

# Reconstruction of Cuneiform Literary Texts as Text Matching

Fabian Simonjetz, Jussi Laasonen, Yunus Cobanoglu,  
Alexander Fraser, Enrique Jiménez

Ludwig-Maximilians-Universität München, Munich, Germany

fabian.simonjetz@caic.badw.de, jussi.laasonen@outlook.com, yunus.cobanoglu@yahoo.de,

fraser@cis.uni-muenchen.de, enrique.jimenez@lmu.de

## Abstract

Ancient Mesopotamian literature is riddled with gaps, caused by the decay and fragmentation of its writing material, clay tablets. The discovery of overlaps between fragments allows reconstruction to advance, but it is a slow and unsystematic process. Since new pieces are found and digitized constantly, NLP techniques can help to identify fragments and match them with existing text collections to restore complete literary works. We compare a number of approaches and determine that a character-level n-gram-based similarity matching approach works well for this problem, leading to a large speed-up for researchers in Assyriology.

**Keywords:** digital humanities, text matching, historical data, digital assyriology

## 1. Introduction

Cuneiform script, one of the oldest known writing systems, was deciphered in the mid-nineteenth century and has been extensively researched for more than a century. Traditionally, the corpus of cuneiform literature has primarily been processed manually, but with the increased availability of cuneiform data in digital form there is a growing interest in the development of automatic approaches to processing the data, ranging from OCR to end-to-end MT.

One of the idiosyncrasies of cuneiform is that it was written on clay tablets which, though much more durable than other media, are prone to breaking. As a result, the majority of the cuneiform corpus consists of fragments of larger tablets, and the whereabouts of the missing parts that would make a complete tablet are often unknown. However, if a text was written down more than once, as is often the case with literary and scientific work, the original text can be reconstructed by assembling partially overlapping fragments of different tablets covering the same content. So far, these matching pieces are accidental discoveries made by experts who come across overlapping passages in the course of transliterating fragments, normally using printed dictionaries or concordances. Although such discoveries happen on a regular basis, it would take decades to complete the literary and scientific works written on dozens of thousands of artifacts which await identification in museum cabinets around the world.

We formulate cuneiform fragment identification as an NLP problem and explore the feasibility of text matching methods to automate and speed up the process. Our main contributions are threefold:

1. We describe the properties and challenges of the cuneiform writing system and the fragmen-

tary nature of the data from a CL and NLP perspective.

2. We define the task of *fragment identification* as a text matching problem.
3. We present initial work on semi-automatic fragment matching. A method to generate synthetic test data from a corpus of partially identified cuneiform fragments is described, and results from a number of experimental matching approaches are compared.

The remainder of the paper is structured as follows. Section 2 presents the background of cuneiform fragments and the challenges of the writing system for NLP. The task of fragment identification is defined in Section 3. We proceed with a description of the dataset and a heuristic to generate test data in Section 4. The matching approaches we explored are defined in Section 5, followed by a discussion of the results in Section 6. Related work is summarized in Section 7, and some concluding remarks and directions for future research are given in Section 8.

## 2. Background

The literary works from ancient Mesopotamia were written on clay tablets inscribed with cuneiform script. If stored under favorable conditions, this medium is enormously durable but prone to breaking, so most works of Babylonian literature only survived in fragmentary form.

Cuneiform script was used for all kinds of written records, including personal letters and other everyday documents. Without any duplicates parts of documents that go missing are lost forever, but *literary and scientific* texts were frequently copied on multiple clay tablets which were often kept together in one and the same library, particularly in the first

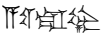
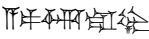
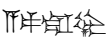



Figure 1: A fragmentary clay tablet of the Epic of Gilgamesh in the Iraq Museum (IM.67564). The tablet originally had three columns per side, of which only two fragmentary ones survive.

millennium BCE. When sections of a text are preserved on fragments of different tablets they may overlap and form partial duplicates, each of which typically contains a few signs that are missing on others. The manual identification of such pieces has traditionally been the key for the reconstruction of Babylonian literature. The process of identifying new fragments is therefore often a matter of luck: For example, a small fragment had been kept in a museum's drawer for over 100 years before it was identified as the beginning of the *Epic of Gilgamesh* (Kwasman, 1998).

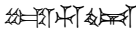
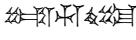

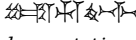
Part of the problem is that a substantial portion of the cuneiform fragments excavated so far is not fully documented. But even if digital records of every fragment in existence were available there are several factors that make their identification a non-trivial task, i.e., (i) a lack of orthographic conventions; (ii) a limited orthographic transparency; and (iii) data sparseness.

**Lack of orthographic conventions** There was never a strict orthography of any language written in cuneiform script. The same word, e.g., Akkadian *aparras* 'I will divide', can be written in different ways, either phonetically with various combinations of signs as in Examples (1i-iii) or by means of a logogram (1iv).






- (1) i.   
*a-par-ra-as*  
 ii.   
*a-pa-ar-ra-as*  
 iii.   
*a-pa-ra-as*

- iv.   
 KUD  
 'I will divide'

The variation is particularly pronounced in manuscripts from the first millennium BCE since in that period the original triptotic declension of the Akkadian language had been lost in the spoken variety but persisted in writing. With the vowels no longer corresponding to the use in previous periods, a high degree of variation, in particular in word-final position, began to emerge. Example (2) shows four different ways the word *lemutta* 'evil' is written in the manuscripts of the *Epic of Creation* I 44 (Heinrich, 2021). Matters are complicated further by the absence of word boundaries and punctuation.

- (2) i.   
*le-mut-ta*  
 ii.   
*le-mut-tu*  
 iii.   
*le-mut-tu<sub>4</sub>*  
 iv.   
*le-mut-ti*  
 'evil'

**Limited orthographic transparency** The readings of cuneiform signs are context-sensitive. The same grapheme can correspond to multiple phonemes, and vice versa, a phoneme can be represented by several distinct signs. Example (3) shows a selection of phonetic renderings of the sign UD, and (4) lists different signs all of which correspond to the phoneme /tu/, alongside the name of the sign and its reading (accompanied by an index which is added in Assyriological research as a means of disambiguating homophones).

- (3)   
*tam/tu<sub>2</sub>/par/ut/hiš/...*  
 (4) i.   
 TU (*tu<sub>1</sub>*)  
 ii.   
 UD (*tu<sub>2</sub>*)  
 iii.   
 DU (*tu<sub>3</sub>*)  
 iv.   
 TUM (*tu<sub>4</sub>*)

Another factor that affects the orthographic transparency of cuneiform is the fact that the repertoire of signs changed over time. For example, there are cases of signs that originally had distinct shapes such as the ones in (5i-iii) which coalesced into a single grapheme, in this case (6).

- (5) i. 𒀭  
 KU  
 ii. 𒀭  
 TUG<sub>2</sub>  
 iii. 𒀭  
 EŠ<sub>2</sub>
- (6) 𒀭  
 ku/tug<sub>2</sub>/eš<sub>2</sub>

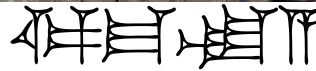
**Sparseness** Although there are cuneiform tablets with extensive amounts of text, fragments tend to be rather short. Some pieces can be as small as a fingernail and may contain no more than a few decipherable signs. With an average length of 111 signs in the database used for this study (cf. Section 4), cuneiform fragments can be considered the tweets of the ancient world. Accordingly, many of the challenges posed by Twitter and other short text data fully apply to cuneiform fragments, too: a limited number of content words, spelling variations, and other “ill-formed words” (Han and Baldwin, 2011, p. 368) make it more difficult to capture similarities between documents (Phan et al., 2008). Moreover, if fragments break unfavorably, the resulting loss of context combined with the polyvalence of signs can make the text difficult or even impossible to decipher, possibly leading to transliteration errors that directly impact matching accuracy. Since cuneiform data consists of low-resource languages these issues cannot be leveraged by using pre-trained models because the amounts of data needed for developing such models are not available.

### 3. Fragment Matching

The aim of the fragment identification task is to recover complete texts given a collection of fragmentary pieces. As discussed above, multiple copies of literary works such as the *Epic of Gilgamesh* are known. Although they feature slight variations (cf. Section 2), their contents are usually structured in a relatively stable line format, i.e., line breaks are generally consistent across different versions. Fragments of different tablets that cover the same textual material can thus be aligned similar to a photo collage: One fragment may hold the beginning of a line, and the rest may occur on another, allowing longer passages of the text to be reconstructed.

In what follows, we will refer to compilations of overlapping cuneiform documents with shared textual content as *chapters*, and to fragments confirmed to be part of a chapter as *manuscripts*.

Given a collection of chapters and a set of fragments that have not been identified yet, the identification problem can be defined as a type of *docu-*



di-ma-tu-a  
 ABZ457 ABZ342 ABZ58 ABZ579

Figure 2: Cropped photograph, Unicode string, ATF transliteration, and sequence of ABZ-signs of an individual cuneiform text line

*ment classification task*: Each chapter represents a category, and each new fragment needs to be labelled with respect to which category (if any) it belongs to. Formally, each cuneiform fragment and manuscript is represented by a sequence of text lines, and each line consists of a sequence of cuneiform *signs*. The aim of fragment identification is therefore to determine whether a given fragment is a manuscript of any known chapter.

Because of the ambiguities of cuneiform script and data sparseness, a fully automatic matching of fragments does not seem feasible at this point. Instead, we opt for a semi-automatic strategy with traditional NLP approaches, and our goal is to find a method that is capable of effectively aiding human experts in finding and validating potentially matching documents.

## 4. Data

### 4.1. Corpus

We use the collection of literary texts made available in the corpus of the *Electronic Babylonian Library* (eBL) project<sup>1</sup> (Jiménez et al., 2018a). As of the time of writing, it comprises 159 chapters containing between 1 and 70 (on average 8.82) manuscripts each, adding up to a total of 1,402 manuscripts. This collection of identified documents is complemented by a growing body of about 27,000 transliterated fragments in the eBL *Fragmentarium* (Jiménez et al., 2018b) that may include so far unidentified manuscripts.

Although there are Unicode blocks for *Cuneiform*,<sup>2</sup> *Cuneiform Numbers and Punctuation*,<sup>3</sup> and *Early Dynastic Cuneiform*,<sup>4</sup> Unicode is an inadequate way of rendering signs for our purposes because it amalgamates several different periods (Studt, 2007) which, in combination with

<sup>1</sup><https://www.ebl.lmu.de/corpus>

<sup>2</sup>[https://unicode.org/charts/nameslist/n\\_12000.html](https://unicode.org/charts/nameslist/n_12000.html)

<sup>3</sup>[https://unicode.org/charts/nameslist/n\\_12400.html](https://unicode.org/charts/nameslist/n_12400.html)

<sup>4</sup>[https://unicode.org/charts/nameslist/n\\_12480.html](https://unicode.org/charts/nameslist/n_12480.html)

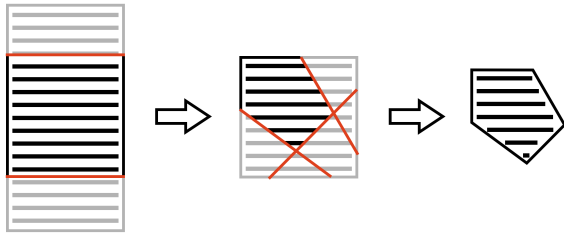


Figure 3: Schematic illustration of test fragment creation by first sampling passages of existing manuscripts and subsequently masking edges to simulate breaking

the diachronic change of the sign repertoire (cf. Section 2), would lead to an ambiguous representation. We therefore base our approach on *transliterations* of the texts.

In Digital Assyriological research, cuneiform is transliterated into Latin script following the so-called *ASCII Transliteration Format* (ATF; see the third row in Fig. 2) introduced by the *Cuneiform Digital Library Initiative* (CDLI; CDLI contributors, 2016) in the late 1990s and early 2000s. The specifications were later modified by the *Open Richly Annotated Cuneiform Corpus* consortium (Oracc; Novotny et al., 2014). The eBL data employs an extended version of ATF which additionally links each sign to an entry in a dictionary of cuneiform specific for the first millennium BCE, Borger's (1988) *Assyrisch-Babylonische Zeichenliste* 'Assyrian-Babylonian sign list' where each sign receives a unique identifier we refer to as *ABZ number*. Readings, logograms, numbers, and compound graphemes in ATF format can be converted to a sequence of ABZ numbers by assigning it the index of a sign with a matching reading or name from Borger (1988). If the sign does not have an ABZ number, the name of the sign is used instead. Our experiments are based on the resulting sequence of ABZ signs which is the only way of digitally representing cuneiform that conforms to the repertoire of cuneiform signs in the first millennium BCE.

#### 4.2. Test Data Generation

From the eBL chapters we derived a test set of artificial fragments by extracting manuscripts from their source chapter and 'breaking' them into smaller fragments via a simple masking heuristic.

100 manuscripts were randomly extracted from the corpus chapters. The documents were masked by first truncating them to extract random subsequences of lines, matching the length distribution of real fragments in the eBL fragment database (i.e., an average length of about 15 lines with a minimum of 2 and a maximum of 20). The extracts were then further fragmented by fitting their contents into

grids of signs and 'breaking away' corners by drawing break lines between random points along the edges of the grids and excluding the signs that fall outside of the resulting boundaries (see Fig. 3). Albeit simple, this strategy results in a large variety of shapes that resemble real cuneiform fragments both in terms of size and the existence of partially missing lines. This process was repeated 10 times to generate 1,000 test documents. After filtering out fragments that ended up empty, we obtained a set of 965 artificial fragments.

## 5. Experiments

We conducted a number of experiments to explore the feasibility of the task and establish baselines. The approaches are briefly described below.

**Bag of words** As a naive way to compute the pairwise similarity between documents the overlap of their respective bags of words (BOW), or in this case bags of cuneiform signs, can be used. It is easy to implement, but all information regarding the order of signs is lost. We used the Jaccard index to rank the intersections.

**Longest common substring** The longest common substring (LCS) of two sequences of symbols is the longest sequence of consecutive symbols that occurs in both of the input strings. In contrast to the longest common subsequence, a related problem, LCS does not allow gaps between the individual symbols. In order to compute the LCS of a fragment and a chapter, it was computed pairwise between the fragment and each manuscript of the chapter, and the longest match of all comparisons was considered.

**Sequence alignment** The Needleman-Wunsch alignment algorithm (Needleman and Wunsch, 1970) computes the optimal local or global alignment between two strings. Finding the best alignment is equivalent to minimizing the edit distance of two input sequences. It has been successfully applied for the alignment of genome sequences in the field of Bioinformatics (Altschul et al., 1990) which has parallels with alignment of natural language data. The algorithm allows for gaps between partially matching sequences and can be modified with a custom matching function, so in principle it can be adapted to account for common, systematic cuneiform sign variations. As for LCS, the alignment was computed for each fragment-manuscript pair per chapter and the best match was considered.

**N-gram matching** Another way to approach the matching problem that conserves sequential information while allowing variations of individual signs are *n-gram models*. The input sequences are partitioned into overlapping subsequences of length  $n$ . For the task of fragment identification it seems particularly adequate because it allows the individual manuscripts of a chapter to be combined into a joint representation by merging their  $n$ -gram distributions. This solves a major problem of LCS and Needleman-Wunsch alignment which only take individual manuscripts into account, potentially leading to systematic misses if a fragment partially matches two or more manuscripts. This could result in a low score even if the total overlap is high. In addition, the creation of, and similarity computation between,  $n$ -gram models are very fast, and  $n$ -gram models can be created once and stored for later usage while LCS and Needleman-Wunsch alignment need to be recomputed every time, making them less adequate for real-time applications.

While the frequencies of  $n$ -grams can be important for general document matching tasks, the overlaps between the manuscripts of a chapter in our case mean that the frequencies of those  $n$ -grams that occur in multiple manuscripts would arbitrarily be inflated, so we use a binary bag-of- $n$ -grams approach disregarding the frequencies. That is, each chapter and fragment is represented as a *set of  $n$ -grams*, merging the sets of each manuscript for a given chapter.

When deciding for a similarity metric it must be considered that fragments are generally expected to be much shorter than the chapters they belong to, and vice versa, a large fragment could also potentially be part of a shorter chapter (e.g., if not many manuscripts have been identified yet, or if the fragment contains excerpts of multiple texts). We want a high score if a large portion of the smaller document is included in the larger one, regardless of any size differences. The similarity is therefore computed as the proportion of  $n$ -grams of the shorter document that occur in the longer one, a metric referred to as *overlap coefficient* or sometimes *Szymkiewicz-Simpson coefficient* (Vijaymeena and Kavitha, 2016). It returns 0 if a chapter  $C$  and a fragment  $F$  share no  $n$ -grams and 1 if all  $n$ -grams of  $F$  are included in  $C$  (or vice versa). It is computed as follows:

$$\text{overlap}(F, C) = \frac{|F \cap C|}{\min(|F|, |C|)} \quad (1)$$

As an extension of the base approach, we implemented two ways of weighting  $n$ -grams. First, overlaps of longer  $n$ -grams should be valued considerably higher than shorter ones. For example, if a fragment and a chapter share three 1-grams, i.e., three individual signs, it is less indicative of

a match than a single shared 3-gram. In order to account for that, we adapt the above formula such that instead of the set sizes we employ the sum of the squared length of each  $n$ -gram in the set.

Second, in addition to the length of shared  $n$ -grams, a differentiation should be made regarding the overall prevalence of the  $n$ -grams in the corpus. As for stopwords in natural language data in general, overlaps of rare  $n$ -grams should be considered more important than overlaps of very common ones. This can be accounted for by *Term Frequency-Inverse Document Frequency* weighting (TF-IDF). Originally proposed by Sparck Jones (1972), TF-IDF assigns a greater importance to terms that occur in fewer documents in a reference collection than more common ones (for a concise and accessible introduction see, e.g., Robertson, 2004). In our specific use case, terms are sign  $n$ -grams and the reference collection are the eBL chapters. TF-IDF is defined as:

$$\text{TF-IDF}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D) \quad (2)$$

where

- $t$  represents a term (in our case  $n$ -gram),
- $d$  represents a document (chapter), and
- $D$  represents the collection of all chapters.

As we are dealing with sets of cuneiform sign  $n$ -grams, document-wise frequencies are irrelevant, so we compute the term frequency  $\text{tf}$  as binary depending on whether the  $n$ -gram occurs in the document or not, i.e.:

$$\text{tf}(t, d) = \begin{cases} 1 & t \in d \\ 0 & t \notin d \end{cases} \quad (3)$$

We compute the smoothed inverse document frequency as per the equation below:

$$\text{idf}(t, D) = \log \left( \frac{1 + |D|}{1 + |\{d \in D : t \in d\}|} \right) \quad (4)$$

where

- $|D|$  is the total number of chapters
- $|\{d \in D : t \in d\}|$  is the number of chapters that contain the  $n$ -gram  $t$ .

The matching tool is publicly available.<sup>5</sup> With the exception of the Needleman-Wunsch algorithm, the tests were run on a MacBook Pro with a 1.4 GHz Quad-Core Intel Core i5 processor and 8GB of RAM. For our dataset, BOW and default  $n$ -gram matching take a few seconds to finish. The  $n$ -gram extraction and TF-IDF computation take about

<sup>5</sup><https://github.com/ElectronicBabylonianLiterature/ngram-matcher>

Approach	Precision@3
BOW	0.14
LCS	0.90
Needleman-Wunsch	0.79

Table 1: Results of the baseline approaches, reported as proportions of chapters which included a correctly assigned fragment among the top 3 matches (Precision@3).

30 seconds each. Computing the overlaps with weights takes less than 2 minutes. LCS is considerably slower at about 30 minutes. The Needleman-Wunsch alignment is computationally expensive and was run on a 10-core cloud-based supercomputer. Even so, the algorithm took about 24 hours to complete.

## 6. Results and Discussion

For the evaluation, the set of artificial fragments (cf. Section 4.2) was compared element-wise with each of the chapters from the eBL dataset (excluding the manuscripts extracted for the creation of the test fragments). Each fragment in the resulting similarity matrix is considered a correct match if its source chapter is among the the top 3 most similar items ( $Precision@k$  with  $k = 3$ ). Although this metric has certain limitations (Manning et al., 2008, p. 148) it appears suitable for the task of cuneiform fragment identification where we want to reduce the number of elements for manual inspection to an amount that is manageable yet maximizes the chance of finding matches. The results of the baseline approaches are summarized in Table 1, and the results of the n-gram-based variants we tested are in Table 2.

Since the order of signs is crucial to find overlapping text segments it is not surprising that the BOW baseline is outperformed by all approaches that take sequential information into account. The low score also indicates that Jaccard index is inadequate because it is sensitive to size differences between the input documents.

LCS shows a very high  $Precision@k$  although it can be expected to degrade if sign variations, deletions, or reordering of signs are present (which are not included in the synthetic test data). The Needleman-Wunsch alignment algorithm suffers less from this problem as it allows gaps to occur in the input sequences. Both approaches have the benefit that they make explicit which parts of the compared documents correspond to each other which facilitates manual evaluation. However, this advantage is outweighed by the fact that fragments can only be aligned against individual manuscripts, not against a chapter as a whole, which could lead

to systematic misses if a fragment partly overlaps multiple manuscripts, resulting in individually low scores without a way to reconcile them. Both approaches are also computationally expensive, in the case of the alignment approach to a degree that it becomes impractical for our scenario of an online matching tool.

The n-gram approach combines the best of both worlds. It retains information about the relative order of signs while orthographic variations or missing segments are not penalized as much as for LCS. Furthermore, it allows for all manuscripts of a chapter to be naturally merged into a single, joint representation. Since n-grams can be precomputed and stored for later usage, the approach is highly performant, too, especially in the unweighted variants, making it ideal for frequent updating and real-time experimentation. Since the different weighting strategies appear to have little effect on the overall precision while adding considerable computational load (especially for the TF-IDF-based weighting), the basic n-gram approach appears to be preferable in applications where processing time is critical. Regarding the choice of n-values, the results suggest that n-grams up to a size of 3 provide the best balance between computational efficiency and precision, with diminishing returns for higher n.

Our results show that a basic n-gram-based matching approach is capable of effectively narrowing down a collection of fragments to a small set of candidates that are very likely to contain matches. Thanks to its low computational requirements it can be utilized, e.g., in a web-based click-through fragment identification tool where a human expert can rapidly check potential matches. Although LCS and Needleman-Wunsch alignment are too slow to be useful for processing the corpus as a whole, they could facilitate manual evaluation as a means of visualizing parallels between pairs of fragments and chapters known to be similar due to a high number of shared n-grams.

It must be noted that the experiment considerably simplifies the real task. We only compare synthetic fragments against the eBL chapters which were drawn from the same manuscripts used for evaluation. Although the original documents were excluded in the evaluation step (because otherwise, each test fragment would be matched against itself), each item is known in advance to match exactly one of the chapters. This is not the case when matching real fragments as they mostly belong to no chapter at all, so the precision of the approach must be expected to be much lower for the real task.

To assess the usefulness of the approach in a more realistic scenario, we matched a newly added eBL chapter (*Maqlû* ‘Burning’, an anti-witchcraft

Approach	[1, 1]	[1, 2]	[1, 3]	[1, 4]	[1, 5]
<b>n-grams</b>	0.55	0.83	0.88	<b>0.91</b>	0.91
<b>n-grams + L</b>	0.55	0.82	<b>0.91</b>	0.92	0.92
<b>n-grams + TF-IDF</b>	0.55	0.83	<b>0.92</b>	0.92	0.92
<b>n-grams + TF-IDF + L</b>	0.55	0.83	<b>0.94</b>	0.94	0.93

Table 2: Results of the n-gram-based approaches, reported as proportions of chapters which included a correctly assigned fragment among the top 3 matches (precision at  $k$  with  $k = 3$ ), for combinations of  $n$ -values with  $n \in [1, 5]$  and weighting by n-gram length (L) and TF-IDF.

text)<sup>6</sup> against the entire eBL collection of approximately 27,000 fragments, with encouraging results: Inspection of the top-ranking 300 fragments (1.2% of all candidates) by a domain expert led to the discovery of three previously unidentified manuscripts (K.18050,<sup>7</sup> Sm.1215,<sup>8</sup> and Sm.1852,<sup>9</sup>).

## 7. Related Work

Computational approaches to processing cuneiform data, so-called *Digital* or *Computational Assyriology*, have seen a growing interest in recent years due to the increasing availability of digital corpora and the desire to speed up traditional methods. Sahala (2021) gives an overview and evaluation of studies between 1960 and 2020, including the application of NLP techniques to linguistic annotation of cuneiform texts (Sahala et al., 2020), conversion of Unicode cuneiform to transliterations (Gordin et al., 2020), MT (Punia et al., 2020; Guthertz et al., 2023), automatic reconstruction of textual gaps (Fetaya et al., 2020; Lazar et al., 2021), and approaches to OCR (Dencker et al., 2020).

Piecing together torn, ripped, or otherwise damaged documents and other objects is an interdisciplinary task with a multitude of applications, e.g., in archaeology, philology, and forensics (Kleber et al., 2009; Papaodysseus et al., 2002). Methods focus primarily on image processing techniques to find matching contours, similar to jigsaw puzzle solving, but in our case the documents to merge do not necessarily match physically. Therefore, it is more appropriate to rely on the linguistic information the tablets provide which makes for an NLP problem with parallels, e.g., in (short) text matching, clustering, text-reuse detection, and token alignment.

The aim of text matching is to determine the degree of similarity between two or more pieces of text which is crucial for countless applications. Matching short text, i.e., documents that consist only of a few words up to about a paragraph, is a subtask

of text matching that is particularly challenging due to the reduced informativeness of short text which limits the effectiveness of techniques developed on longer text forms (Hu et al., 2019; Jin et al., 2011). Mitigation strategies include normalization of the input (Han and Baldwin, 2011), utilization of pre-trained neural networks (Hu et al., 2019), and enhancement of the short texts with auxiliary data (Jin et al., 2011). Although the latter is partly applicable, these techniques generally presuppose the existence of larger amounts of preprocessed data which are not available for cuneiform. For this reason, traditional rule-based approaches appear to be the more promising route to take at this point, especially due to the structural aspects of matching cuneiform fragments (i.e., the task is to match highly parallel sequences rather than less structured types of data).

Text reuse or plagiarism detection is a core NLP task with a rich body of literature we cannot fully acknowledge here. Closely related to our task is work on historical and other non-standard data. E.g., Büchler et al. (2014) explore text re-use detection in different variants of Bible verses which pose very similar challenges to cuneiform literary work in that there are relatively fixed units of segmentation (verses in biblical texts vs. lines in cuneiform) and a high degree of structural parallelism among textual variants caused, e.g., by different writing schools, dialectal influence, or diachronic change. Büchler et al. also utilize n-grams which belong to the standard repertoire of NLP techniques that proved effective, among others, for text classification in general (Cavnar and Trenkle, 1994), text re-use detection (e.g., Clough et al., 2002; Bensalem et al., 2014), and authorship attribution (Wright, 2017; Sari et al., 2017).

Orthographic variations are not exclusive to cuneiform but are also found, e.g., in Hebrew: Shmidman et al. (2016) present a method that uses only the two most infrequent letters for comparison purposes. This system accounts well for systematic orthographic variations in Hebrew script since the most frequent signs are those that appear most frequently in orthographic variations. It is, however, less appropriate for the more arbitrary variations that are prevalent in cuneiform, which do not in-

<sup>6</sup>[www.ebl.lmu.de/corpus/Mag/1/1](http://www.ebl.lmu.de/corpus/Mag/1/1)

<sup>7</sup>[www.ebl.lmu.de/K.18050](http://www.ebl.lmu.de/K.18050)

<sup>8</sup>[www.ebl.lmu.de/Sm.1215](http://www.ebl.lmu.de/Sm.1215)

<sup>9</sup>[www.ebl.lmu.de/Sm.1852](http://www.ebl.lmu.de/Sm.1852)

volve the systematic insertion or removal of any sign in particular.

Since matching cuneiform fragments are expected to overlap structurally at the line and token level, the task is also related to alignment problems. Alignment of words and phrases is relevant, e.g., in multilingual corpora, and has also been applied to plagiarism detection (Sanchez-Perez et al., 2014; Momtaz et al., 2016; Manjavacas et al., 2019).

A two-stage process where first the documents to be matched are filtered using n-grams, followed by the application of an alignment algorithm that detects inserted or deleted passages is proposed by O'Neill et al. (2021).

## 8. Conclusion and Future Work

We introduced cuneiform fragment identification as a type of text matching problem. The idiosyncrasies of clay tablets and cuneiform script make for a unique set of challenges that lend themselves to be tackled with NLP methods. We have generated synthetic test data by simulating breaking of fragments and explored experimental approaches to simplifying the process of finding matching text fragments. We found that n-grams offer an elegant and efficient way to significantly reduce the number of potential matches from 1,000 to 3 candidates that contain a match up to 94% of the time.

We believe that the task of fragment identification is an intriguing problem that opens up a whole array of research opportunities. The data offers much more than just the sign sequences, e.g., the arrangement of the contents and information about fracture lines can be leveraged to perform fully automatic matching. Many of the fragments in the eBL database include photographs, too, which have the potential for developing multi-modal approaches.

## Acknowledgements

This work was supported by the projects “Cuneiform Artefacts of Iraq in Context” (CAIC)<sup>10</sup> funded by the Bavarian Academy of the Sciences (project number II.C.29), and “electronic Babylonian Literature” (eBL)<sup>11</sup> funded by a Sofja Kovalevskaja Award (Alexander von Humboldt-Stiftung).

## 9. Bibliographical References

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. Ba-

sic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Imene Bensalem, Paolo Rosso, and Salim Chikhi. 2014. [Intrinsic plagiarism detection using n-gram classes](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1459–1464, Doha, Qatar. Association for Computational Linguistics.

Rykle Borger. 1988. *Assyrisch-babylonische Zeichenliste*. Ugarit-Verlag, Neukirchen-Vluyn.

Marco Buehler, Philip R. Burns, Martin Müller, Emily Franzini, and Greta Franzini. 2014. [Towards a historical text re-use detection](#). In *Text Mining*.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, volume 161175, pages 161–169. Las Vegas, NV.

Paul Clough, Robert Gaizauskas, Scott S.L. Piao, and Yorick Wilks. 2002. [Measuring text reuse](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 152–159, Philadelphia, Pennsylvania, USA.

Tobias Dencker, Pablo Klinkisch, Stefan M. Maul, and Björn Ommer. 2020. [Deep learning of cuneiform sign detection with weak supervision using transliteration alignment](#). *PLOS ONE*, 15(12):1–21.

Ethan Fetaya, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. 2020. [Restoration of fragmentary Babylonian texts using recurrent neural networks](#). *Proceedings of the National Academy of Sciences*, 117(37):22743–22751.

Shai Gordin, Gai Gutherz, Ariel Elazary, Avital Romach, Enrique Jiménez, Jonathan Berant, and Yoram Cohen. 2020. [Reading Akkadian cuneiform using Natural Language Processing](#). *PLOS ONE*, 15(10):1–16.

Gai Gutherz, Shai Gordin, Luis Sáenz, Omer Levy, and Jonathan Berant. 2023. [Translating Akkadian to English with neural machine translation](#). *PNAS Nexus*, 2(5). Pgd096.

Bo Han and Timothy Baldwin. 2011. [Lexical normalisation of short text messages: Makn sens a #twitter](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

Adrian C. Heinrich. 2021. [Poem of Creation \(Enūma eliš\)](#). With contributions by Zs. J. Földi,

<sup>10</sup><https://caic.badw.de/>

<sup>11</sup><https://www.ebl.lmu.de/>



- A. Hättinen, E. Jiménez and T. Mitto. Translated by Benjamin R. Foster. Electronic Babylonian Library.
- Weiwei Hu, Anhong Dang, and Ying Tan. 2019. [A survey of state-of-the-art short text matching algorithms](#). *Communications in Computer and Information Science*, 1071:211–219.
- Ou Jin, Nathan N. Liu, Kai Zhao, Yong Yu, and Qiang Yang. 2011. [Transferring topical knowledge from auxiliary long texts for short text clustering](#). *International Conference on Information and Knowledge Management, Proceedings*, pages 775–784.
- Florian Kleber, Markus Diem, and Robert Sablatnig. 2009. [Torn document analysis as a prerequisite for reconstruction](#). In *2009 15th International Conference on Virtual Systems and Multimedia*, pages 143–148.
- Theodore Kwasman. 1998. A new join to the Epic of Gilgameš Tablet I. *Nouvelles Assyriologiques Brèves et Utilitaires*, 1998/99.
- Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. [Filling the gaps in ancient Akkadian texts: A masked language modelling approach](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4682–4691, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Enrique Manjavacas, Brian Long, and Mike Kestemont. 2019. [On the feasibility of automated detection of allusive text reuse](#). In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 104–114, Minneapolis, USA. Association for Computational Linguistics.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
- Mozhgan Momtaz, Kayvan Bijari, Mostafa Salehi, and Hadi Veisi. 2016. Graph-based approach to text alignment for plagiarism detection in Persian documents. *CEUR Workshop Proceedings*, 1737:176–179.
- Saul B. Needleman and Christian D. Wunsch. 1970. [A general method applicable to the search for similarities in the amino acid sequence of two proteins](#). *Journal of Molecular Biology*, 48.
- Helen O’Neill, Anne Welsh, David A. Smith, Glenn Roe, and Melissa Terras. 2021. [Text mining Mill](#). *Computationally detecting influence in the writings of John Stuart Mill from library records*. *Digital Scholarship in the Humanities*, 36(4):1013–1029.
- C. Papaodysseus, T. Panagopoulos, M. Exarhos, C. Triantafillou, D. Fragoulis, and C. Doumas. 2002. [Contour-shape based reconstruction of fragmented, 1600 BC wall paintings](#). *IEEE Transactions on Signal Processing*, 50(6):1277–1288.
- Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100.
- Ravneet Punia, Niko Schenk, Christian Chiarcos, and Émilie Pagé-Perron. 2020. [Towards the first machine translation system for Sumerian transliterations](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3454–3460, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Stephen Robertson. 2004. [Understanding inverse document frequency: on theoretical arguments for idf](#). *Journal of Documentation*, 60(5):503–520.
- Aleksi Sahala. 2021. *Contributions to Computational Assyriology*. Helsingin yliopisto.
- Aleksi Sahala, Miikka Silfverberg, Antti Arppe, and Krister Lindén. 2020. [Automated phonological transcription of Akkadian cuneiform text](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3528–3534, Marseille, France. European Language Resources Association.
- Miguel A. Sanchez-Perez, Grigori Sidorov, and Alexander Gelbukh. 2014. The winning approach to text alignment for text reuse detection at PAN 2014: Notebook for PAN at CLEF 2014. *CEUR Workshop Proceedings*, 1180:1004–1011.
- Yunita Sari, Andreas Vlachos, and Mark Stevenson. 2017. [Continuous n-gram representations for authorship attribution](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 267–273, Valencia, Spain.
- Avi Shmidman, Moshe Koppel, and Ely Porat. 2016. [Identification of parallel passages across a large Hebrew/Aramaic corpus](#). *CoRR*.
- Karen Sparck Jones. 1972. [A statistical interpretation of term specificity and its application in retrieval](#). *Journal of Documentation*, 28(1):11–21.

Margret Studt. 2007. A practical and critical guide to the Unicode blocks 'Cuneiform' and 'Cuneiform Numbers' of Unicode Standard Version 5.0.

M.K. Vijaymeena and K. Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(1):19–28.

David Wright. 2017. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics*, 22:212–241.

## 10. Language Resource References

CDLI contributors. 2016. *Cuneiform Digital Library Initiative (CDLI)*. Cuneiform Digital Library Initiative. PID <https://cdli.mpiwg-berlin.mpg.de/>.

Enrique Jiménez and others. 2018a. *Electronic Babylonian Library: Corpus*. Electronic Babylonian Literature (eBL) Project. PID <https://www.ebl.lmu.de/corpus>.

Enrique Jiménez and others. 2018b. *Electronic Babylonian Library: Fragmentarium*. Electronic Babylonian Literature (eBL) Project. PID <https://www.ebl.lmu.de/fragmentarium>.

Jamie Novotny and Eleanor Robson and Steve Tinney and Niek Veldhuis. 2014. *Oracc: The Open Richly Annotated Cuneiform Corpus*. Oracc Consortium. PID <http://oracc.museum.upenn.edu>.