Comparing "parallel passages" in Digital Archives

July 16, 2019

Abstract

Keywords

Digital Archives Information Retrieval Statistical Language Models Suffixes Trees

1 Introduction

The term "parallel passage" refers to identical, or approximate text patterns of variable length, which could be regarded as semantically equivalent. "Parallel passages" represent alternative surface representations that exhibit identical wording, such as those representing reported speech and direct quotations, or with some small variation in grammatical structure, or vocabulary choice as a result of paraphrasing. On the one hand, differences in vocabulary choice may be the result of synonymy, or hyperonymy where a general or higher-level concept has been selected [Madnani and Dorr, 2010]. On the other hand, paraphrasing on the part of the author may provide evidence of text-reuse, or intertextuality [Fairclough, 1992], where the author has summarised the main concepts, or meaning, encoded by one or more texts that preceded it.

Further differences between passages may arise due to a shift in authorship, dialect, the natural evolution of language over time [Büchler, 2010], and errors introduced by optical character recognition (OCR) during the digitisation. The task of comparing equivalent or similar shared text patterns in text corpora stored in digital archives, has become increasingly challenging and time consuming due to the current scale of digital text data, which makes the task of comparing shared text patterns across multiple documents practically impossible to do manually. Identifying parallel passages, such as those exemplified by paraphrases, also supports a range of natural language tasks, including text generation, information retrieval and extraction, and summarisation.

This paper presents an overview of the text mining tools developed to compare parallel passages, which were deployed in a system known as the *Samtla* (Search And Mining Tools for Language Archives), which was developed to support the research of

historic and cultural heritage collections of documents stored in digital archives. The paper is organised as follows, in Section 2, we review the related work. Section 3 describes the corpora used as test cases to explore the results generated by our proposed approach. We provide a description of the model used as a basis for extracting and scoring the contents of documents according to their shared-text patterns in Section 4. In Section 5, we describe the approach used for identifying related documents according to our proposed model, where we measure the similarity of pairs of documents based on their character-level n-gram probability distributions. Section 6 presents an approach for visualising local similarities between the content of related documents in the form of variable length parallel passages extracted from the document content. We briefly discuss the motivation behind the user interface in Section 7, and some of the language and corpus dependent issues that the document comparison tool addresses to demonstrate the flexibility of the approach to different domains, languages, authors, and time periods in Section 8. We conclude the paper with a summary of the work in Section 9, and future research and development.

2 Related Work

Books, web pages, articles, and reports are all examples of unstructured text data where relevant information exists potentially anywhere within the document. Unstructured text data is often managed and retrieved via a search engine [Levene, 2010]. Search engines provide the means to retrieve information but not to analyse it, this is where text mining techniques are useful, as they provide different views of the data to facilitate the discovery and subsequent analysis of textual patterns [Aggarwal and Zhai, 2012]. These patterns can then be examined more closely through traditional research techniques such as close-reading of the text, but this is generally only possible for small scale digital archives.

One text analysis problem that is of great interest to researchers, particularly those analysing the content of digital archives, is to find *parallel passages* — text segments describing the same concept (entity or event etc.) over large corpora. Parallel passages are semantically similar and could exhibit identical wording, but quite often they exhibit some small variation in structure, or vocabulary choice. The differences are due to the normal rephrasing of the text within the same context, but may also arise from the use of reported speech, a change in authorship, dialect, the natural evolution of language over time, and errors introduced by optical character recognition (OCR). In this paper, the concept of parallel passage is defined in the general sense. Roughly speaking, one can regard parallel passages as a variable length structural text pattern. However, the term parallel passage probably originated from Christian theology, where the comparison of parallel passages, or *hermeneutics*, in the context of the Bible is a major area of Biblical scholarship (see [Strauss and Eliot, 1860]). The aim of *hermeneutics* is to render a translation of a text by comparing examples of the same word, phrase, or context across several texts. The researchers' task is to identify whether the differences present between one or more texts that are regarded as similar, is significant or relevant

Table 1: An example "parallel passage", appearing in [de Jong, 2007].

2 Kings, Chapter 16	*Isaiah, Chapter 7*
Then	In the days of Ahaz son of Jotham son of Uzziah, king of Judah
came up	came up
King Rezin of Aram and	King Rezin of Aram and
King Pekah son of Remaliah of Israel	King Pekah son of Remaliah of Israel
to wage war on Jerusalem;	to Jerusalem to attack it,
they besieged Ahaz	
but could not	but could not
prevail over him.	mount an attack against it.

to the research hypothesis whether the focus of research is on the stylistic differences between authors, or providing evidence of the evolution of textual sequences over time. The technique often involves comparing corresponding passages located across more than one text, by laying out the texts side-by-side. The Bible often describes the same event from different perspective across different canonical books, which can yield a more complete picture of the event than a single passage, or point of view on the subject. The example presented in Table 1 is from the *King James Bible*, and illustrates two "parallel passages" that would be regarded as highly similar by researchers of the Bible. These texts discuss the same event, and it has been proposed that *Isaiah, Chapter 7* was derived from the text of *2 Kings, Chapter 16* [de Jong, 2007]. The similarity between these two texts is not easily identifiable with current tools developed for search and mining of digital archives, due to the variability in the structure and choice of vocabulary, where *2 Kings, Chapter 16* adopts the phrase "wage war" over the word "attack" in *Isaiah, Chapter 7*, and the latter text is also more specific about the time of the event, as described in the introductory section of the text.

Text mining tools developed for the purpose of literary analysis of texts have actually existed since the 1940s, when researchers saw the immediate benefit of using computers to produce concordances of specific text patterns [Rockwell, 2003, Rommel, 2007, Schonfeld and Rutner, 2012, Sweetnam et al., 2012].

Document comparisons tools that highlight patterns of textual reuse can help researchers identify or describe cultural patterns. Without these types of comparison tools, the process of manually annotating each instance of the parallel passage, or short phrase, in question can be both time consuming and error prone depending on the complexity and volume of texts. This means that some corpora may remain unstudied due to the manual process involved in annotating instances of textual reuse across a body of text that has evolved over time [Lee, 2007].

Textual reuse was common in antiquity [Hoek, 1996], and authors rarely acknowledged the original source, which was in part due to the scarcity of published works from which to refer, meaning they would quote the source from memory resulting in an approximation of the original text. In addition, depending on the author, there

would be a tendency to insert new portions of text or paraphrase the original source to suit their specific style or purpose [Lee, 2007]. The ability to identify textual reuse and their original source can help establish the date of authorship showing the evolution of cultural practices or ideas. In addition, the presence and absence of a portion of quoted text provides some clues surrounding the thoughts and motivation of the author.

Consequently, comparing examples of textual reuse across a large body of related text in an automated, or semi-automated way, would be desirable for textual analysis and interrogation of written sources spanning several periods or dialects. Plagarism detection is also another potential application since it addresses similar issues, where portions of a source text have been copied and adapted to form a new text with no acknowledgement of the original author or source [Stamatatos, 2009].

Despite the early interest in the benefit of computational tools, there are still very few available for finding and analysing textual patterns of significance, in order to assist humanities researchers. A few notable examples exist, including the *Logos Bible* [log, 2016] and *BibleWorks* [bib, 2016] systems, which allow users to display and compare parallel passages extracted from the Bible as interlinear text. The *SHEBANQ* [SHE, 2016] toolkit for the study of the Hebrew Bible, and the *Chinese Text Project* [chi, 2016] contain integrated databases of hundreds of thousands of manually compiled parallel passages. Furthermore, a number of systems stand out as being related to the work introduced in this paper.

The *Tesserae project* [Coffee et al., 2012, tes, 2018] is a web application designed to provide digital tools for exploring inter-textual parallels between Latin literary works. The underlying model is word-based, but still provides a level of flexibility in the matching of similar lexical items. The tool produces a list of shared text patterns based on two or more words, and ignores any differences in syntax, and the position of the constituents of the passage, for instance,"committunt semina sulcis" versus "sulcis committas semina" are regarded as similar parallel passages. An approach developed in [Büchler, 2010], suggests an unsupervised method to identify and extract instances of textual reuse in the form of syntactic and semantic similarities identified between documents in ancient Greek. Their research on textual-reuse is also the focus of a project entitled *eTrap* [Franzini et al., 2015]. The project established language-independent approaches for identifying and quantifying text-reuse in historic documents. An output of the project is *TRACER*, which is a *Text Reuse Detection Software* developed to identify the differences and similarities between texts through analysis and manual compilation of short folklore-motifs for comparison of textual re-use cross-linguistically.

However, many of the approaches are reliant on language-dependent natural language processing tools to segment the text and reduce the words to a common representation through normalisation of case, lemmatisation, and the identification of synonyms. There also exists a handful of commercial software tools available for performing document comparison, such as *ABBYY Fine Reader* [ABB, 2018], *Diff Checker* [dif, 2018], and *DiffDoc* [dif, 2016], but the focus of these tools is on locating and highlighting the differences between documents, as opposed to their similarity.

Our approach differs in several important ways. Firstly, the approach is language-agnostic, which is achieved by representing the documents as a collection of variable length character *n*-grams stored in a Statistical Language Model [Zhai, 2008] over the collection as a whole, as well as each individual document. This means there is no

language-specific pre-processing of the texts required. Furthermore, the approach allows us to locate both exact and approximate string patterns using the the *n*-grams of the documents as starting seeds, as well as compensating for errors in the document text, introduced during the digitisation process. For instance, consider the difficulty in extracting words from the following text:

> "They are eafily ge- nerated ; but their extindion is a work of time and difficulty. Let us, therefore, (efpecially when we cc hold the mirror up to nature at home,) not only forgive, but even forget, if poffible, all the unpleafant treatment our citizens have experienced."[1]

Many of the words in the text are either corrupted by errors in the recognition of characters, for instance, 'f' versus 's', and 'd' replacing 't', or divided by white-space and punctuation due to the flow of the text in the original document. In this paper, we present a digital infrastructure that greatly facilitates the finding and comparing parallel passages in any domain or language. Here the notion of "finding" is divided into *recommendation* (i.e., locating documents discussing the same subject or topic in relation to a source document), and the other refers to *discovery* (i.e., identifying potentially related text segments shared between documents). In the next section, we introduce the digital archives and text corpora, which act as the case studies for generating a list of recommended documents, and extracting parallel passages.

3 The Digital Archives

The document recommendation and comparison tools have currently been deployed over the archives listed in Table 2. The *Aramaic Magic Bowls and Amulets from Late Antiquity* (6th to 8th CE) are the focus of research for a team of historians from Israel and the UK led by the University of Southampton [VMB, 2014],[Levene, 2002]. The texts were written in ink on clay bowls using a number of related dialects including Aramaic, Mandaic, and Syriac. The focus of research is on finding formulaic parallel passages that provide an insight into the evolution and transmission of liturgical forms over the centuries, which is valuable for understanding both Jewish society, and the history of magic in late antiquity. For instance researchers have identified passages that appear a few centuries later in the *Book of Ezra* from the Hebrew Bible.

The document comparison was initially developed to support the textual analysis of the Aramaic Magic Bowl corpus, where the research focus is on identifying the historical and cultural context of the texts by comparing texts found in the same location, culture, and time period. Since many of the Magic Bowls were discovered with no archaeological context the work looks to establish their meaning, significance, authorship, and time period through analysis of the Aramaic, Hebrew, and Syriac, and Mandaic languages and dialects [Ginsberg, 1936] recorded in the texts. The document comparison tool assists in identifying verbose and formulaic phrases that have been paraphrased or inserted as a whole or in part across multiple texts over several periods.

[1] Quoted from *A Short Account of the Malignant Fever: Lately Prevalent in Philadelphia... To which are Added, Accounts of the Plague in London and Marseilles* [Carey, 1794], *Wellcome Trust UK Medical Heritage Library*

Archive	Domain	Period(s)	Language	Documents	Size
Aramaic Magic Text from Late Antiquity (Southampton University)	Religion	6AD - 10AD	Aramaic, Hebrew, Syriac	539	830.5KB
English Bibles (Archive.org)	Religion	15th - 20th Century	English	185	947.2KB
UK Medical Heritage Library (Wellcome Trust)	Medicine	18th - 20th Century	English, French, German, Italian Spanish, Russian, Malagasy	75,973	35.1GB
Financial Times Historic Archive (British Library)	Finance	18th - 20th Century	English	70,640	230MB

Table 2: **A summary of the digital archives supported by the document comparison tool.**

The *UK Medical Heritage Library* is curated by the Wellcome Trust and contains documents in a broad range of topics, including medical articles, health reports, books on diet and nutrition, and historical documents relating to medical practices (e.g. phrenology), although we found that more than 54,000 of the documents are composed of health reports on public health by year. The archive is also multilingual with documents in English, French, German, Spanish, Italian, and Russian [wel, 2017a, wel, 2017b].

The *Financial Times historic newspaper archive* (curated by the British Library) covers news articles published in the years 1888, 1939, 1966, and 1991) in English, and formed the basis of a pilot study organised by the British Library and the Financial Times to explore ways in which the corpus could be used to improve access and gain insights into the content of the archive [2].

Lastly, we have also applied the system to a collection of Bibles in the English language, including the the *Tyndale Bible* (1526), *DouayRheims Bible* (1582), *King James Bible* (1611), *Noah Websters Bible* (1833), and *The American Standard Bible* (1901), which provide a record of the evolution of the English language over a long time period of several centuries.

4 Statistical Language Models

The digital infrastructure developed to address the recommendation of documents, and extraction of "parallel passages", is constructed from a character-based *Statistical Language Model* (*SLM*). An *SLM* is a mathematical model representing the probabilistic distribution of words, or sequences of characters, found in the natural language represented by text corpora [Zhai, 2009], which provides a consistent methodology for comparing documents according to the underlying principles and structure of the language.

Recently *Neural Network Language Models* (*NNLM*) have been proposed to address some of the short-comings of *SLM*s [Bengio et al., 2003], particularly the data-sparseness issue and to compensate for the occurrence of unseen words in the data. An *NNLM* addresses the data-sparsity issue by encoding words as vectors, or word-embeddings, as the input to the neural network [Mikolov et al., 2013]. The current state-of-the-art approaches are able to identify synonyms related to individual words, making them suitable for a wide range of Natural Language tasks. *NNLM*s are parametric models requiring some experimentation to achieve good performance, which means they require a certain degree of human engineering to design the network architecture, training approach, provide an appropriate set of training examples, and encode the input to the network in a suitable way to accommodate the data, or application. In addition, the computational costs required to train an *NNLM* is much greater than for an equivalent language model.

A further issue is that languages such as Hebrew, and Russian attach affixes to a root morpheme, which represent the syntactic constituents of the language. This compli-

[2] Samtla received the 2017 runners-up award for research for the work achieved on this archive: `https://blogs.bl.uk/digital-scholarship/2018/02/-bl-labs-2017-symposium-samtla-research-award-runner-up.html`

cates word-based models as the documents must first be pre-processed using language-dependent stemming and part-of-speech tagging algorithms to identify all instances of a word to create an accurate model. The character-based n-gram model resolves many of these issues, and they have been shown to outperform raw word-based models in practice, when the language is morphologically complex [Mcnamee and Mayfield, 2004], which has made them effective in spam-email filtering [Kanaris et al., 2006], authorship attribution [Houvardas and Stamatatos, 2006], neural networks [Kim et al., 2015], and named entity recognition (NER) [Klein et al., 2003], compared to the word-level models, due to the fact that the character-level model captures the syntax, and to some degree, the semantics of text by modelling sub-word features that are not available at the word-level. Recent work demonstrates that the number of parameters required of *NNLM* models can be optimised through the adoption of a character-level representation to model higher-level linguistic units, such as words at the output layer [Kim et al., 2015]. While character *NNLM* models have shown to outperform word-level models, they require morphological tagging as part of the preprocessing step, which may not be available for little known languages, and particularly ancient languages.

We adopted an *SLM* approach, since they are non-parametric, and do not require much in the way of heuristic design due to the underlying probabilistic framework adopted by the model, which makes them simpler to implement and they have been shown to perform well empirically [Zhai, 2008]. Furthermore, the way in which an *SLM* computes the probability of a given sequence is transparent nd easily explainable to researchers with a non-technical background. Conversely, the *NNLM* model is generally treated as a black-box whereby there is no direct way to show how the resulting probabilities were generated by the model given the original input. We adopt a character-based n-gram *SLM*, rather than the more conventional word-based model since it requires very little in the way of language-specific stemming or text normalisation, aside from ignoring strings representing punctuation marks.

A *SLM* is computed over the whole archive of documents, called the *collection model* C, and over each document individually, which we refer to as the *document model* D. The collection model C provides a archive-specific, or global, probability for every sequence of characters according to the language, whereas the document model D provides a local probability, which enables us to model, to some degree, the topic of the document. The *SLMs* are generated by extracting overlapping character-level n-grams from the document content, where n varies from a single character to a pre-determined maximum, using a sliding window of fixed length n. We selected n=15, based on the observation that many languages tend to have an average word length of approximately 15 characters [wol, 2016]. This allows us to capture character-sequences representing both single words and short phrases. The character n-grams are further reduced to lower-order n-grams by iteratively removing a character from the front of the sequence for each order n. The motivation for this is that higher-order n-grams tend to represent collocations or phrases, and can occur less often than smaller n-grams represented by words, meaning that there would be less information on which to calculate a reliable probability when constructing the *SLM*. By approximating the probability of the higher-order n-grams through interpolation [Chen and Goodman, 1999, Zhai and Lafferty, 2004], using the more reliable es-

timates of the lower-order n-grams, we are able to preserve the dependency between large sequences and small sub-sequences, that is, the dependency or relationship between higher-order structures (such as clauses, and collocations), which encode a level of semantics, and lower order n-grams capturing the syntax of the language. We store the n-grams of the documents in a k-truncated [Schulz et al., 2008] on-disk suffix tree [Barsky et al., 2010] data structure with the depth of the tree limited to a maximum of $k = 15$ nodes, equal in size to the sliding window to reduce the memory requirements of the suffix tree.

Scale is achieved through horizontal-scaling by partitioning the input according to the finite-alphabet of the language and storing each sub-tree separately on disk during construction. The resulting generalised suffix tree represents a compressed "trie" data structure where the suffixes of the string act as the keys and the positions of the string are the values. The leaf nodes store the individual document ID and start positions of the character string, which are retrieved as the index for the documents when the tree is traversed to extract n-grams of a given length during comparison (see Section 6).

We store the count of each character of the n-gram at the node for calculating the conditional probability of any n-gram. Once constructed the probability of a given n-gram submitted to the suffix tree as a query, which we denote as q, can be computed through Bayes theorem [Gill, 2014], where we have

$$P(D|q) = \frac{P(q|D)P(D)}{P(q)} \ . \tag{1}$$

The right-hand side of Equation (1) represents the standard query model $P(q|D) = P_D(q)$ multiplied by $P(D)$, the *prior* probability of the document D which is often assumed to be uniform (i.e. the same for all documents) and we therefore ignore for the purpose of scoring. As the prior is uniform, we score the n-grams of the documents on the basis of the query model $P_D(q)$ for each document retrieved from the suffix tree. Smoothing is an important component of a *SLM*, as it reduces the influence of terms representing the syntax of the language, which are not good descriptors of the topic defined by the users query [Zhai, 2008], and at the same time, compensates for terms that are missing in the documents, which can occur when extracting n-grams from two separate documents for comparison.

We adopt Jelinek-Mercer smoothing [Bahl et al., 1983] to adjust the conditional probability of the n-grams in the documents given information about their frequency according to the collection model C. Defined as follows,

$$P_D(q) \approx \lambda_1 \hat{P}_D(q) + \lambda_2 \hat{P}_C(q), \tag{2}$$

where we replace P_M with $\hat{P_M}$ to show that we are approximating the probability inferred from the given *SLM*, and λ_i is a weighted term that defines the contribution of each *SLM* with the document model D contributing $\lambda_1 = 0.6$, and the collection model C contributing $\lambda_2 = 0.4$ to the final smoothed score for the query. A more detailed description of the *SLM* and suffix tree are presented in [Harris et al, 2014]. The smoothed probability for a n-gram is used to generate a probability distribution for each document, which we use to find similar documents through the document recommendation tool discussed next, in Section 5.

5 Document Recommendation

Document recommendation is the process of recommending documents to the user that discuss the same or similar topic in relation to a *target* document. The target document is the document the user selects, either from a list of search results or through browsing. The task of document recommendation is to identify documents that are most similar to the target document. This requires that the documents are reduced to a common representation that can be measured in order to assess the degree of similarity between document pairs. Documents can be represented by feature vectors describing information, such as URL, date of publication, language, topic, and author, or the content can be extracted in the form of n-grams. We identify related documents based on the statistics drawn from the character-level n-gram probability distributions stored in the document model D. After identifying a set of related documents, we apply local-sequence alignment methods to visualise the text sequences shared between the documents (see Section 6).

We identify related documents by measuring the similarity between the character-level n-gram probability distributions of the documents stored in the document model D. The size of the n-gram is fixed, where small values for n correspond to a finer-grained document similarity measure and a high setting for n is best for identifying documents that share long verbose sequences, representing a coarser-grained analysis. We have found $n=7$ corresponding to a 7-gram model provides a good balance between small and large shared-sequences, based on our 15-gram language model. That is, half the length of the sliding window used for processing the n-grams during indexing. This was supported by anecdotal evidence supplied by researchers who empirically assessed the output of the document recommendation and comparison tools.

The similarity between the probability distributions is computed with the *Jensen-Shannon Divergence* (*JSD*) measure, which is the symmetric version of the well-known *Kullback-Liebler Divergence* (*KLD*) [Endres and Schindelin, 2003]. Each document model M_d is extracted from the document model D, and the *KLD* is computed between two n-gram probability distributions, P and Q, corresponding to $document_1$ and $document_2$, respectively. The *KLD* is defined as follows

$$D_{KL}(P||Q)\sum_i P(i)\log_2\frac{P(i)}{Q(i)}, \tag{3}$$

where i is the probability for a 7-gram drawn from the smoothed distribution for $document_1$, or $P(i)$, and $document_2$, defined as $Q(i)$. The *JSD* is then derived from the resulting *KLD* score, as follows:

$$JSD(P||Q) = 1 - \sqrt{\frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)}, \tag{4}$$

where M is the average of the two distributions P and Q, which is defined as $\frac{1}{2}(P+Q)$. The *JSD* is then one minus the square root of the interpolation between the two distributions which are interpolated with a weighted term corresponding to a 50% contribution from each *KLD* score for P and Q. The resulting *JSD* returns a value between 0 and 1, where a score of 1 means the documents are identical. For each document, the *JSD* scores are ordered in descending-order according to their similarity to

the target document so that the most similar documents are ranked at the top of the related document result list. These ranked lists are displayed alongside the corresponding document when users view a document, either through search or browsing.

6 Document Comparison

Document comparison is the task of comparing the difference, or similarity, between the content of two or more documents through analysis of shared vocabulary or features. The document comparison tool developed for *Samtla* identifies text regions of similarity between documents that could be widely divergent overall. Divergence between parallel passages defines the permitted tolerance between the two sequences before they are no longer classed as being similar, or identical.

The underlying algorithm for identifying shared text patterns is a tailored variant of the *Basic Local Alignment Search Tool* (*BLAST*) algorithm, commonly used in bioinformatics for comparing *DNA* sequences [Ma and Zhang, 2011]. The method uses a *local sequence alignment* approach that identifies a series of short sequences called seeds extracted from the document model D, that are common to the documents being compared. The start seeds are expanded a character at a time to produce a larger sequences in each document, up to a predefined threshold. More precisely, the seeds are composed of the unique set of 3-gram character strings shared between two documents, one representing the "target" document, and the other a document drawn from the list of related documents (refer to Section 5).

We selected the 3-gram as the best starting length for the seeds as it provides a finer-grained comparison and flexibility in the extension process compared to higher-order n-grams. We extend the 3-gram seeds one character at a time, first from the left, and then from the right, through an iterative extension process. This extension process captures the left and right contexts for the seed. The motivation behind the document comparison tool is that sequences sharing the same or similar left and right contexts may be classed as syntactically or semantically related. We score each pair of (approximately) matched sequences according to their *Levenshtein edit distance* [Levenshtein, 1966, Gusfield, 1997]. Given an expanded seed s_1 from $document_1$, and s_2 extracted from $document_2$, we score each sequence as follows:

$$ed(s_1, s_2) \leq \lfloor m\delta \rfloor \tag{5}$$

where the term $\lfloor m\delta \rfloor$, on the right-hand side, defines the threshold determining the limit of the extension process. The limit is met when the edit distance is greater than the floor of the length of the shorter sequence m, multiplied by a tuneable tolerance parameter δ. The default setting for the tolerance is $\delta = .2$, which allows the two sequences to differ by as much as 20%, before the extension stops and moves on to the next seed. The extension is also terminated once the seed start position reaches zero and end position is extended to the length of the document. In this case, the documents are considered practically identical. The output produced by the algorithm is a list of start and end positions identified by a unique seed *id* for each instance of the parallel passage for the two documents. The algorithm for extending and subsequently scoring the initial seeds, is presented in Algorithm 1. As an example of the output generated

Algorithm 1 The algorithm for extending and measuring the edit distance between the seeds, represented as 3-gram character sequences, shared between document pairs. The index of start and end positions are passed as an argument to this function for each instance of the seed, which is subsequently extended up to the maximum tolerance limit.

procedure SEED–EXTENSION
  Retrieve each shared 3-gram $s_1 \in D_1$ and $s_2 \in D_2$
  Retrieve the start and end position (i,j) for all s_1 and s_2 as the initial start and end position of the seed.
  for each (start$_i$, end$_i$) in seed s_1 **do**
    for each ($start_j$, end_j) in seed s_2 **do**
      $m = |s_1|$
      n = 0
      ed = editdistance(s_1, s_2)
      while $ed \leq \lfloor m\delta \rfloor$ **do**
        s_1 = s_1 ($start_i$-n, end$_i$)
        s_2 = s_2 ($start_j$-n, end$_j$)
        m = min($|s_1|$,$|s_2|$)
        ed = editdistance(s_1, s_2)
        $s_1 = s_1(start_i, end_i + n)$
        $s_2 = s_2(start_j, end_j + n)$
        m = min($|s_1|$,$|s_2|$)
        ed = editdistance(s_1, s_2)
        n = n + 1
      end while
      Store the resulting sequence for s_1 and s_2
    end for
  end for
end procedure

by the algorithm, if we were to extract the initial 3-gram seed string "ham" from the *Book of Genesis, Chapter 10* appearing in both the *Douay-Rheims Bible* (1582), and the *King James Bible* (1611), and iteratively extend it up to the threshold score, we obtain the following shared-sequence:

Douay-Rheims Bible (1582) – Genesis, Chapter 10:
"Noe: Sem, C ham , and Japheth"

King James Bible (1611) – Genesis, Chapter 10:
"Noah; Shem, Ham , and Japheth".

The edit distance between these two sequences is calculated as follows:

- The strings *Noe* and *Noah* have an edit distance of two, since one substitution ($a \rightarrow e$) and one insertion (final character h) is required to translate the strings.
- the conversion of the string *Sem* to *Shem*, requires an insertion of character h, equal to an edit distance of one.
- the sequence *Cham* is converted to *Ham* with the deletion of character C at the beginning of the string for a total edit distance of four.

Despite the differences in the spelling of the names appearing in the two texts the two sequences might be regarded as semantically equivalent to a researcher of Bible scripture.

7 User interface

The interface for the document comparison tool was developed as part of a collaborative design process with the users of the *Aramaic Magic Bowl archive* (see Figure 1, and Figure 2). The central idea behind the current design was to replicate the process of comparing the physical documents, where the researcher places the two documents side-by-side and highlights the parallel passages of interest. The interface was designed to maintain the structure of the digital texts so that occurrences of parallel passages could be identified.

To access the document comparison the user selects a document from a sidebar composed of a list of related document considered to be similar to the target document. The main area of the user interface is designed to emulate the researchers' process of comparison by rendering the content of the documents next to each other. Above each document is a small horizontal map, which displays a preview of the parallel passages present in the documents in light-blue, and the currently selected parallel passage in a darker shade of blue. We use a similar scheme in the document content windows to highlight the passages that are selectable.

In addition, we provide a filter at the bottom of the documents to allow the user to filter the documents for large and short sequences. The default behaviour of the tool is to display the largest parallel-passage extracted. When the user selects a parallel passage in the body of the text, the related document window automatically scrolls to the first instance of the passage.

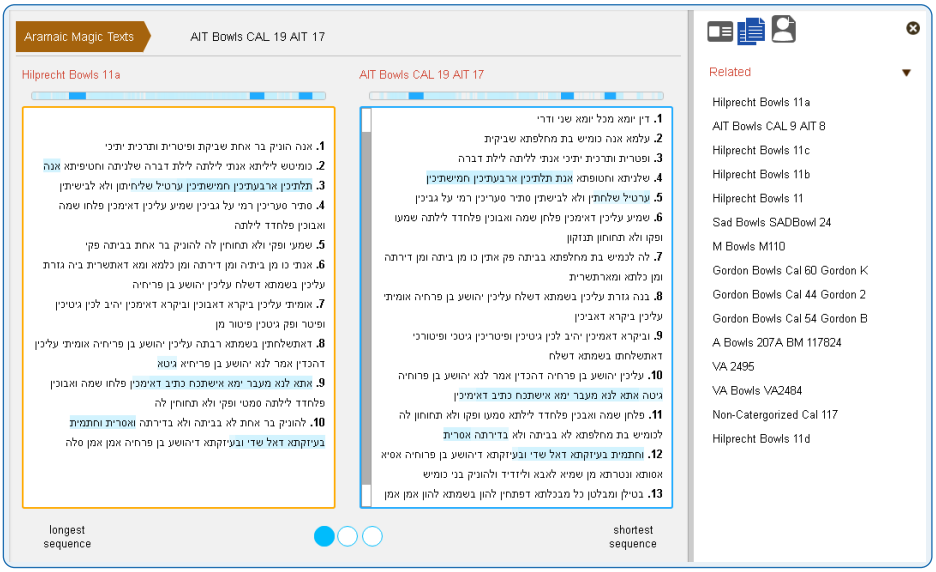

Figure 1: A document comparison between multiple parallel passages extracted from the *Aramaic Magic Bowl archive*.

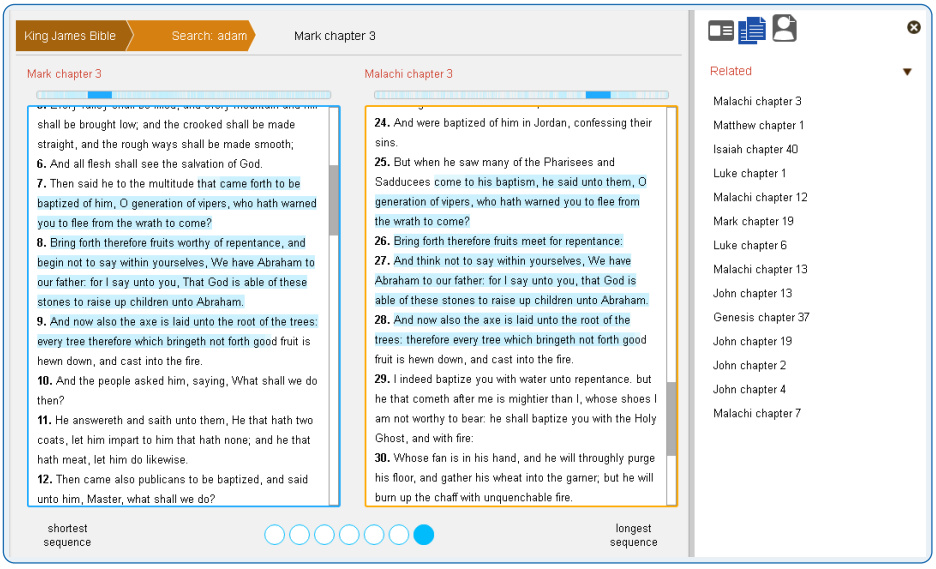

Figure 2: A document comparison between two parallel passages extracted from the *King James Bible*.

8 Discussion

The core motivation of the tool is to find sequences of text that are syntactically and semantically similar to other texts, whilst ignoring the small nuances created by digitisation and spelling errors, vocabulary choice, paraphrasing, and surface damage. The tool has been applied to a range of corpora, which differ with respect to language, domain, and time period, where some collections contain a record of the language across at least two centuries. Each collection has an associated *SLM* constructed over the documents, and many of the corpora fall into the historic text domain, which was driven by the needs of researchers working in collaboration to design and test the tool to support their literary analysis or research on social and religious cultural contexts recorded by the texts. The example result are composed of pairs of texts considered the most similar to each other, and we highlight the parts of the shared-sequences that differ to illustrate the degree of tolerance permitted by the approach.

We discuss the issues that the document comparison tool has been designed to resolve, or mitigate, and present several example sequences extracted by the tool mainly from the English Bible corpus due to their accessibility to the reader, the fact that they provide some of the best examples of textual reuse, are the work of several authors, and they record the evolution of a language over a long period of time capturing changes to the orthography and spelling conventions.

8.1 Digitisation and text representation issues

The document comparison tool compensates for issues associated with the representation of texts generated through digitisation. Issues range from damage to the physical surface of the text for instance in the case of paper, parchment, and velum media there may be some deterioration due to the conditions of storage, or the age of the original. In the case of the Aramaic Magic Bowls, which were made of ceramic, the text may be incomplete due to missing fragments, and there are instances where the ink has faded due to exposure to light, and moisture. In addition, the process of digitisation often removes the structure of the text causing image captions, footnotes, headings, and page numbers to be rendered within the main text, causing a corruption of the text.

Furthermore, the precision of the OCR output may be poor due to the typeface of the printed text, which can result in poor recognition causing some characters to be replaced by others e.g. $v \rightarrow u$, $d \rightarrow a$, $j \rightarrow I$, and $n \rightarrow i$. The example from the *Financial Times archive of historic newspapers*, presented in Table 3, shows two parallel passages extracted from an article entitled "Copper Joins in General Metal Malaise", which is considered to be the top-related document to the article "Vietnam a Factor in Both Tin and Copper". In this example, the quality of the OCR affects the ability of text comparison tools to extract all instances of the same word or phrase, for instance the word "America" has several permutations, such as "Aeuican", and "Americai", present in earlier issues of the Financial Times (the years 1888, 1939, and 1966). These spelling errors are attributed to insertions, deletions, and substitutions generated by the digitisation process:

Vietnam a Factor in Both Tin and Copper, Feb 01, 1966, issue 23,840.
... of 5 tens each. cotton Liverpool-American contract closed quiet, unchanged. American i-inch middling spot 23.35d., Sudan bar Sakel five 39d, Lambert six 32.50d. Closing rices (basis American i-inch middling): March 22.20d. May 22.20d., July 22.20d, Oct. 20.90d. Dec. 20.90d, March 20.90d, May 20.90d. Sudan contract closed quiet and unchanged. Freights dry cargo-grain chartering was continued at a ...
Copper Joins in General Metal Malaise, Feb 10, 1966, issue 23,848.
... ch..... 2251 12g5.6l cotton Liverpool-Aeuican contract closed squiet, unchanged. American 1-inch i middling 23.35d Sudan Sakel five 39d, Lambert six 32.50d. Closing prices (basis h Americai-inch middling): March 22.20d, May 22.20d. July 22.20d. Oct. 20.90d,. Dec. 20.90d, March 20.90d, May 20.90d. i Sudan contract closed quiet, unchanged. sFreights dry cargo- the overall weight of chartering rem...

Table 3: An example parallel passage identified in two separate issues in a column published in the Financial Times newspapers archive 10 days apart.

When performing a comparison across this particular collection of digital texts the task is made more difficult by the differences attributed to OCR errors, which can be an issue for automatic text analysis approaches based on words. After the digitisation process, there may still be issues to overcome when identifying textual re-use. Language is dynamic, and as a result when analysing a body of texts covering a long period of time we can expect to find variability in the language and its orthography as it evolves over time, which compound the problem of identifying all instances of the same word or phrase.

8.2 Issues attributed to vocabulary choice and paraphrasing

Some texts such as the Bible, and Aramaic Magic Bowl corpus contain a high-level of textual re-use. This is attributed to the religious genre of the text where important figures and key events are reiterated, or used to frame the context of a new text. For instance, the Aramaic Magic Texts from Late Antiquity are similar with respect to

Egyptian offering formulae [Franke, 2003], where the same or similar introductory text and conclusion are inserted as a matter of necessity to imbue the text with authority or spiritual power. In the context of the Bible we can find examples of different vocabulary choices made by the author, as well as the influences of the original source text from which the new text was translated.

A good example of this is presented in Table 4, were the texts span several centuries, and illustrate the presence of a core text e.g. "*and the King of Sodom went out to meet him, after ..., and of the Kings that were with him, ..., which is the King's ...*". With differences between the two exhibited by the spelling of named entities such as the names of people and locations mentioned in the text. These variations may have been influenced by the original source used for translation. Translators depending on different editions of the Bible. The Bible has been translated several times from sources in Hebrew, Greek, and Latin, which have resulted in different transcriptions for names of people and locations, for example in Table 4, nouns denoting the names for people are transcribed with different vowels, e.g. *Chodorlahomor* versus *Chedorlaomer*, and noun phrases such as titles *Melchisedech the King of Salem* when compared to *Melchizedek King of Salem*. A more extensive example can be found in the names for location, where we observe differences in vocabulary choice and spelling *in the Vale of Save* versus *at the Valley of Shaveh*. Parts of the text are identical between the two editions, whereas other parts have been adapted or paraphrased and potentially driven by decisions made by the author during the construction and subsequent editing of the text thereafter. When we compare shared-sequences forming larger linguistic structures such as sentences, we can see the further effects of authorship and textual reuse. For example, consider the challenge of comparing the first sentence in the *Douay-Rheims Bible* from 1752: "*the substance, and lot his brother, with his substance, the women also the people*", with that of the *King James Bible* 2003 edition: "*and he brought back all the goods, and also brought again his brother lot, and his goods, and the women also, and the people*". The sentences describe a series of events in sequential order, and are semantically similar. However, the text of the *King James Bible* (2003) is more verbose and there exists additional textual material where *the substance* is replaced by the phrase *and he brought back all the goods*.

Another example of paraphrasing can be seen in Table 5, where there is not only a difference in how the names for entities are transcribed, but also the constituents of the phrase have been changed e.g the phrase Agar *her Egyptian maid* is rendered as "Hagar *her maid the Egyptian*", and "*Abram her husband* to be his wife" compared to *her husband Abram*. The instances would be regarded as semantically similar in terms of their interpretation, however, these examples demonstrate how difficult the process of document comparison can be when tackled manually. In the case of software tools, rules are often defined to capture the numerous ways in which a text is constructed, edited, and evolving as a consequence of natural language processes overtime.

Douay-Rheims Bible, the Challoner Revision (1752)
... the substance, and lot his brother, with his substance, the women also the people. and the King of Sodom went out to meet him, after he returned from the slaughter of Chodorlahomor, and of the Kings that were with him in the Vale of Save, which is the King's Vale. but Melchisedech the King of Salem, bringing ...
King James Bible (2003)
... and he brought back all the goods, and also brought again his brother lot, and his goods, and the women also, and the people. and the King of Sodom went out to meet him after his return from the slaughter of Chedorlaomer, and of the Kings that were with him, at the Valley of Shaveh, which is the King's Dale. and Melchizedek King of Salem brought forth ...

Table 4: The longest sequence shared between two chapters of the Bible from different periods, with the differences between the two sequences highlighted.

Thomson's Septuagint by SF Pells (1808)
took Agar her Egyptian maid, after Abram had dwelt ten years in the land of Chanaan, and gave her to Abram her husband to be his wife. And he went in unto agar
King James Bible (2003)
took Hagar her maid the Egyptian, after Abram had dwelt ten years in the land of Canaan, and gave her to her husband Abram to be his wife. And he went in unto Hagar...

Table 5: The longest sequence shared between two chapters of the Bible from different periods, with the differences between the two sequences highlighted.

8.3 Language change over time

The document comparison tool is also tolerant to language change in texts that span several periods. The Bible provides another good example of this as illustrated in Table 6, where the variability between the extracted sequences is quite extensive where the language and orthography of English has since evolved between the two editions. A researcher might regard these two texts as semantically equivalent despite the difference in the surface form of many of the words, which include character substitutions, insertions, and deletion. To illustrate, compare the surface forms of the word *buttelarshipe* in the *William Tyndale* Bible and its equivalent *butlership* in *The American Standard Version of the Holy Bible*. A further example is the substitution $u \rightarrow v$, which is motivated by changes in the orthographic conventions of the English language. Around the time of the *William Tyndale Bible*, it was common to use the character *v* as an allograph for the sound *u* in certain contexts, such as the beginning of words e.g. *unto* versus *vnto*. The same can be said of the substitution *i* for *y* in words *lyfted* versus *lifted*, and *thyrde daye* versus *third day*, not to mention the insertion of *e* at the end of the two words. These small differences are not easily captured with word-based approaches, and would result in out of vocabulary items [Woodland et al., 2000]. This degree of variability is also difficult to capture using generalised rules, since some rules may be specific to individual words rather than a general rule or pattern of the orthography and language.

William Tyndale Bible (1534)
... daye .iij. dayes shall Pharao take thy heade from the and shall hange the on a tree and the byrdes shall eate thy flesh from of the. And it came to passe the thyrde daye which was Pharaos byrth daye that he made a feast vnto all his servauntes. and he lyfted vpp the head of the chefe buttelar and of the chefe baker amonge his servauntes. And restored the chefe buttelar vnto his buttelarshipe agayne and he reched the cupp...
The American Standard Version of the Holy Bible(1901)
... ee days shall Pharaoh lift up thy head from off thee, and shall hang thee on a tree; and the birds shall eat thy flesh from off thee. And it came to pass the third day, which was Pharaohs birthday, that he made a feast unto all his servants: and he lifted up the head of the chief butler and the head of the chief baker among his servants. And he restored the chief butler unto his butlership again; and he gave the cup int...

Table 6: The longest sequence shared between two chapters of the Bible from different periods.

In addition, when comparing historic texts with their modern equivalents, there are often a number of overlapping processes occurring simultaneously involving both language change over time and decisions influencing how the author constructed the new text.

The document comparison tool was designed to facilitate the discovery of parallel passages representing textual reuse exhibited by documents identified as similar by their *JSD* score, and recommended to the user. The character-based approach provides a simple language and corpus agnostic method and is tolerant to OCR errors, variability introduced by authorship through style and vocabulary choices, and changes to the orthography and syntax of the language over time. As illustrated in the discussion, the tailored variant of the *BLAST* algorithm is capable of identifying parallel-passages that can be quite divergent overall, providing a starting point for researchers to identify textual reuse in a potentially large collection of documents that would otherwise involve a large amount human effort to compile manually, even for a relatively small collection of documents.

In terms of the limitations of the approach, the extracted text sequences may not be aligned to words, meaning that the beginning and end of the shared-sequences may begin or end word internally. We experimented with extending the seeds to capture

whole words, but the approach was ad-hoc and caused some sequences to appear more significant than they were, especially for morphologically rich languages. In addition, the size of the starting seed determines the level of granularity. The starting seeds are 3-gram sequences representing the exact matches shared between two documents. The initial seeds should not be too large otherwise some sequences may not be identified during the iterative extension process as the tolerance threshold would have been met. Greater flexibility can be achieved through a higher setting for the tolerance parameter, which can help to address different levels of morphological complexity present in different languages. In other words, languages like Aramaic require a higher setting $\delta \leq .2$, whereas languages with low morphological complexity, such as English, require a smaller setting, in our study we found $\delta \leq .1$ to be appropriate based on anecdotal evidence from our users. In general however, a tolerance of $\delta = .2$, performs well for the majority of languages, particularly if the corpus contains a record of the language over a period of time or across several dialects. This value can be discovered through experimentation by comparing the results with known shared-sequences in published research or by consulting domain experts.

9 Conclusion

The paper presented an outline of a generic document comparison tool currently developed for the *Samtla* system with application to mining text patterns. The main research contribution is a novel combination of *character-based n-gram language models*, *space-optimised suffix tree*, *generalised edit distance* and *local sequence alignment*, which is relatively simple to implement and agnostic to the language, script, and domain of the text. The tools provide the means for researchers to interact with and explore the documents through and comparison of parallel passages representing repeated text patterns of interest to researchers in the humanities who are looking to make use of the large volume of textual data, such as letters, witness accounts, reports, and monographs stored in cultural heritage archives.

The related documents tool was constructed from the *SLM n*-gram probability distributions for pairs of documents and we measured their similarity through the well-known *Jensen-Shannon Divergence JSD* measure, which is a popular method for assessing the similarity between two probability distributions. The recommended document tool provides the access point for a further tool, the document comparison tool (discussed in Section 6), which enables researchers to explore the "relatedness" of the recommended documents through visual mining of large and small parallel passages. The techniques complement each other and work well together to provide a domain and language agnostic digital infrastructure for the search and discovery of parallel passages of importance to historians and linguists researching the reuse of cultural contexts recorded in the documents.

The document comparison tool is designed to provide an automated approach for extracting parallel-passages to provide a starting point for researchers to explore a collection of documents. Using a combination of simple approaches we have demonstrated that the document comparison tool can compensate for language and author-specific choices, which can cause issues for researchers performing a large scale com-

parative analysis across multiple documents. The tool will help researchers of text corpora reduce the time investment incurred in manually annotating and comparing these sequences for research with the potential for identifying novel sequences that have not previously been recorded. The proposed digital infrastructure for finding parallel passages is not necessarily restricted to historic document collections, but can be straightforwardly extended to other application domains such as medical and legal document collections.

As part of future work we will explore methods for evaluating the results of the document recommendation and comparison tools. We have evaluated the underlying *SLM* model used to score and retrieve the *n*-gram sequences [**?**], and the next step is to evaluate the sequences generated by the output of the document comparison. There would not appear to be an established standard of measures nor a gold-standard data-set on which to evaluate document comparison methods, and so we will be exploring previous approaches based on crowd-sourcing and non-parametric correlation measures to assess user rankings of the generated recommendations and comparisons.

References

[VMB, 2014] (2014). Vmba: Virtual magic bowl archive. http://www.southampton.ac.uk/vmba/. [Online; accessed 28-January-2014].

[bib, 2016] (2016). Bibleworks - bible software. http://www.bibleworks.com/classroom/1_10/. [Online; accessed 18-May-2016].

[chi, 2016] (2016). Chinese text project – parallel passages. http://ctext.org/tools/parallel-passages. [Online; accessed 18-May-2016].

[dif, 2016] (2016). Diff doc tool. http://www.softinterface.com/MD/Document-Comparison-Software.htm. [Online; accessed 8-February-2016].

[log, 2016] (2016). Logos bible software series x tour: Parallel passages. https://www.logos.com/media/tour/ParallelPassages.htm. [Online; accessed 18-May-2016].

[SHE, 2016] (2016). Shebanq (system for hebrew text: Annotations for queries and markup). https://shebanq.ancient-data.org/. [Online; accessed 18-May-2016].

[wol, 2016] (2016). Word length distribution in various languages. https://reference.wolfram.com/language/example/WordLengthDistributioninVariousLanguages.html. [Online; accessed 11-January-2016].

[wel, 2017a] (2017a). Wellcome trust collections – uk medical heritage library. http://wellcomelibrary.org/collections/digital-collections/uk-medical-heritage-library/. [Online; accessed 7-January-2017].

[wel, 2017b] (2017b). Wellcome trust uk medical library project. `wellcomegrant`. [Online; accessed 19-June-2017].

[ABB, 2018] (2018). Abbyy finereader 14. `https://www.abbyy.com/en-gb/finereader/compare-documents/`. [Online; accessed 16-March-2018].

[dif, 2018] (2018). Diffchecker. `https://www.diffchecker.com/`. [Online; accessed 16-March-2018].

[tes, 2018] (2018). Tesserae. `http://tesserae.caset.buffalo.edu/index.php`. [Online; accessed 16-March-2018].

[Aggarwal and Zhai, 2012] Aggarwal, C. and Zhai, C. (2012). *Mining Text Data*. Springer-Verlag New York Inc.

[Bahl et al., 1983] Bahl, L. R., Jelinek, F., and Mercer, R. (1983). A maximum likelihood approach to continuous speech recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-5(2):179–190.

[Barsky et al., 2010] Barsky, M., Stege, U., and Thomo, A. (2010). A survey of practical algorithms for suffix tree construction in external memory. *Softw. Pract. Exper.*, 40(11):965–988.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155.

[Büchler, 2010] Büchler, M, G. A. E. T. H. G. (2010). Unsupervised detection and visualisation of textual reuse on ancient greek texts. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science (JDHCS)*, pages 14–16.

[Carey, 1794] Carey, M. (1794). *A Short Account of the Malignant Fever: Lately Prevalent in Philadelphia... To which are Added, Accounts of the Plague in London and Marseilles...* author.

[Chen and Goodman, 1999] Chen, S. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–394.

[Coffee et al., 2012] Coffee, N., Koenig, J.-P., Shakti, P., Ossewaarde, R., Forstall, C., and Jacobson, S. (2012). Intertextuality in the digital age. 142.

[de Jong, 2007] de Jong, M. (2007). *Isaiah Among the Ancient Near Eastern Prophets: A Comparative Study of the Earliest Stages of the Isaiah Tradition and the Neo-Assyrian Prophecies*. Supplements to Vetus Testamentum. Brill.

[Endres and Schindelin, 2003] Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.

[Fairclough, 1992] Fairclough, N. (1992). Discourse and text: Linguistic and intertextual analysis within discourse analysis. *Discourse & Society*, 3(2):193–217.

[Franke, 2003] Franke, D. (2003). The middle kingdom offering formulas: A challenge. *The Journal of Egyptian Archaeology*, 89:39–57.

[Franzini et al., 2015] Franzini, G., Franzini, E., Bchler, M., and Moritz, M. (2015). etrap [electronic text reuse acquisition project]: A research group implementing the ehumanities a.c.i.d. paradigm.

[Gill, 2014] Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach*, volume 20. CRC press.

[Ginsberg, 1936] Ginsberg, H. L. (1936). Aramaic dialect problems. ii. *The American Journal of Semitic Languages and Literatures*, 52(2):95–103.

[Gusfield, 1997] Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

[Harris et al, 2014] Harris et al (2014). The Anatomy of a Search and Mining System for Digital Humanities. In *Proceedings of the 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 165–168, London, UK.

[Hoek, 1996] Hoek, A. V. D. (1996). Techniques of quotation in clement of alexandria. a view of ancient literary working methods. *Vigiliae Christianae*, 50(3):223–243.

[Houvardas and Stamatatos, 2006] Houvardas, J. and Stamatatos, E. (2006). *N-Gram Feature Selection for Authorship Identification*, pages 77–86. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Kanaris et al., 2006] Kanaris, I., Kanaris, K., and Stamatatos, E. (2006). Spam detection using character n-grams. In Antoniou, G., Potamias, G., Spyropoulos, C., and Plexousakis, D., editors, *Advances in Artificial Intelligence*, pages 95–104, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Kim et al., 2015] Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2015). Character-aware neural language models. In *AAAI 2016*.

[Klein et al., 2003] Klein, D., Smarr, J., Nguyen, H., and Manning, C. D. (2003). Named entity recognition with character-level models. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 180–183, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lee, 2007] Lee, J. (2007). A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479, Prague, Czech Republic. Association for Computational Linguistics.

[Levene, 2002] Levene, D. (2002). *Curse Or Blessing: What's in the Magic Bowl?* Parkes Institute pamphlet. University of Southampton.

[Levene, 2010] Levene, M. (2010). *An Introduction to Search Engines and Web Navigation*. John Wiley & Sons, Hoboken, New Jersey, 2nd edition.

[Levenshtein, 1966] Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

[Ma and Zhang, 2011] Ma, J. and Zhang, L. (2011). Modern BLAST programs. In *Problem Solving Handbook in Computational Biology and Bioinformatics*. Springer US.

[Madnani and Dorr, 2010] Madnani, N. and Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Comput. Linguist.*, 36(3):341–387.

[Mcnamee and Mayfield, 2004] Mcnamee, P. and Mayfield, J. (2004). Character n-gram tokenization for european language text retrieval. *Inf. Retr.*, 7(1-2):73–97.

[Mikolov et al., 2013] Mikolov, T., Yih, W., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter: Human Language Technologies*, pages 746–751. Association for Computational Linguistics.

[Rockwell, 2003] Rockwell, G. (2003). What is text analysis, really? *Literary and Linguistic Computing*, 18(2):209–219.

[Rommel, 2007] Rommel, T. (2007). Literary studies. In *A Companion to Digital Humanities*, pages 88–96. Blackwell Publishing Ltd.

[Schonfeld and Rutner, 2012] Schonfeld, R. C. and Rutner, J. (2012). Supporting the changing research practices of historians. *Final Report from Ithaka S.*

[Schulz et al., 2008] Schulz, M. H., Bauer, S., and Robinson, P. N. (2008). The generalised k-truncated suffix tree for time-and space-efficient searches in multiple dna or protein sequences. *IJBRA*, 4(1):81–95.

[Stamatatos, 2009] Stamatatos, E. (2009). Intrinsic plagiarism detection using character n-gram profiles. *threshold*, 2:1–500.

[Strauss and Eliot, 1860] Strauss, D. and Eliot, G. (1860). *The Life of Jesus: Critically Examined*. Number v. 1 in The Life of Jesus. C. Blanchard.

[Sweetnam et al., 2012] Sweetnam, M., Agosti, M., Orio, N., Ponchia, C., Steiner, C., Hillemann, E., Ó Siochrú, M., and Lawless, S. (2012). User needs for enhanced engagement with cultural heritage collections. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL)*, pages 64–75.

[Woodland et al., 2000] Woodland, P. C., Johnson, S. E., Jourlin, P., and Jones, K. S. (2000). Effects of out of vocabulary words in spoken document retrieval (poster session). In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 372–374, New York, NY, USA. ACM.

[Zhai, 2008] Zhai, C. (2008). Statistical language models for information retrieval: A critical review. *Found. Trends Inf. Retr.*, 2(3):137–213.

[Zhai, 2009] Zhai, C. (2009). *Statistical Language Models for Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Francisco.

[Zhai and Lafferty, 2004] Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214.