evident that there were more shared sentences between the undated draft and the December 21 draft (21 versus 12 and 11), and these shared sentences were of much greater homology than were the shared sentences in pair-wise arrangements between the undated draft and the January 7 and January 17. While not conclusive proof of the pre-ordering of the undated draft relative to the December 21 draft, it adds constructively to the evidence supporting that scenario.

4. DISCUSSION

4.1 Scope and Limitations

We discuss in this work three concepts: 1) sequence homology analysis (SHA) which is a well-established technique heretofore not applied to writing research applications, 2) an entirely novel procedural algorithm by which SHA can be streamlined into a writing research activity, and 3) a network visualization approach, which is common to many research fields, not limited to biology but somewhat new to writing research. Readers are referred to more comprehensive reviews of SHA for an in-depth discussion of its limitations (e.g. Mount, 2001), but a few are highlighted here. Firstly, SHA is a parameterized analysis: the User defines the search scope (i.e. the length of characters that might constitute a homology: at the level of individual characters, words, sentences, paragraphs, etc.), and the threshold at which a homology is determined to exist (e.g. filtering singletons, or singletons + dyads, etc.). The nature of the analysis and its interpretation may change depending on these parameters, as changing these parameters will impact the speed, sensitivity, and specificity of the SHA. Therefore, in this process –like any other parameterized analysis– great care must be exercised in setting parameters, reviewing and interpreting the output. Furthermore, whereas the SHA paradigm is well-understood for its use in the 4-letter genetic alphabet, its application to written texts –with a much more extensive alphabet and nature of writing– is as yet unknown. A facile utilization of SHA (and interpretation of its results) will require a great deal of community research; this work can only serve as a promulgation of the idea into writing research.

The algorithm described here was developed in service of the present need to analyze the four drafts of the EFS, and may not necessarily be ideal for another research activity. Furthermore, this routine involved supervision at several junctures; there will be many researches for which the level of supervision must be intensified or reduced. At the present time, there is no clear path to a fully automated analysis that reproduces the full results presented here. However, potential avenues for automation are discussed in a subsequent section.

The network visualization is a routine graphical action with well-known pros and cons, which can only be briefly discussed here. As indicated previously, there are dozens of ways that a network-style graphic can be customized to convey important information about the draft analysis, and there are a wide range of quantitative measures that are commonly used in other research arenas to describe qualities of the network. Many of these activities can be carried out within the Cytoscape environment via freely downloadable toolkits (“plug-ins”). However, whereas the application of network theory to writing analysis is still somewhat new (Caporossi & Leblay, 2001), the opportunities and limitations are yet to be fully defined. This is, of itself, a limitation: a considerable amount of basic research in the application of network theory remains to be completed before we can begin to use these concepts to rigorously test hypotheses about written documents.

Furthermore, the methodology described here is fundamentally different than the tools currently available for assessing the evolution of a single draft as it is being composed (e.g., Van Was & Leitjen, 2006; Lindgren, Leitjen and Van Waes, 2011, Ahlsén & Strömqvist, 1999). These “online” tools are ideal for capturing edits in real time so that they can be assessed for a valuable insights into the physical act of creation of a written work. Rather, the approach taken here captures changes made between drafts and not during the editing process. The details of the physical act itself are therefore lost, and in their place: an analysis of the lexical content. The online tools are appropriate for when the placement (or removal) of each character can be measured;

the SHA-based tools are best for when one or more drafts is complete and no other information is available.

4.2 Algorithmic automation

It is not uncommon to identify matches within DNA sequences (akin to tagging here) based on blind thresholding. Whereas the tagging is the crux of the algorithm presented here, and the primary outcome of the SHA, with additional researches into the probability scores found in written texts, it may be possible to automate this step. However, this could only reasonably follow a validated benchmark study performed on many documents, and possibly a variety of probability scoring rules.

The document analysis algorithm described here is a "supervised" routine in that no sentence is tagged or categorized without authorization from the User. This paradigm is one of several possible approaches with varying degrees of supervision. The four EFS drafts are a relatively small dataset (260 sentences), so a supervised review did not pose a large burden, and assured minimum chance at miscategorizing sentence topics. However, for larger datasets, or for applications where the tolerance for subjectivity is low, this routine can be automated. Three arenas for automation are discussed briefly here:

1. Transcription of .PDF content to digital text can be made very high-throughput via optical character recognition (OCR). There are a multitude of proprietary softwares, as well as freewares that can be used for OCR.

2. Sentences could be categorized manually via Natural Language Processing (NLP). NLP is an emergent rules-based technique which automates classification of text based on keywords and lexical relationships. While NLP at present requires a fair experience in computer programming, it is likely that applications will be developed in the near future that will make NLP accessible to a broader audience.

3. Sentence tagging could be left to a simple winner-take-all criteria, i.e. remove the solicitation of User input. This would eliminate a potentially intensive aspect of any document analysis enterprise, but runs the risk of mis-categorizations. New sentences (i.e. text inserted into a later draft) typically require new tags (as opposed to applying an old tag), so a threshold homology score would have to be pre-set. On account of the novelty of this method, particularly with respect to its application to the analysis of written text, it is advised a great deal of thought would have to go into how such an automation would be structured. We pilot this method of automation in the following section.

As with any juncture where automation is a feasible alternative to User action, the substitution of algorithmic decision making for human supervision requires a great deal of consideration, preparation, execution, and proof-checking. Indeed, it may be optimal to retain as much User input as is practicable while streamlining the operations which support the supervision.

4.3 Reliability of SHA in identifying matches

Pursuant to an automation of the sentence tagging, it is appropriate to ask: "How quickly does SHA find the optimal match?" This is the equivalent question to: "How reliably could an unsupervised SHA find the correct tag-match if taking the top-ranked match out of all candidates?" We performed a *post hoc* benchmark test in order to answer this question. A brief summary of our method and results follows.

Among the 260 sentences in the EFS draft series, 127 sentences were not "new" sentences, i.e. there were 127 sentences matched to previously existing tags. All candidate matches were selected by a human User (i.e. the algorithm was supervised), and for an arbitrary sentence, e.g. sentence *N* out of 360, all previous sentences (1 through *N* − 1) were measured for the homology of each to the yet untagged (*N*th) sentence; thus *N* − 1 homology matrices and scores were computed. An exemplar of this calculation is shown in the Figure 6, using Sentence #196 (sentence 22 in Draft 4, i.e. "To strive for less would be unworthy of a free and religious people."), which was found in the supervised analysis to match to Sentence #125 (i.e. Sentence 14 in Draft 3: "To strive for less would be unworthy of a free and Christian people").

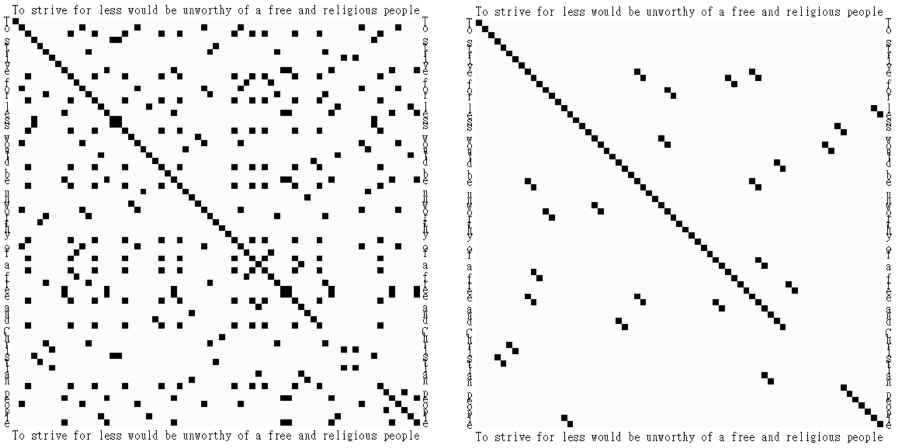
To strive for less would be unworthy of a free and religious people
To strive for less would be unworthy of a free and Christian people
To strive for less would be unworthy of a free and Christian people
To strive for less would be unworthy of a free and religious people
To strive for less would be unworthy of a free and religious people
To strive for less would be unworthy of a free and Christian people
To strive for less would be unworthy of a free and Christian people
To strive for less would be unworthy of a free and religious people

Figure 6. Sequence homology matrix for Sentence #196 (D4:22) and Sentence #125 (D3:14) with all character matches shown (Left) and after filtering for singletons (Right). This matrix yields the highest homology score of the 195 candidates searched. However, the homology score h is confounded by the false-positive dyads remaining following filter for singletons.

Following the exemplar, we show the homology scores for all 195 sentence candidates to D4:22, sorted in descending order (Figure 7).

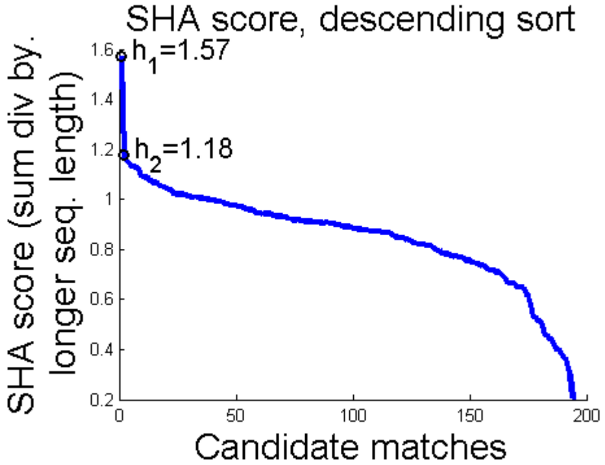

Figure 7. Sorted homology scores for match routine for Sentence #196. Best-match score (h1 = 1.57) was substantially greater than the next-highest score (h2 = 1.18).

Here, we see that the top-ranked homology scores (h_1 = 1.57) was substantially greater than the next-highest score (h_2 = 1.18); the top-score did indeed correspond to the correct tag match (i.e. to D3:14). Among the 127 simulated unsupervised matchings, the correct tag was the top-identified match in 78 cases (61.4%).

Despite the good performance of this routine, it can be seen that filtering for singletons may not be sufficient for routines requiring high precision: there are many surviving character match pairs (Figure 6b), and the second-best match (an incorrect pairing), still yielded a homology score greater than 1 (meaning that the number of match characters exceeded the number of characters being considered, viz. Figure 7). In response to this observation, we re-ran this benchmarking analysis with a filtering for dyads as well as singletons. We show the updated homology matrix in Figure 8.

From this new filtered homology matrix, two long character runs offset a single word replacement: ("religious people" for "Christian people."). There are no stray matches otherwise.

Following this increased filter length, we report that the correct sentence was identified in 109 cases (85.8%). We conclude that even modest filtering may prove reliable in producing proper candidates for matching, and may create opportunities for automating the routine, e.g. to automatically accept the first result if it is above a given threshold (say h_1 > x), or sufficiently greater than the next alternative (say $h_1 - h_2$ > y). Sample code used to create the homology matrices shown in Figures 7b and 8a is provided in Appendix B.

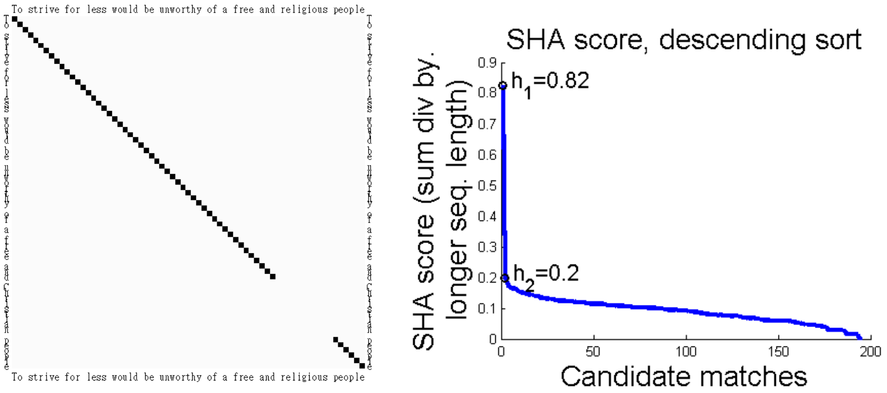

Figure 8. Update of the sequence homology matrix for Sentence #196 and Sentence #125, after filtering for runs <3 characters in length. Sorted homology scores for match routine for Sentence #196 following more aggressive homology matrix filtering. Best-match score (h_1 = 0.82) now more faithfully reflects the true homology, and is even more distinct from next-highest score ($\underline{h}_2$= 0.20).

4.4 Contrast to Keystroke Logging

A great deal of emphasis has been placed in recent years to measure the subtleties of written craftworks as they emerge during the creative process. Keystroke logging, in particular, capitalizes on the relative ease with which data can be captured from the keyboard of an author, retaining information on the timing of all events at the man-machine interface: frequency of deletions, time between inputs and edits, navigation within the document, etcetera (Lindgren, Leijten & Van Waes, 2011; Leijten & Van Waes, 2013; Ahlsén & Strömqvist, 1999). There are several similarities between the approach described here and keystroke-logging, including the detailed capture of document edits and the yielding of datasets that lend easily to a quantitative analysis. Moreover, the network-analytic visualization paradigm approach proposed here approximates the method proposed by others (Caporossi & Leblay, 2011) in response to the need for interpretable output from the keystroke-logging paradigms.

However, there are many important differences between keystroke logging and SHA and the algorithm described here. Broadly, keystroke logging is a powerful tool by which to measure the *performance* of a writing or revision task; the SHA algorithm described here is a methodology for measuring the *outcome* of a revision process. This tool is not designed to capture online edits; indeed, the impetus for this work was borne out of a need to gain information from long-completed documents written 50 years ago.

From this algorithm, we propose that one of the most impactful results is a practical view into the differences between two written products. When these documents are serial drafts, the information approximates a virtual markup of the original draft, where text alterations (insertions, deletions, embellishments, etc.) are rapidly quantified and converted into an interpretable graphical format. When these documents are (ostensibly) unrelated works, this nature of this information transforms into that of a simple similarity measure, which can be used in the service various editing activities, e.g. plagiarism detection (see subsequent section). Most importantly, this method yields research-caliber window into the architectural and content changes of a document's structure. From the network visualization shown here, far-flung edges and interspersed node colors indicate changes in organization of sentences and topics, respectively. The thinness or thickness of the edges show were sentences have undergone extensive or superficial revision, as measured by an objective rater. The objective of this tool is to manufacture an additional dimension of insight from any two documents. Especially where two "flat" drafts of a single document are compared against one another, the algorithm here provides information where no other tools can: similarity or dissimilarity within and across documents in a parametrizable, minimally supervised routine.

Keystroke logging is ideal when writing can be obtained from a "live subject" writing in real-time. For researches involving historical documents, or where the primary hypotheses do not involve temporal data and transient editing activities, SHA may prove a suitable alternative.

4.5 Contrast to Microsoft Document Comparison Tool

Perhaps the most widely-used utility for comparing changes between two documents is the Document Comparison (DC) tool featured in the Microsoft (MS) Word software, intended to serve as a document mark-up for projects where Track Changes could not be applied, e.g. as might be the case for two documents authored by someone else. The DC tool is potentially powerful and certainly handy: MS Word is used on a great many computers, and the convenience of an inline utility for document comparison –no less on that requires no expertise to use– cannot be underestimated.

However, unlike the routine introduced here, a tool like MS DC is not intended for use in a research setting: It can only compare two documents at a time, the output is entirely "temporary" (i.e. there is no way to export the information for *post facto* quantitative analysis), there is no ability to visualize the changes to the document except for via the inline markup, and the tool is not at all customizable. For a very limited set of circumstances and perhaps best applied to only small segments of text, very few of which would involve rigorous research-based document analysis, the DC tool provides a streamlined, efficient, totally unsupervised document comparison. For all other pursuits, the DC tool furnished within the MS suite is rather limited.

One particularly important shortfall of the DC tool is that it does not capture translocated text: there is no markup for a text translocation, only for insertion or deletion; this entire component of data is not only missing, but is mis-categorized. Whereas sentence rearrangement is commonplace in revision, were an attempt made to make inferences about the revision of a document based on the DC tool, the conclusions made would be utterly unreliable on this basis alone. Furthermore, though the DC algorithm is proprietary and therefore cannot be dissected, it is apparent that the DC tool attempts to match document contents based on the order in which they appear. This is a fundamentally different approach than is taken here, where all sentences are evaluated at the same time, giving equal chance to find a match to the first sentence as to the last. As an example, consider a snippet of text from the December 21 draft (Sentences 64-68 within the draft) and from the January 7 draft (Sentences 64-69 within that draft).

Table 5. Copy of the output of the Microsoft Word Document Comparison tool (DC) for two passages copied from the Eisenhower Farewell Speech set: December 21 draft, sentences 64-68 (96-100 of 260; *at Top*) and January 7 draft, sentences 64-69 (167-172; *at Middle*). Strikeout text = deletion, Underline text = insertion, Plain text = Unaffected in revision. Compare to Figure 9, (output from SHA + network visualization on same passages)

Passage 1:

Members of the Congress, my prayer for the future is that the world in which we live can be turned from a community of fear into a confident confederation of mutual trust and respect. The conference table may be marked by a sense of frustration and disappointment with the past, yet scarred though it may be, we must not forsake it for the certain terrors of nuclear war. The tools of the open society are still available to us. We dare not fail to use them. Believing as I do in the fullness of the American character, I have every confidence we shall.

Passage 2:

We want democracy to survive for all generations to come, not to become the insolvent phantom of tomorrow. America's heartfelt yearning for the future is that this world of ours, ever growing smaller can avoid becoming a community of dreadful fear and hate, and instead a proud confederation of mutual trust and respect. Protected by our moral, economic, and military strength, we can advance to the world's conference table with confidence. That table, scarred though it may be by many frustrations and disappointments, must not be abandoned for the certain agony of a mutually-destructive, preposterous war. Believing as I do in the sturdiness and understanding of the American people, and in the abiding desire of people everywhere for peace with justice, I have every confidence we can sustain the free world security and hold fast to our democratic ideals. So – as I say goodnight to you on the eve of my departure from official life, I thank you for the opportunities you have given me for public service in war and peace.

Document Comparison:

~~Members of~~ We want democracy to survive for all generations to come, not to become the ~~Congress, my prayer~~insolvent phantom of tomorrow. America's heartfelt yearning for the future is that ~~the~~ this world ~~in which we live~~of ours, ever growing smaller can ~~be turned from~~avoid becoming a community of dreadful fear ~~into~~and hate, and instead a ~~confident~~proud confederation of mutual trust and respect. ~~The~~Protected by our moral, economic, and military strength, we can advance to the world's conference table ~~may be marked by a sense of frustration and disappointment~~ with ~~the past, yet~~confidence. That table, scarred though it may be, ~~we~~ by many frustrations and disappointments, must not ~~forsake it~~be abandoned for the certain ~~terrors~~agony of ~~nuclear war. The tools of the open society are still available to us. We dare not fail to use them.~~ a mutually-destructive, preposterous war. Believing as I do in the ~~fullness~~sturdiness and understanding of the American ~~character~~people, and in the abiding desire of people everywhere for peace with justice, I have every confidence we ~~shall.~~can sustain the free world security and hold fast to our democratic ideals. So – as I say goodnight to you on the eve of my departure from official life, I thank you for the opportunities you have given me for public service in war and peace.

We highlight a few observations based on this example (following Table 5):

1. The DC tool conflates sentences: This appears to be labile to sentence ordering and suggests that the passages are read "left-to-right" (or, equivalently, "first-to-last") instead of as being taken as complete text objects with constituent sentences co-existing simultaneously. For example: Sentence 1 in the Revised Passage ("We want democracy…") is a completely new addition following revision; Sentence 2 in the Revised Passage ("America's heartfelt yearning…") is the actual revision of Original Sentence 1 (both contain the "community of fear" clause).

2. Edits are all-or-none at the word level: Edits cannot be visualized at a finer resolution (by character) or gross resolution (sentences). The extent of revision is not known: a change in single letter will appear as a substitution of the entire word (which would only be realized as a spelling error upon visual inspection). Because the resolution of the DC search cannot be altered (to review the passages for their sentences instead of their individual words), translocated sentences are likely to be broken into separate clauses each with their own insertion-deletion, e.g. "frustrations and disappointment."

3. Translocation of a sentence would obscure its finer edits: As described above, moving of a sentence (e.g. to a different paragraph) would appear as a wholesale deletion-insertion even if the sentence content remains unaltered otherwise. However, if the sentence were translocated and a legitimate word substitution made within the sentence, this second change would not be captured, as all the words in that sentence would appear as newly inserted text.

In contrast these sentences were compared using SHA and plotted as part of the larger network of sentences comprising these drafts. From the graph, a few features are immediately clear: Sentence 64 in the earlier draft and Sentence 65 in the later draft are the same sentence (sentences ending in "…mutual trust and respect."), as are Sentences 65 (earlier) and 67 (later; sentences describing a conference table marked with "frustration and disappointment."). Additionally, we see that Sentences 66-68 were dropped in the revision, sentences 64, 66, and 69 added, and one sentence changed substantially (Sentence 69 in earlier draft into Sentence 68 of later draft). This sentence, contains "Believing as I do," "of the American", and "I have every confidence" phrases in both drafts, but otherwise shows an intense revision. While many edits were identified of themselves in the Microsoft DC tool, their relationship and their impact on the architecture of the manuscript can only be captured via the SHA algorithm proposed here.

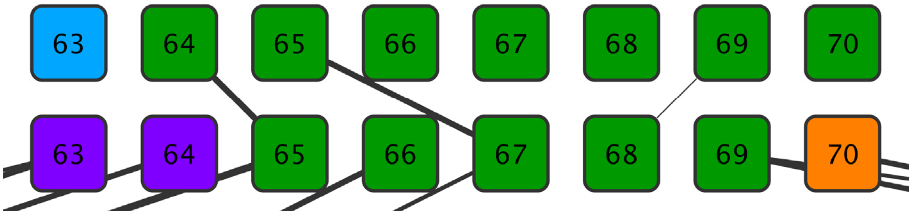

Figure 9. Network representation of two passages copied from the Eisenhower Farewell Speech set: December 21 draft, sentences 64-68 (96-100 of 260; *at Top*) and January 7 draft, sentences 64-69 (167-172; *at Middle*). Edge thickness = sentence-wise homology score. Compare to Table 9, bottom row (output from MS DC tool on same passages).

Lastly, we note here that it is possible to implement this algorithm and the SHA analysis within the Microsoft Word environment, as a complement or substitute for the DC tool. For interactive utilities ("Macros"), the MS Office Suite uses Visual Basic, which can perform all of the requisite actions of the algorithm outline here, including a) for-loops, c) text splitting by sentences or at other resolution, c) quantitative text comparison, d) storage, and e) plotting. Thus, while it seems that a researcher might prefer to perform this kind of document analysis in a computational environment, e.g. Matlab, R, or from a command line utility, those seeking to incorporate SHA into the MS environment would be able to do so with no additional labor.

4.6 Extension on this application

There are many facets of sequence homology analysis which facilitate the quantification of text-based data in extended research areas. A few are highlighted here.

4.6.1 Document forensics.

Though not a formally tested hypothesis of this work, we showed here how the a network-style visualization (supported by the SHA) creates a platform upon which basic assumptions about the nature of the document creation process can lead to inferences about the proper ordering of serial drafts (i.e. ordering the undated EFS draft ahead of the December 1960 draft); this process was facilitated by adding categorical information. In addition to organizing undated drafts, it may be further possible to place certain documents within a general time frame of creation, or to identify the approximate geographical region where its author was raised or educated, using turns of phrase and grammatical styles as landmarks. This would very likely incorporate NLP, would involve a good deal of assumption-making, and would require generous "margins" for error. However, to the extent that written language analogizes the "text" of our genome, this kind of analysis is commonplace among ethnographers, anthropologists, and historians.

4.6.2 Plagiarism detection.

As evidenced by the examples shown here (Figures 1-3, 6 and 8a) it is a straight-forward endeavor to identify homologous sequences between any two drafts. However, this technique can be expanded to find similar character runs within a much larger database. For instance, the basic local alignment search tool (BLAST) is a suite of tools used by researchers in biomedical fields to cross-reference a sample of DNA against a comprehensive database of known disease markers. This is a very high-throughput activity: most BLAST applications allow for the searching of thousands of samples against millions of database records, in the span of minutes. The BLAST tools available for genomic research come in many varieties, specializing variously in optimizing search sensitivity, specificity, speed, and scope of search.

This paradigm can be applied easily to writing research applications, for example, in comparing a manuscript against a database of historical and contemporary published works, with an interest in identifying similar passages above a threshold length. Whereas many professional writing activities still ask authors to provide assurance of their professional conduct in avoiding plagiarism, plagiarism is typically only detected by editors, or savvy readers after a manuscript is in print; SHA provides an efficient and cost-effective means by which to screen entire documents for similarity.

While there are several methods already available for plagiarism detection, they vary in their intensiveness and relationship to SHA. For instance, Finger-printing may be the approach most like SHA, with greatest flexibility to document structuring and combination of probabilistic modeling and supervision. However, one major difference is the incorporation of metadata into the sequence analysis, e.g. the total number of data "chunks" (Brin, Davis, and Garcia-Molina, 1995): the methodology proposed here does not include any metadata in the decisioning of a match; this may increase the intensity of supervision, but reduces the risk of errant conclusions about the match quality. Similarly, "string matching" which uses the *R*-measure, which is the normalized sum of the lengths of all the suffixes of the text repeated in other documents of the collection; this method not only incorporates metadata (see previous remarks; Khmelev & Tahan, 2003). Other approaches, e.g. citation analysis and stylometry are yet farther from the SHA approach.

There are commercially available routines for plagiarism detection, including those used by academic publishers. For instance, the iThenticate group's CrossCheck software is used by more than 200 publishers. Its algorithm is proprietary and therefore its scoring approach is not publicly available; however, the paradigm is generally similar to that proposed here: a similarity score is produced and sections are flagged for review by the Editor. Again, this particular software is parameterized for length, e.g. to allow for Review articles (which often incorporate extended copy-paste of passages) to have overall higher similarity scores than Original Research articles. Readers are directed to a recent review of plagiarism detection services for a discussion of both free and paid detection services (Garner, 2011).

4.6.3 Self-plagiarism and repetition identification.

In addition to homology across two putatively unrelated texts (e.g. plagiarism analysis), a single document could be analyzed for repetition within the work, e.g. self-plagiarism or repetition. Avoidance of self-plagiarism is widely recognized as a "best practice" in scientific writing, particularly in texts that do not contain highly complex methodological descriptions, say as might be repeated across multiple scientific papers using the same laboratory techniques (Roig, 2005; Bretag & Mahmud, 2009). This issue is specifically addressed by the American Psychological Association (APA, 2010). To this point, consider the following single sentence from Draft 4 of the EFS:

> But each proposal must be weighed in the light of a broader consideration: The need to maintain balance in and among national programs -- balance between the private and the public economy, balance between cost and hoped for advantage -- balance between the clearly necessary and the comfortably desirable; balance between our essential requirements as a nation and the duties imposed by the nation upon the individual; balance between actions of the moment and the national welfare of the future. Good judgment seeks balance and progress; lack of it eventually finds imbalance and frustration.

Here, the word "balance" and the phrase "balance between" are used in an epistrophe. While this is obviously an intentional replication on the part of the speech writer, it provides an illustrative example for how unintentional repetitions might be found. Consider the homology matrix for this sentence taking the same sentence as "Passage 1" (column) and "Passage 2" (rows; Figure 10a; with zoom into the start of the repetitions in Figure 10b).

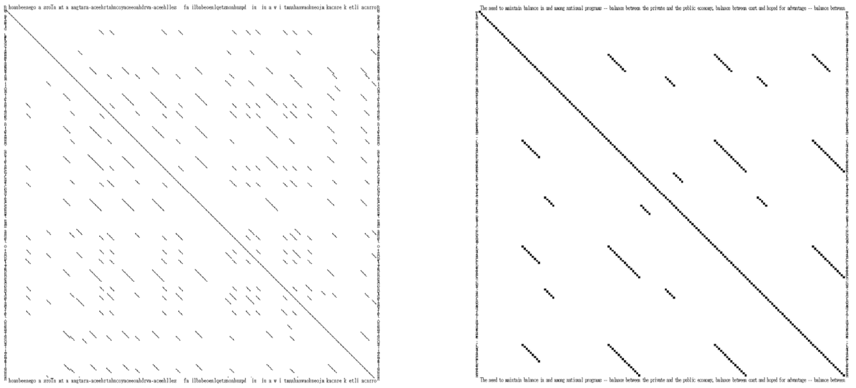

Figure 10. Auto-homology matrix of Sentence #207 (Draft 4, Sentence #25): Full homology matrix (*Left*), and zoom window of characters 77-253 ("The need to maintain balance…"), i.e. the first window in which the word "balance" appears (*Right*). Text annotation around the entire homology matrix (*Left*) is limited to every third character for clarity. Homology matrices filtered for runs less than 5 characters.

The first obvious feature is the diagonal stretching the length of the matrix: this reflects the identicality of the two passages being compared (the same sentence; by design). This diagonal assures that the homology score will be at least unity. But there are also many smaller off-diagonal "runs" of text, corresponding to text matches other than the trivial identity and which will increase the homology score. Here, we filtered for runs fewer than 5 characters long, thus any three-letter word match (plus the two offsetting spaces) survive the filter. In this way, it can be seen (from Figure 10b, especially) that the words "balance", "balance between", "and" and "the." Naturally, extending the filter to obscure longer stretches of matched characters would leave only the longer words and phrases intact.

Extracting these matched passages is a straight-forward operation on the homology matrix. Operationally, a repetition occurs wherever the columnar sum is greater than 1 (i.e. ignoring the diagonal, wherever there is at least one additional character match). We can call these repetition sequences as the characters corresponding to any locus in the matrix with a columnar sum greater than 1; repeated phrases are any string of such characters offset by a break, i.e. s_j for all $k<j<m$ such that $\Sigma_i F_{ij} > 1$ and $\Sigma_i F_{ik} = \Sigma_i F_{im} = 1$. Calling these from the windowed matrix (Figure 10b), we return:

1. balance (chars: 97-105)
2. and (chars: 108-112)
3. balance between the (chars: 139-159)
4. and the (chars: 167-175)
5. balance between (chars: 191-206)
6. and (chars: 212-216)
7. balance between (chars: 239-254)

Again, this is a partial list, as we limited our search to the zoomed window. Obviously, this list could be further filtered for matched sequences, e.g. to return only those repeated phrases longer than x characters. Sample code for this repetition extraction is provided in Appendix C.

4.6.4 Voice identification.

A great many works, including the EFS, are known or suspected to have been the collaborative effort of multiple authors. Can individual voices could be identified within a single- or multiple drafts? The answer is: possibly yes. While it would require approaches similar to those described for forensic and plagiarism analyses (see Above), it may be possible to identify separate voices within a single draft. Several research groups have demonstrated terrific success in identifying anonymous or pseudonymed publications based on lexical overlap to other published works (Liptak, 2000). The SHA analysis would be flexible to this arena of research, and so it is plausible that with adequate supporting information, individual voices could be distinguished within a single manuscript.

4.6.5 Sympatico with the author.

The network-style depiction of the four EFS drafts presented a window into how pockets of text are inserted, deleted, moved, and broken apart.

Consider, for example, the pocket of text in Draft 3 grouped under category 9 (goals of government, Figure 10). In Draft 4, the last sentence of this category is broken off from the first 5 sentences, and two additional sentences are added (Movement of Draft 3, Sentence 26 to Position #21 in Draft 4, Table 6).

While this finding, of itself, does not allow any particular insight into the thought process of Eisenhower or his speech writers, the SHA analysis and network visualization have brought into the light that a change was made. Though beyond the purview of this work, it seems reasonable that the movement of D3:26 points to its importance, and the addition of D4:22-23 (and indeed, the nature of the changes between D3:26 and D4:21) help to provide additional context and meaning.

Table 6. Highlight of bifurcated passage in Drafts 3 and 4

Draft D	Sentence S	Text
3	26	But never must we fail to meet every crisis with steadfastness, courage, and understanding, so that we may remain, despite every provocation, on charted course toward permanent peace.

4	21-23	Only thus shall we remain, despite every provocation, on our charted course toward permanent peace and human betterment. Crises there will continue to be. In meeting them, whether foreign or domestic, great or small, there is a recurring temptation to feel that some spectacular and costly action become the miraculous solution to all difficulties.

Obviously, niche applications of this approach, e.g. to understanding the thought process of Eisenhower and his collaborators in the drafting of the EFS, requires cooperative engagement with investigators with the appropriate expertise. Nevertheless, we present here a powerful tool by which text alterations can be visualized and earmarked for further investigation.

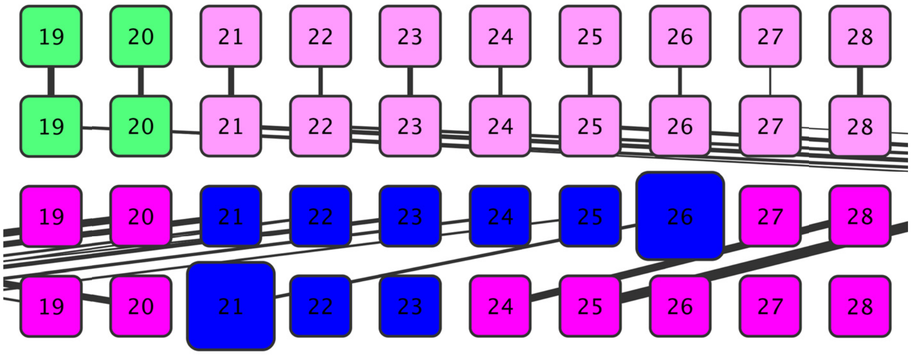

Figure 11. Highlight of tagged pair D3:26 :: D4:21.

The algorithm described here uses SHA to facilitate the tagging of sentences and categorization into topics; the result is a novel window into the evolution of a document that simply cannot be known from just the final draft, and --even with access to the text of previous drafts-- would be next to impossible to create without the elements described here. For the vast majority of historical documents, which likely do not contain such annotations, interrogating serial drafts for these types of changes may be the closest approximation possible.

5. Summary and Conclusion

We present here a technique common to molecular biologists, for novel application to research questions in a fundamentally non-biological context, i.e. evolution of a creative product viz. revision of a written document, namely sequence homology analysis (SHA). This methodology has been extensively validated for its intended application, i.e. genomic and proteomic analysis, but has heretofore not been applied to the analysis of lexical datasets, i.e. human writing. Sequence analysis has been applied in biological contexts to leverage new information about species evolution, forensic identification, and medical disorders. By converting SHA into a tool for use in the analysis of written documents, we propose that a great new set of informations can be extracted efficiently from the vast pool of available data.

SHA services a broad need in several academic communities for a reliable means by which to quantify aspects of creative processes, for which there are many existing hypotheses that would benefit from a robust objective methodology for testing, and even more as yet unrealized hypotheses that could be borne out of a familiarity with this kind of approach. For example: the analysis of serial drafts provides novel insight into the processes underlying creation of a document, including the creative milestones and thematic developments that occur across the drafting process. SHA as it is presented here, can be applied on its own to rapid categorization of text within a written document, based on a probabilistic matching routine supervised by the investigator. As a stand-alone analysis, SHA allows for facile descriptive analysis of the basic properties of a revised document, i.e. the extent of text insertion, deletion, and translocation. Coupled with other tools, e.g. network analysis or natural language processing, the utility of SHA is extended into broader, more sophisticated domains that are likely to present valuable window into the cognitive processes of the writer.

References

Ahlsén, E., &Strömqvist, S. (1999). ScriptLog: A tool for logging the writing process and its possible diagnostic use. In Loncke F, Clibbens J, Arvidson H, Lloyd L (Eds.), *Argumentative and Alternative Communication: New Directions in Research and Practice*. Wiley Publishers, London, p. 144-199. ISBN: 978-1-86156-143-5/1

Alamargot, D., & Lebrave, J.-L. (2010). The study of professional writing: A joint contribution from cognitive psychology and genetic criticism. *European Psychologist, 15*(1), 12-22. DOI: 10.1027/1016-9040/a000001

Albors, J., Ramos, & J. C., Hervas, J. L. (2008). New learning network paradigms: Communities of objectives, crowdsourcing, wikis and open source. *International Journal of Information Management, 28*(3), 194-202. DOI: 10.1016/j.ijinfomgt.2007.09.006

APA: The publication manual of the American Psychological Association. Fifth Edition, Washington, D.C., 2010.

Belda, S., Boni, A., Peris, & J., Terol, L. (2012). Rethinking capacity development for critical development practice. Inquiry into a postgraduate programme. *Journal of International Development, 24*(5), 571-584. DOI: 10.1002/jid.2850

Bisaillon, J. (2007). Professional editing strategies used by six editors. *Written Communication,* 24(4), 295-322, 2007. DOI: 10.1177/0741088307305977

Bretag, T., & Mahmud, S. (2009). Self-plagiarism or appropriate textual re-use? *Journal of Academic Ethics, 7*, 193-205. DOI: 10.1007/s10805-009-9092-1

Brin, S., Davis, J., & Garcia-Molina, H. (1995). Copy detection mechanisms for digital documents *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data* (ACM), pp: 398-409. DOI: 10.1145%2F223784.223855

Caporossi, G., & Leblay, C. (2011). Online writing data representation: A graph theory approach. *In Proceedings of the 10th International Conference on Advances in Intelligent Data Analysis X* (Porto, Portugal, October 29-31), 80-89. DOI: 10.1007/978-3-642-24800-9_10

Cress, U., & Kimmerle, J. (2008). A systemic and cognitive view on collaborative knowledge building with wikis. *International Journal of Computer-Supported Collaborative Learning,* 3(2), 105-122. DOI: 10.1007/s11412-007-9035-z

Garner, J. R. (2011). Combating unethical publications with plagiarism detection services *Urologic Oncology: Seminars and Original Investigations 29*(1), 95-99. DOI: 10.1016/j.urolonc. 2010.09.016

Groenendijk, T., Janssen, T., Rijlaarsdam, G., & Van den Bergh, H. (2013). The effect of observational learning on students' performance, processes, and motivation in two creative domains. *British Journal of Educational Psychology, 83*(1), 3-28. DOI: 10.1111/j.2044-8279.2011.02052.x

Ji, H., Favre, B., Lin, W. P., Gillick, D., Hakkani-Tur, D., & Grishman, R. (2013). Open-Domain multi-document summarization via information extraction: Challenges and prospects. In T. Poibeau, H. Saggion, J. Piskorski, & R. Yangarber *Multi-source, Multilingual Information Extraction and Summarization* (pp. 177-201). Springer Berlin Heidelberg.

Khmelev, D., & Teahan, W. J. (2003). A repetition-based measure for verification of text collections and for text categorization. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* pp:104-110. DOI: 10.1145/860435.860456

Lindgren, E., Leijten, M., & Van Waes, L. (2011). Adapting to the reader during writing. *Written Language & Literacy, 14*(2), 188-223. DOI: 10.1075/wll.14.2.02lin

Leijten, M., & Van Waes, L. (2013). Keystroke Logging in Writing Research Using Inputlog to Analyze and Visualize Writing Processes. Written Communication, *30*(3), 358-392. DOI: 10.1177/0741088313491692

Liptak, A. (2000). Paper Chase: An English professor tells how he tracks down anonymous authors. *New York Times Book Review*, November 26, 2000. Direct: NY Times

Lonergan, D. C., Scott, G. M., & Mumford, M. D. (2004). Evaluative aspects of creative thought: Effects of appraisal and revision standards. *Creativity Research Journal, 16*(2-3), 231-246. DOI: 10.1080/10400419.2004.9651455

MacArthur, C. A. (2009). Reflections on research and writing and technology for struggling writers. *Learning Disabilities Research & Practice, 24*(2), 93-103. Mount, D. W. (2001). Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press. ISBN: 0879696087

Mount, D. W. (2008). Comparison of the PAM and BLOSUM amino acid substitution matrices. *Cold Spring Harbor Protocols*. DOI: 10.1101/pdb.ip59

Mumford, M. D., Medeiros, K. E., & Partlow, P. J. (2012). Creative thinking: Processes, strategies, and knowledge. *Journal of Creative Behavior, 46*(1), 30-47. DOI: 10.1002/jocb.003

The National Archives. (2007). NARA existing digital copies to be gathered and made available online. Press Release. Retrieved form [http://www.archives.gov/comment/nara-digitizing-plan.pdf], September 10, 2007.

Pifarré, M., & Fisher, R. (2011). Breaking up the writing process: How wikis can support understanding the composition and revision strategies of young writers. *Language and Education, 25*(5), 451-466. DOI: 10.1080/09500782.2011.585240

Roig, M. (2005). Re-using text from one's own previously published papers: An exploratory study of potential self-plagiarism. *Psychological Reports, 97*(1), 43-49. DOI: 10.2466/pr0.97.1.43-49

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, & B., Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research, 13*(11), 2498-2504. DOI: 10.1101/gr.1239303

Silveira, R. (1999). The relationship between writing instruction and EFL students' revision processes. *Linguagem & Ensino* 2(2), 109-127. URL: UCPEL Library

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). Analysis of collaborative writing processes using revision maps and probabilistic topic models. *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK 2013, Leuven, Belgium*, pp. 38-47, DOI: 10.1145/2460296.2460307

The White House, Office of the Press Secretary. (2000). President Clinton announces the completion of the first survey of the entire human genome. Retrieved from [web.ornl.gov/sci/techresources/Human_Genome/project/clinton1.shtml], June 25, 2000.

Widlund, H. R., Cao, H., Simonsson, S., Magnusson, E., Simonsson, T., Nielsen, P. E., Kahn, J. D., Crothers, D. M., Kubista, M. (1997). Identification and characterization of genomic nucleosome-positioning sequences. *Journal of Molecular Biology, 267*, 807-81. DOI: 10.1006/jmbi.1997.0916

Wingate, U., & Tribble, C. (2011). The best of both worlds? Towards an English for academic purposes/academic literacies writing pedagogy. *Studies in Higher Education, 37*(4), 481-495. DOI: 10.1080/03075079.2010.525630