4-1-2011

Comparing nearly identical treaty texts: A note on the *Treaty of Fort Laramie with Sioux, etc., 1851* and Levenshtein's edit distance metric

Charles D. Bernholz
University of Nebraska-Lincoln, cbernholz2@unl.edu

Brian L. Pytlik Zillig
University of Nebraska-Lincoln, bzillig1@unl.edu

Published in *Literary and Linguistic Computing* 26:1 (April 2011), pp. 5–16; doi: 10.1093/llc/fqq016

Published online September 21, 2010.

Comparing nearly identical treaty texts: A note on the *Treaty of Fort Laramie with Sioux, etc., 1851* and Levenshtein's edit distance metric

Charles D. Bernholz and Brian L. Pytlik Zillig

Love Memorial Library, University of Nebraska–Lincoln, Lincoln, NE 68588, USA

Corresponding author – Charles D. Bernholz, Love Memorial Library, University of Nebraska–L:incoln, Lincoln, NE 68588, USA. **email** cbernholz2@unl.edu

Abstract

Vladimir Levenshtein's edit distance algorithm is used to reveal disparities between delimiter stripped texts of the Senate amended *Treaty of Fort Laramie with Sioux, etc., 1851* as corrected in a previous study, and of other federal copies of this transaction. All of the latter deviated markedly from that newly created version, reflecting errors of exclusion, of the absence in some transcripts of the Senate modification, of editorial decisions made by Charles J. Kappler during the preparation of his treaty compilations at the beginning of the twentieth century, and of spelling. These results confirmed that the instrument was until now never published in its complete formal state. This study may serve as a model for future text analyses that might benefit from the employment of Levenshtein's metric.

1 Introduction

More than 80 years ago, Wroth (1926, p. 749) observed that American Indian treaties are "a literary type that has been neglected by readers and teachers of early American literature," and he proposed that "these printed documents [are] the single original American contribution to the types of literary expression." Clearly, Wroth had in mind those early instruments created between Britain and the tribes, wherein discourse centered upon creating enduring friendships. Later transactions involving the new federal government appeared as more contract-like certificates, each conveying—directly, yet at some social expense—the law of the land. In total, 375 such contacts are acknowledged today by the Department of State (see, for example, *Ratified Indian Treaties, 1722–1869*, 1966).

In a previous analysis of the *Treaty of Fort Laramie with Sioux, etc., 1851* (see Kappler, 1904b, pp. 594–6; henceforth *Fort Laramie*), the published federal texts for this instrument were examined in order to produce a final, correct version of this transaction (denoted *Laramie09*), reflecting both the original document's material as well as an amendment made by the Senate to one of its articles (Bernholz and Pytlik Zillig, 2009). As one result of that study, it was concluded that *Fort Laramie* had never been published in an error-free state.

The text of this contract with nine tribes of American Indians consisted of a preamble, eight articles, and a testimonium that announced forthcoming signatures. Seven sources were assessed during these comparisons: that furnished by digital images of the original 1851 treaty itself, now held at the National Archives and Records

Administration (*Laramie51*); by the Senate's 1852 Confidential Executive Document used during the ratification process (Articles of a treaty, 1852, pp. 1–3; *Laramie52*); by two nineteenth century summary examples, one from each of *A Compilation of All the Treaties Between the United States and the Indian Tribes Now in Force as Laws* (1873, pp. 1047–50; *Laramie73*) and *Laws of the United States Relating to Indian Affairs* (1884, pp. 317–20; *Laramie84*); by the two items published in Charles J. Kappler's 1903 (pp. 440–2; *Laramie03*) and 1904 (pp. 594–6; *Laramie04*) second volumes of *Indian Affairs: Laws and Treaties*; and by the passage offered in his 1929 fourth volume of that series (pp. 1065–7; *Laramie29*). The 1903 example did not contain the signatures of the event participants, so Kappler removed the testimonium found in the other reports.

Comparing content is a fundamental process in text analysis and during the formation of *Laramie09*, many questions arose regarding the discrepancies found among the array of available alternatives for each textual term. Setting aside the understandable changes induced by the Senate's annuity modification—wherein it was ordered that two words of Article 7 were to be replaced by thirty to describe that alteration—and given the reasonable expectation of identical or nearly identical subsequent accounts following that mandate, the observed dissimilarities presented interesting insights into the provenance of these documents. This was especially so when an evaluation was conducted between the treaty's original 1851 wording and that of each of the three renditions found in Kappler's *Indian Affairs* (1903, 1904b, and 1929). These later copies had been compiled in response to a Senate request for an up to date catalog of relevant Indian affairs materials (Compilation of Indian Affairs, 1902), where his second *Treaties* volume in each of the 1903 and 1904 series was reserved for these specific instruments alone.

Unfortunately, and as the result of errors committed by the Department of the Interior, *Fort Laramie* was never appropriately published in the *Statutes at Large*, as is required by law (1 *Stat.* 187), so only unofficial copies endure. The general availability of Kappler's *Indian Affairs* therefore assured during the last century that his collations—and in particular, the 1904 one— became the most frequently used record of this event's parameters, as well as for those of almost all of the other federally recognized treaties with the tribes. Francis Paul Prucha stated that Kappler's 1904 *Treaties* edition "follows a chronological arrangement and relies for its texts on the *Statutes at Large* (although it prints the Fort Laramie Treaty from 1851, which the *Statutes* omits), and in most cases, it prints the treaty *as amended* instead of the original treaty with amendments printed at the end, as the *Statutes* does" (1994, pp. 523–4; emphasis original).

Besides the later expected alteration motivated by the introduction of the annuity adaptation, the observation of a variety of text exclusions, incursions, and replacements in all other versions relative to the original 1851 material led to an additional matter. *Laramie09*, instead of the 1851 instrument, could potentially serve as the standard against which all previous descriptions might be compared, especially if it could be established that the content of that revised entity was much closer to that of the original 1851 document than it was to any of the other remaining federal renditions. Under this scenario, the need to index inherent dissimilarities was critical, since Kappler had incorporated the Senate adjustment into each of his three twentieth century transcripts, thereby approximating the expected accurate account. Ultimately, *Laramie09*—the 1174 word, 5730 byte record fashioned during the text analysis study—captured the original 1851 text (including punctuation) which was then altered solely by integrating the allowance modification into Article 7.

Absent punctuation, spelling, and insignificant replacement differences across versions, the pivotal inconsistencies are identified in Table 1. First, just *Laramie51*, *Laramie52*, and *Laramie84* contained a boundary sub-specification—"*… thence from the mouth of Powder River….*"—found in the original Gros Ventre, Mandan, and Arikara portion of Article 5 (henceforth the Gros Ventre exclusion). Second, *Laramie73* and the three Kappler descriptions were the only ones that presented the updated annuity clause of Article 7 that the

Table 1. Indicated presence of the Gros Ventre boundary phrase, the revised payment parameter, and the testimonium in the seven federal versions of the *Treaty of Fort Laramie with Sioux, etc., 1851*, for the years 1851 through 1929, and of the proposed corrected rendition.

	Laramie51	*Laramie52*	*Laramie73*	*Laramie84*	*Laramie03*	*Laramie04*	*Laramie29*	*Laramie09*
Article 5 boundary text for Gros Ventre, Mandan, and Arikara	√	√		√				√
Senate-revised annuity parameters			√		√	√	√	√
Testimonium	√	√	√	√		√	√	√

Senate had developed during ratification. Finally, this Table further confirms that *Laramie09* employed the relevant Gros Ventre boundary phrase, the revised payment parameter, and the testimonium, thereby more closely forming the expected—and legally required—final narrative that should have been part of the *Statutes at Large*.

2 Levenshtein's Edit Distance

Vladimir Levenshtein, the 2006 IEEE Richard W. Hamming Medal winner "for exceptional contributions to information sciences, systems and technology," wrote 40 years ago (1966) on the transfer of information, where the three operations of deletion, insertion, and substitution may be engaged to correct, at minimum cost, errors contained in a transmitted string. Kruskal (1983, p. 216), in an overview of sequence comparison applications, used the two terms *industry* and *interest* to illustrate two ways to use six single byte deletions, substitutions, and insertions—or six substitutions alone—to transform the first word into the second. Thus, the minimum cost to "correct" *industry* into *interest* would be the sum of the costs associated with those six elementary processes, wherein each action has some given weight greater than or equal to zero. At an assigned weight of one for any such deletion, substitution, or insertion, the calculation provides a direct insight into the expense required to effect the change.

In a multiple element example, Soukoreff and MacKenzie (2001) used the model *quick brown fox* as a prototypic presented text and *quixck brwn fox* as the corresponding transcribed one. Even though there might be as many as six errors in such a communication—as shown by the failure of the transcribed substring *xck br* to match the presented one *ck bro*—the two most likely errors committed during this text entry task were those of the insertion of the character *x*, and of the omission of the character *o*. Levenshtein's edit distance (LED) algorithm, which considers a byte as the logical unit of measurement, would thus yield a score of 2: the sum of the deletion operation of *x* and of the insertion one of *o*. It follows that for such string pairs that prove to be identical, the computed LED value would be zero. Furthermore, the observed edit distance amount must be less than or equal to the maximum of the two string lengths; replacing a missing sequence with one of length *n* bytes costs no more than *n* operations. Tennison (2007) offers one version of LED implementation code that served as the basis for this examination.

For the assessment conducted here, the original 1851 treaty served as the base document. The test data were taken from the earlier study and consisted of a vertically aligned joint set of the various transaction texts, excluding delimiters and constructed to a uniform length, where a specific document's alignment was augmented by blank rows to fill in any absent subsections of that version; see Table 1 of the Web site The *Treaty of Fort Laramie with Sioux, etc., 1851*: Revisiting the Document Found in Kappler's *Indian Affairs: Laws and Treaties* (Bernholz and Pytlik Zillig, 2009; http://treatyoffortlaramie1851.unl.edu/). As an example of this procedure, a subsection of two parallel instruments might have a token difference such as *head men* versus *headmen* which, when placed in these arrays, required their sequences to occupy first two lines, and then only one location, supple-

mented by a subsequent blank corresponding to the former's *men* term. The application of the LED algorithm to this portion of the data would compute an edit distance for each of the *head* versus *headmen* and the *men* versus blank rows, and would return individual values of 3 and 3, respectively. Note that a contrast between *head men* and *headmen* as two strings would generate an LED of just 1, i.e. through the deletion cost of the blank separating the two words in the first target. The vertical text distribution format employed here thus maximized the potential cumulative LED scores.

In addition, the process returns evidence of all text differences, including those of capitalization. Inspection verified that the tokens in these treaty documents fluctuate markedly on this characteristic. Since the main objective was concerned with the *contents* of the accord rather than with its presentation, all materials were normalized to lower case prior to similarity testing.

Kruskal (1983; see also the subsequent volume by Sankoff and Kruskal, 1999) made particular reference to Levenshtein's work and noted that textual comparisons may be either simple or complex. The application of this algorithm has been found useful in the development of spell checkers (Kukich, 1992) and plagiarism detectors (Zini *et al*., 2006). Far more complex undertakings have included coordinating more than fifty versions of Chaucer's *The Canterbury Tales* (Spencer *et al*., 2003; see also Bordalejo, 2002 and 2003); nucleic acid sequencing (Waterman, 1984); and handwriting analysis (Seni *et al*., 1996). More recently, Levenshtein's algorithm has been cited as an efficacious approach and/or as a component in a variety of US patents, addressing such diverse systems for aggregating traveler information (Gueziec, 2005); for resolving an incorrectly entered uniform resource locator (Chudnovsky and Chudnovsky, 2008); and for the syntactic pattern recognition of sequences (Badr and Oommen, 2010). In reality, one of the claims in the Badr and Oommen specifications was that the "present invention can be used to locate subsequences in sequences when the former are inaccurately represented. Thus, the invention has potential applications in the human or other genomic projects, in the detection of targets for diseases, and ultimately in the drug-design process," thereby expanding further upon the nucleic acid research from three decades earlier.

The application of LED to the materials in the present study might be considered similar to the Chaucer examination. In fact, ours was a far less complex endeavor, primarily because the *Fort Laramie* text is a treaty between sovereigns composed of an identifiable sequence of ordinally arranged articles. Aust (2007, p. 16) remarked that "most treaties are drafted according to standard forms and processed according to long-established procedures." Absent the question of promulgation, there is nothing in the history of the construction of *Fort Laramie* that would suggest that it departed from the manner in which any of the other treaties with the tribes was formed by the United States. In contrast, *The Canterbury Tales* material is, according to Spencer and his colleagues, "a series of looselyconnected stories … [that] show many different orderings of the tales and linking passages … largely due to rearrangements of items (tales and links) by scribes, who found it difficult to establish an appropriate order even in the earliest manuscripts." In order to attack this problem, "methods developed for the analysis of gene order [were used] to produce a stemma based on the order of tales and links in *The Canterbury Tales*" (Spencer *et al*., 2003, pp. 97–8). Furthermore, Van Reenen and Van Mulken (1996, p. ix) spoke to the degree of computational analysis required within stemmatology for "cases of extensive manuscript traditions or highly contaminated traditions" by observing that "[i]t comes as no surprise that those who wish to solve the problems of large and entangled manuscript traditions invoke the help of computer science." Their conclusions proposed that "in order to detect the kinship relations among manuscripts three stages must be discerned: the unrooted deep structure, the underlying intermediate structure, and the rooted oriented structure."

Fort Laramie did not have such a convoluted past, nor would any federal employee, in either transferring or reproducing this instrument, have taken the liberties that those scribes apparently took as they rearranged the *Tales*. The rigidity of the treaty format expedited simpler ed-

itorial activities, and permitted an avoidance of Van Reenen and Van Mulken's scenario, as well as the one engendered in Spencer and Howe's statement that, in some collation studies, investigators "may not be able to identify the corresponding locations in witnesses if each has been compared to a base text" (2004, p. 255). In *Fort Laramie*, uncomplicated procedures to detect element errors and/or inexact passage reproduction were paramount. These activities were advanced by that formal arrangement: its stiffness led to the selection of an easy yet beneficial design to concatenate, and then to visualize and assess, the data. The vertically aligned joint set of texts immediately made evident exclusions or incursions among the sources. Under these less difficult conditions, the calculation of LEDs was straightforward, avoiding a setting in which the processing software sometimes "gets totally lost" due to manuscript variance (Gilbert, 1973, p. 145). As a windfall within this examination, this algorithm was far more intuitive than some of the proposals enumerated by Spencer and Howe (2001).

3 Assignable Costs

The known differences among these *Fort Laramie* documents, explicitly exemplified by the chronological inability of the original contract to hold the final Senate annuity proposal, meant that there was a finite set of expected deviations across these examples. Incorporating that Senate update thereby induced a degree of divergence between the wording of the 1851 and the 1852 pre-adjustment transcripts, and of those of some later versions. This introduction also guaranteed that the value of the cumulative LED, computed between the original passages and those of both the ensuing accounts and *Laramie09*, would never be zero, but would instead reflect at a minimum one or more fixed offsets of some non-zero length.

As one specific instance of this situation, revising the original 1851 rendering to reflect the legislated annuity change required the amendment activity specified by the Senate: "Article 7, strike out the words 'fifty years,' and insert in lieu thereof the following: *the term of ten years, with the right to continue the same, at the discretion of the President of the United States, for a period not exceeding five years thereafter*" (Journal of the Senate of the USA, 1852, p. 703; emphasis original). Note, however, that the punctuation in the specifications of the Senate's annuity correction varies in official presentations. The cited 1852 Journal of the Senate statement contains two commas—one after each of the terms *same* and *States*—that appear in no other renditions. Volume 8 of the *Journal of the Executive Proceedings of the Senate of the United States of America* eliminated those two markings as well (1887, p. 389). The correct 1852 passage, as approved by the Senate, was applied to *Laramie09*. The individual LED cost for this operation consisted of a 134 byte maneuver that substituted those thirty words for the initial two terms *fifty years*. Similarly, inserting the Gros Ventre boundary parameter into the 1904 text, from which it is absent, necessitated a thirty-one byte repair, while the replacement of the deleted testimonium from Kappler's shortened 1903 version of *Fort Laramie* amounted to a 195 byte procedure.

These three established problems—the Gros Ventre exclusion, annuity amendment, and testimonium—formed the potential major offsets required by members of this document suite. The adjustments represented 0.5, 2.3, and 3.4%, respectively, of *Laramie09's* total byte count. A specific comparison's cumulative LED cost, with or without such offset expense(s), was then supplemented by the dissimilarity costs of such tokens as *Yellow Stone* versus *Yellowstone*, *head men* versus *headmen*, and *southwesterly* versus *south-westerly*, if any. This latter noise, mirroring perhaps editorial decisions and/or transcription errors as well as the nature of the vertical data format used in these inquiries, further clouded the true underlying divergence among the texts. Table 1 catalogs problems appearing in the various models of *Fort Laramie* and shows that only *Laramie03* had two missing elements: the Gros Ventre exclusion and the testimonium.

4 LED Testing and Results

An initial trial was performed on *Laramie51* and *Laramie52*, in order to ascertain whether the Senate had worked with a markedly different doc-

Table 2. LED error table for six federal versions of the *Treaty of Fort Laramie with Sioux, etc., 1851*, for the years 1852 through 1929, when compared with the proposed 2009 rendition, enumerating the total number of comparative textual errors; their cumulative byte amounts; the expected cost for text replacement, given the known absences identified in Table 1; and an index of the resulting document disparity noise, derived by subtracting the expected cost from the observed cumulative LED

	Laramie09	*Laramie52*	*Laramie73*	*Laramie84*	*Laramie03*	*Laramie04*	*Laramie29*
N errors	30	107	99	102	135	94	124
Cumulative LED	134	457	418	461	595	393	518
Expected cost	134	134	31	134	226	31	31
Noise	0	323	387	327	369	362	487

ument than the original treaty. The first assessment concerned both full texts, including their capitalizations. There were 135 observed errors generating a cumulative LED of 387 bytes in this comparison; 64 bytes were due to capitalization differences, i.e. to examples in the parallel data such as *Affairs* versus *affairs* or *Territory* versus *territory*. When the lower case contents were readdressed, the cumulative LED was automatically reduced by these sixty-four bytes to a sum of 323, but the new error count—seventy-seven—was not diminished by a similar integer amount because capitalization was just one type of error that affected the cumulative LED scores in these comparisons. The first examination had illustrated both problems of capitalization and bifurcated word inequality by the frequent occurrence of disagreements, such as *Head men* versus *headmen*.

Succeeding judgments were made between the lower case texts of the original treaty and the proposed corrected model, i.e. between *Laramie51* and *Laramie09*, and of *Laramie09*—employed as the new standard for *Fort Laramie*—with each of *Laramie52*, *Laramie73*, *Laramie84*, *Laramie03*, *Laramie04*, and *Laramie29*. Table 2 conveys for these trials the detected error counts; the cumulative LED scores; the expected costs sustained to reintroduce one or more missing subsections of relevant matter; and noise estimates founded upon the differences between the cumulative LED scores and the expected costs to bring each pair of test documents into register.

Finally, the preamble and articles alone of *Laramie03* and *Laramie04* were examined to detect the degree of reliability between Kappler's two editions of *Fort Laramie*. He had declared in the preface to the second edition (1904a, p. v) that "[t]he new edition has afforded the compiler an opportunity … to add the signatures subscribed to each treaty which was omitted in the first edition to save space," and reinstituted the testimonium into the latter.

Along with the observed LED score, Table 2 directly quantifies the expected costs attributable to the shortfalls identified in Table 1, as well as, by subtraction, amounts of additional textual noise. These data offer an intuitive understanding of the magnitude of the divergences among these renditions. With particular reference to the three fixed offset values in mind, the LED test results are briefly summarized as follows.

4.1 Laramie09 versus Laramie51

The disparity between the original transaction and *Laramie09* is immediately evident in the expected offset cost of 134 bytes for the thirty word annuity amendment, but there were no further costs— acknowledged by the noise score of zero—since *Laramie09* was based on a direct copy of *Laramie51*.

4.2 Laramie09 versus Laramie52

Similarly, the Senate's working copy only considered the 1851 wording, and might in fact be the only account directly reproduced from the original parchment. The 457 byte cumulative LED cost included the expected annuity offset of 134 bytes, but it also demonstrated—through a total of 107 errors—that there were many more induced changes. This rendition and *Laramie84* are the only two beyond the original text that maintained the Gros Ventre boundary specification,

thereby saving the modest expense of thirty-one bytes, but this revelation also supports the contention that at least the former was taken *directly* from the 1851 *Fort Laramie* contract itself. As possible evidence to support the application of this hypothesis to the second transcript, the 1884 *Laws of the United States Relating to Indian Affairs* volume was, according to its title page, "compiled by the Indian Office" where at least one of the "fifteen handwritten copies" of the treaty must have been accessible (see VanDevelder, 2009, p. 196, who also provided Kappler's 1904 *Fort Laramie* in an appendix on pp. 245–7).

4.3 Laramie09 versus Laramie73

Ninety-nine errors, at a penalty of 418 bytes, illuminated a substantial divergence between this pair, especially when it is considered that the latter already held the annuity amendment and was only missing the thirty-one byte Gros Ventre exclusion. The noise was apparently caused in part by numerous examples of *Yellow Stone* versus *Yellowstone*, *head men* versus *headmen*, *head waters* versus *head-waters*, and *Twenty five Yard Creek* versus *Twenty-five Yard Creek* entries in the two files, respectively.

4.4 Laramie09 versus Laramie84

The expense of 461 bytes involving 102 errors in this test was driven by the fact that the 1884 report must be a reproduction of the original and/or of the 1852, pre-Senate amended material. Thus, the required 134 byte annuity cost is a recognized portion of the instrument's returned overall LED score. Clearly, the editors of *A Compilation of All the Treaties Between the United States and the Indian Tribes Now in Force as Laws* (1873) failed to provide a robust account. In 1900, and again in 1901, the Commissioner of Indian Affairs, William A. Jones, criticized the overall *Compilation* when he remarked that it was "inaccurate" (Annual reports of the Department of the Interior for the fiscal year ended 30 June 1900, 1900, p. 50, and Annual reports of the Department of the Interior for the fiscal year ended 30 June 1901, 1901, p. 47). The observed cumulative LED found in the present comparison may offer corroborating evidence for that accusation.

4.5 Laramie09 versus Laramie03

Kappler's decision to remove the testimonium from his 1903 version now necessitated an expense to absorb that final section from *Laramie09*'s complete transcript. The total of 135 errors and a cumulative LED of 595 incorporated the known costs for *two* missing components: the thirty errors and 195 bytes associated with providing the deleted testimonium, and the seven word, Gros Ventre exclusion worth thirty-one bytes. The prompt convergence to an outlay of 400 bytes, following the subtraction of the expenditure for the testimonium to be inserted, brings the cumulative LED cost for *Laramie03* very close to that of *Laramie04*. Table 1 shows that both failed to supply the Gros Ventre exclusion. See the supplementary assessment between these two versions below.

4.6 Laramie09 versus Laramie04

The observed variation between the new proposed standard for *Fort Laramie* and Kappler's 1904 well-used one yielded the lowest cumulative LED cost. There were still several reasons for the remaining 250þ byte dissimilarity—the Gros Ventre exclusion was just one—but when all the LED scores in Table 2 were contrasted, Kappler's *Laramie04* provided the best approximation to the true, complete rendering of the event as postulated by *Laramie09*.

4.7 Laramie09 versus Laramie29

The LED score for this comparison was amplified by the incursion of the sixteen-word and sixty-two byte phrase *thence up the north fork of the Platte River to the forks of the Platte River* in *Laramie29*'s Sioux reservation parameters of Article 5. Thus, the cumulative noise score of 487, that far outdistanced that calculated for all other renditions, was affected by this production error; by the shortcut use of numeric values for *$50,000* and *10* in the amount and duration aspects of the annuity definition; and by the use of *Art* for introducing each Article. The expected cost was only thirty-one bytes for the absent Gros Ventre exclusion. Clearly, the lack of precision in reproducing the 1929 version is unfortunate, but the use of *Fort Lara-*

mie in Kappler's volume 4 was more for general information regarding the evolution of the tribal assents required for the Senate's annuity adjustment, and less for the legal text demands of Congress; the latter need was addressed properly by the earlier 1904 edition of *Treaties*.

4.8 Laramie03 versus Laramie04

Following the general testing against the *Laramie09* model, an additional investigation involved Kappler's two main editions of Fort *Laramie*. The *Laramie09* results had indicated that the 1903 and the 1904 versions converged to within just seven bytes, after subtracting the testimonium offset of 195 bytes from *Laramie03*'s returned cumulative LED score. Such similarity—a 400 net cumulative LED for *Laramie03* versus one of 393 for *Laramie04*—validated the hypothesis that Kappler reproduced the former for use in the second edition. Upon closer inspection, there was an actual difference of thirteen bytes found during a comparison of only the preamble and article sections of *Laramie03* and *Laramie04*, wherein the latter's testimonium segment was removed from that analysis. The outcome revealed that *Laramie04*'s initial total cumulative LED score relative to *Laramie09* had been inflated by six bytes, due to a common spelling inconsistency—*headmen* versus *Head men*, respectively—located in its testimonium.

4.9 Laramie04 versus Laramie29

The final examination assessed the differences between *Laramie04* and *Laramie29*. The test was conducted between these two—and not between *Laramie03* and *Laramie29*—because *Laramie03* does not possess the testimonium section. The cumulative LED for the *Laramie04* and *Laramie29* consideration produced forty-seven errors encompassing a total of 175 bytes. Removing the known cost for *Laramie29*'s unique sixteen-word incursion from these sums, the final tally is thirty-one errors with a cost of 113 bytes. Both amounts suggest a poor reproduction of *Laramie04* and a significant departure from the accuracy demonstrated in the conversion of *Laramie03* into *Laramie04*.

5 Conclusions

Text analysis can be a complex undertaking when disparate contents are compared. Roos and Heikkila (2009, p. 417) cited Jorge Luis Borges—"No book is published without some discrepancy in each one of the copies. Scribes take a secret oath to omit, to interpolate, to change"—to highlight their evaluation of stemmatological methods. If the underlying fundamental questions pertaining to provenance (or, perhaps, to malicious intent) are disregarded for the moment, the twists and turns of such allegedly (yet rarely) identical materials are still capable of providing interesting insights. Indeed, this was one of the outcomes of the initial study described here: it was found that the *Treaty of Fort Laramie with Sioux, etc., 1851* had never been published in a complete and accurate form during the last century and a half. This discovery was expedited by Levenshtein's algorithm that promptly underlined the divergences and—just as importantly—their magnitudes.

For other research efforts into the realms of punctuation and of capitalization, the LED process may be even more useful, because it recognizes *every* distinction between compared documents. Brossard (1945) has a discussion of the use of punctuation in statutes, while *Ewing v. Burnet*, before the US Supreme Court, observed that "[p]unctuation is a most fallible standard by which to interpret a writing; it may be resorted to when all other means fail. . ." (1837, p. 54). It is also relevant to acknowledge that paragraph 2 of Article 3, section 3 of the US Constitution, pertaining to the punishment of treason, suffered from questionable punctuation (Boutwell, 1895, pp. 321–3). American Indian materials are a rich mine of such data: there were eighty-eight commas, thirty-eight en dashes, thirty-five periods, six semicolons, and four colons, as well as nineteen *River* and twenty-five *river* terms, among the original text of *Laramie51*. These are the very typographical characteristics towards which all other copies of *Fort Laramie* should have converged, if the intent was to truly reproduce the original, and so the quest for fidelity required *Laramie09* to be constructed directly from the fabric of *Laramie51*,

and tailored only to furnish the amended annuity parameters.

In terms of consistency, the observed noise scores might mirror the effect of compositor carelessness. Boutwell (1895, p. 322) noted that such errors were well known and that federal "engrossing clerks ma[d]e mistakes not only in punctuation, but even in words and paragraphs," necessitating remedies that included legislation such as "An act to perfect the revision of the statutes of the United States, and of the statutes relating to the District of Columbia" (1877; 19 *Stat*. 240). Unfortunately, this fourteen page act began with the statement "Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, That for the purpose of correcting *errorors* and supplying omissions in the act entitled…" (19 *Stat*. 240; emphasis added). Additionally, a standardized approach might have been taken by government editors to tackle the orthography of instruments like *Fort Laramie*, where contents were sometimes purposely reformatted. Such strategies included deploying the "Names of Indian Tribes and Bands" list, published in 1900, that "[t]he Bureau of Ethnology and the Indian Bureau [had] undertaken to secure uniformity in the spelling of the names of Indian tribes and bands," and which the Government Printing Office published for use by federal agencies (Annual reports of the Department of the Interior for the fiscal year ended 30 June 1900, 1900, p. 687). Kappler used this new tribe name array, but chose to employ it only for the treaty titles in his 1903 and 1904 compilations and to leave unmodified the final texts derived from the *Statutes at Large* (Bernholz, 2010).

Indeed, outcomes such as these have already served elsewhere to stimulate similar text analyses, where the rewards might be significant for uncovering minute discrepancies between alleged copies, or among successive editions. While a contrast between the preamble and articles texts of *Laramie03* and *Laramie04* uncovered just a thirteen byte divergence contained in two errors, discovered differences may be far less limited. After removing punctuation, a cumulative LED score of 14 was generated in an evaluation of just the first line of Walt Whitman's 1855 original and of his later 1891 revision of *I Sing the Body Electric*, i.e. for a test between "The *bodies* of *men and women* engirth me and I engirth them" and "The *armies* of *those I love* engirth me and I engirth them". The italicized terms enumerate the four pairs of errors across just these twelve words.

The observed differences between the 1852 Journal of the Senate entry for the annuity redefinition and the other *Fort Laramie* accounts are a pertinent materialization of such recompense. In the present instance, the non-existence of the testimonium, or the failure to present those recognized annuity parameters following ratification, yielded a blatant lack of correspondence; even the Gros Ventre exclusion—at just 0.5% of the overall byte length for the document—subtended a highly visible thirty-one bytes. Additionally, the average cumulative noise LED score across the lower case test versions of *Laramie52* through *Laramie04* was 354, representing fully 6% of the total document's span and an amount almost equivalent to the sum of the offset costs for all three known major text faults. This latter finding alone validated the decision to create *Laramie09* directly from the original *Fort Laramie* itself, instead of by selecting terms from the pool of all previous noisy reproductions.

To sum up, even though the Gros Ventre boundary omission or the testimonium removal from Kappler's 1903 presentation of *Fort Laramie* has apparently violated neither the spirit nor the letter of federal law, assessing these American Indian materials through Levenshtein's algorithm and uncovering these difficulties has been an effective manifestation of its use to fathom such variances. *Fort Laramie* was a useful model, but the observation of a nonzero LED score is an immediate indication that *any* text may be ajar. As such, these data would echo two pertinent statements made by Kruskal (1983). First, he declared that "the Levenshtein distance between two sequences is a *plausible* indicator of the amount of actual historical change between them" (p. 232; emphasis added). Misspellings and/or corrections to texts subsequent to the original treaty induced such transitions. Second, in his attempt to address the validity of this approach—"Why Levenshtein dis-

tance?"—within any sequence comparisons, he proposed that a sufficient usage rationale might be formed upon "an application of the universal *post hoc* justification process. If we use any particular definition for distance, and find that this kind of distance supplies the information we want, that 'it works' when we check its performance, then the satisfactory performance justifies the definition. Every well-made application of distance contains such checking and supports this rationale" (p. 234). The relative cost analysis measurements made in the comparisons of these various renditions of the same instrument are a credible indicator of change *and* of departure from the initial document: the observed 124 errors and 487 bytes of noise in *Laramie29* should promptly instill some concern about that material's ability to reflect truly the 1851 text.

As noted above, Roos and Heikkila made use of Borges' *The Lottery in Babylon* to underscore the prevalence of imperfect textual materials. However, an additional sentence from the same story is especially relevant to this *Fort Laramie* examination: "The scribe who writes a contract almost never fails to introduce some erroneous information" (Borges, 2007, p. 35). Those very faults, immediately evident in the preamble of the original 1851 treaty document, helped form the conclusions derived from this study. These findings substantiate Kruskal's general observation—that Levenshtein's algorithm is a useful tool in such textual research—and recommend that future investigations might benefit from a similar deployment of this simple yet effective approach.

Acknowledgments

The authors thank Brett Barney, at the Center for Digital Research in the Humanities at the University of Nebraska-Lincoln, for suggesting Walt Whitman's *I Sing the Body Electric* passages to enrich this presentation.

References

A Compilation of All the Treaties Between the United States and the Indian Tribes Now in Force as Laws. (1873). Washington, DC: Government Printing Office.

Annual reports of the Department of the Interior for the fiscal year ended June 30, 1900. Indian Affairs. Commission to the Five Civilized Tribes. Indian Inspector for Indian Territory. Indian contracts. Board of Indian Commissioners. (1900). House of Representatives. 56th Congress, 2nd session. House Document No. 5, part 2.2 (*Serial Set* 4102). Washington, DC: Government Printing Office.

Annual reports of the Department of the Interior for the fiscal year ended June 30, 1901. Indian Affairs. Part I. Report of the Commissioner, and appendixes. (1901). House of Representatives. 57th Congress, 1st session. House Document No. 5, pt. 2-1 (*Serial Set* 4290). Washington, DC: Government Printing Office.

Articles of a treaty. (1852). Senate. 32nd Congress, 1st session. Senate Confidential Executive Document No. 11. Washington, DC: Government Printing Office.

Aust, A. (2007). *Modern Treaty Law and Practice.* 2nd ed. New York: Cambridge University Press.

Badr, G. and Oommen, J. B. (2010). *Method of syntactic pattern recognition of sequences. U.S. Patent No. 7,689,588*. Washington, DC: U.S. Patent and Trademark Office.

Bernholz, C. D. (2010). Standardized American Indians: The "Names of Indian tribes and bands" list from the Office of Indian Affairs. *Government Information Quarterly*, 27: 272–9.

Bernholz, C. D. and Pytlik Zillig, B. L. (2009). The *Treaty of Fort Laramie with Sioux, etc., 1851*: Revisiting the Document Found in Kappler's *Indian Affairs: Laws and Treaties,* http://treatyoffortlaramie1851. unl.edu/ (accessed 6 September 2010).

Bordalejo, B. (2002). *The Manuscript Source of Caxton's Second Edition of the Canterbury Tales and Its Place in the Textual Tradition of the Tales,* http://www.bordalejo.net/theses.html (accessed September 6, 2010).

Bordalejo, B. (2003). *The Phylogeny of the Tale-Order in the Canterbury Tales,* http://www.bordalejo.net/theses.html (accessed September 6, 2010).

Borges, J. L. (2007). *Labyrinths: Selected Stories & Other Writings*. New York: New Directions.

Boutwell, G. S. (1895). *The Constitution of the United States at the End of the First Century*. Boston, MA: D. C. Heath & Co.

Brossard, E. E. (1945). Punctuation of statutes. *Oregon Law Review*, 24: 157–72.

Chudnovsky, D. V. and Chudnovsky, G. V. (2008). *Method to resolve an incorrectly entered uniform resource locator (URL). U.S. Patent No. 7,376,752*. Washington, DC: U.S. Patent and Trademark Office.

Compilation on Indian Affairs. (1902). *Congressional Record*, 35: 5664–5.

Ewing v. Burnet, 36 U.S. 41 (1837).

Gueziec, A. (2005). *System for aggregating traveler information. U.S. Patent No. 7,161,497*. Washington, DC: U.S. Patent and Trademark Office.

Gilbert, P. (1973). Automatic collation: A technique for medieval texts. *Computers and the Humanities*, 7: 139–47.

Journal of the Executive Proceedings of the Senate of the United States of America, Vol. 8. (1887). Washington, DC: Government Printing Office.

Journal of the Senate of the United States of America, being the first session of the Thirty-second Congress; begun and held in the City of Washington, December 1, 1851, in the seventy-sixth year of the independence of the United States. (1852). Senate. 32nd Congress, 1st session (*Serial Set* 610). Washington, DC: Government Printing Office.

Kappler, C. J. (1903). Indian affairs. Laws and treaties, Vol. 2. Treaties. Senate. 57th Congress, 1st session. Senate Document No. 452, pt. 2 (*Serial Set* 4254). Washington, DC: Government Printing Office.

Kappler, C. J. (1904a). Indian affairs. Laws and treaties, Vol. 1. Laws. Senate. 58th Congress, 2nd session. Senate Document No. 319, pt. 1 (*Serial Set* 4623). Washington, DC: Government Printing Office.

Kappler, C. J. (1904b). Indian affairs. Laws and treaties, Vol. 2. Treaties. Senate. 58th Congress, 2nd session. Senate Document No. 319, pt. 2 (*Serial Set* 4624). Washington, DC: Government Printing Office.

Kappler, C. J. (1929). Indian affairs. Laws and treaties, Vol. 4. Laws. Senate. 70th Congress, 1st session. Senate Document No. 53 (*Serial Set* 8849). Washington, DC: Government Printing Office.

Kruskal, J. B. (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Review*, 25: 201–37.

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24: 377–439.

Laws of the United States Relating to Indian Affairs: Compiled from the Revised Statutes of the United States enacted June 22, 1874, and from Statutes at Large from that date to March 4, 1883: Also, Special Acts and Resolutions Previous to the Enactment of the Revised Statutes, not Embraced in or Repealed by the Revision: Also, List of all Ratified Treaties and Agreements Made with the Several Indian Tribes, 3rd edn. (1884). Washington, DC: Government Printing Office.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10: 707–10.

Prucha, F. P. (1994). *American Indian Treaties: The History of a Political Anomaly*. Berkeley, CA: University of California Press.

Ratified Indian Treaties, 1722–1869. (1966). Washington, DC: National Archives and Records Service.

Roos, T. and Heikkila, T. (2009). Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing*, 24: 417–33.

Sankoff, D. and Kruskal, J. (1999). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Stanford, CA: Center for the Study of Language and Information.

Seni, G., Kripasundar, V., and Srihari, R. K. (1996). Generalizing edit distance to incorporate domain information: Handwritten text recognition as a case study. *Pattern Recognition*, 29: 405–14.

Soukoreff, R. W. and MacKenzie, I. S. (2001). Measuring errors in text entry tasks: An application of the Levenshtein string distance statistic. *Companion Proceedings of the ACM Conference on Human Factors in Computing Systems – CHI 2001*. New York: Association of Computing Machinery, pp. 319–20.

Spencer, M. and Howe, C. J. (2001). Estimating distances between manuscripts based on copying errors. *Computers and the Humanities*, 16: 467–84.

Spencer, M. and Howe, C. J. (2004). Collating texts using progressive multiple alignment. *Computers and the Humanities*, 38: 253–70.

Spencer, M., Bordalejo, B., Wang, L.-S. *et al.* (2003). Analyzing the order of items in manuscripts of "The Canterbury *Tales*." *Computers and the Humanities*, 37: 97–109.

Tennison, J. (2007). Levenshtein distance on the diagonal. http://www.jenitennison.com/blog/node/12 (accessed January 28, 2010).

Van Reenen, P. and Van Mulken, M. (1996). *Studies in Stemmatology*. Philadelphia, PA: John Benjamins Publishing.

Van Develder, P. (2009). *Savages and Scoundrels: The Untold Story of America's Road to Empire through Indian Territory*. New Haven, CT: Yale University Press.

Waterman, M. S. (1984). General methods of sequence comparison. *Bulletin of Mathematical Biology*, 46: 473–500.

Wroth, L. C. (1926). The Indian treaty as literature. *Yale Review*, 17: 749–66.

Zini, M., Fabbri, M., Moneglia, M., and Panunzi, A. (2006). Plagiarism detection through multilevel text comparison. *Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution*. Washington, DC: IEEE Computer Society, pp. 181–5.