

Automatic Identification of Cuneiform Fragments Using String Alignment Algorithms

Anonymous ACL submission

Abstract

The literature from ancient Mesopotamia is still riddled with textual lacunas. Scores of fragments which could potentially fill those lacunas lie unidentified in museums's cabinets, but their identification has traditionally been slow and laborious due to the ambiguities of cuneiform script. The present article presents a novel method for dealing with these ambiguities by using a string alignment algorithm adapted for cuneiform, which makes identification much easier and speeds up the process dramatically. The availability of this algorithm and of corpora on which to use it will advance significantly the reconstruction of Mesopotamian literature.

1 Introduction

The literary works from ancient Mesopotamia are in a state of reconstruction. The vehicle in which they were written, clay tablets inscribed with cuneiform script, is enormously durable but also brittle, so the masterpieces of Babylonian literature have come down to us in a fragmentary form. Since their rediscovery in the mid-nineteenth century, cuneiformists have striven to put together the classics of Mesopotamian literature from scores of fragments scattered throughout the world's museums, but the process is far from finished. Most texts are still riddled with textual lacunae and tens of thousands of fragments lie unidentified in museums' cabinets.

Most literary texts were copied on more than one clay tablet (so-called "manuscripts" of a text), often kept together in one and the same library. Frequently sections of texts are preserved in fragments from different manuscripts (referred to as "(partial) duplicates"), which partially overlap. Each fragment typically preserves a few signs not preserved on its duplicate; with the help of this signs other duplicates can be identified. The identification of these duplicates has traditionally been the

key for the reconstruction of Babylonian literature. Their potential in this respect is far from exhausted, since many hitherto unidentified fragments will, once identified, turn out to be duplicates of other fragments.

The identification of new fragments is rendered particularly challenging by the ambiguities of cuneiform script. Each sign has several possible phonetic renderings (so-called "readings"), and most syllables can be represented by means of different signs. There was never a strict orthography of any language written in cuneiform script, so the same word, e.g. Akkadian *aparras*, "I will divide", can be written phonetically with various combinations of signs (such as *a-par-ra-as*, *a-pa-ar-ra-as*, and *a-pa-ra-as*), and also by means of a word-sign (so-called "logogram", such as KUD). The variation is particularly pronounced in manuscripts from the first millennium BCE: the original triptotic declension of the Akkadian language had been lost in the spoken variety, but was still preserved in the written language. The vowels represented in the written language, however, no longer correspond with the use in previous periods, and a large degree of variation occurs at word end: for instance, the word *lemutta* in the 'Epic of Creation' I 44 is written in four manuscripts in four different ways, as *le-mut-ta*, *le-mut-tu*, *le-mut-tu₄*, and *le-mut-ti*.

The identification of fragments has traditionally been done manually, making use of existing dictionaries (in particular the *Chicago Assyrian Dictionary* = [Oppenheim et al. 1956-2011](#)), or else of concordances compiled ad-hoc (see [Borger 1991](#), p. 51). Due to the peculiarities of cuneiform script, however, it is often difficult to make out individual words from sequences of signs bereft of context: if no unequivocal word can be made of the signs preserved, or if the word is a common one, such dictionaries and concordances are of limited use. For this reason, the identification has always depended on chance, and the success has, consequently, been

082 limited: thus, the catalogue of small fragments in
083 the Nineveh collection, whose compilation took
084 almost twenty years of painstaking work of the
085 foremost cuneiformist at the time (Lambert, 1992),
086 succeeded in identifying some 335 of the 5,400
087 pieces catalogued, i.e. 6.2%. Due to the low suc-
088 cess rate, the reconstruction is an extremely slow
089 process: it took, for instance, over 100 years to
090 identify the beginning of the ‘Epic of Gilgameš’
091 in a small fragment kept in a museum’s drawer
092 (Kwasman, 1998).

093 The potential in the use of computers for auto-
094 matic identification of cuneiform fragments has
095 long been noticed, but never realised, perhaps be-
096 cause of the skepticism with which some of the
097 foremost specialists regarded the process.¹

098 We obtained the corpus of literary texts that have
099 been made available to the community by the “Elec-
100 tronic Babylonian Literature” project². The avail-
101 ability of this data enables the development of algo-
102 rithms to account for the ambiguities of cuneiform
103 script, described above. Such ambiguities are best
104 addressed by means of alignment algorithms that
105 account for gaps and variations in the aligned se-
106 quences. The present paper presents the first use of
107 string alignment algorithms for the identification
108 of cuneiform fragments. We have created a search
109 system that allows searches of cuneiform fragments
110 using the sign mapping described below (3). More-
111 over, we have integrated the string alignment algo-
112 rithms into the search system, thus enabling special-
113 ists to perform alignments on cuneiform fragments
114 entered using the lightweight transliteration system
115 referred to as ATF,³ the standard in the field.

116 2 Previous Work

117 We review previous work on solving similar align-
118 ment problems. Multiple sequence alignment has
119 been applied successfully in linguistics (Prokić
120 et al., 2009; List, 2011), but it does not produce
121 optimal alignment, and our use case only requires
122 pairwise alignment. Sanchez-Perez et al. (2014)
123 propose a graph-based approach to text alignment
124 for plagiarism detection, but the algorithm does

¹“It is impracticable to use computers to identify such pieces, since the ambiguities of cuneiform script and the lack of reconstructed texts to be used as a basis for machine identifications rule out any such method. It is accordingly necessary to rely on the human memory and on aids such as dictionaries and glossaries” (Lambert, 1992, p. ix).

²eBL, <https://www.ebl.lmu.de/>

³<http://oracc.org/doc/help/editinginatf/index.html>

125 not account for gaps or lacunae and does not out-
126 put the actual alignment. Williams et al. (2014),
127 Williams (2015), and Brusuelas (2016) describe
128 the adaptation of the BLAST algorithm for the
129 Greek alphabet (“Greek-BLAST”) in the frame-
130 work of the Ancient Lives project⁴, which involved
131 the implementation of a Greek Letter Oriented Sub-
132 stitution Matrix (GLOSUM) in lieu of the BLO-
133 SUM (BLOcks Substitution Matrix) substitution
134 matrix. Greek manuscripts, however, do not con-
135 tain the same sort of orthographic variation that
136 can be found in cuneiform texts: this variation,
137 described above, means that the background fre-
138 quency that is at the core of GLOSUM cannot be
139 used for cuneiform. Momtaz et al. (2016) use a
140 sentence similarity measure for text alignment for
141 text reuse detection. The approach works on the
142 sentence level, but many fragments do not con-
143 tain complete sentences or lines. Shmidman et al.
144 (2016) present a method that uses only the two
145 most infrequent letters for comparison purposes:
146 this system accounts well for orthographic variation
147 in Hebrew script, since the most frequent signs are
148 those that appear most frequently in orthographic
149 variations (e.g. י in דלמא vs. דילמא). It is, how-
150 ever, inappropriate for the sorts of variations that
151 are common in cuneiform, which do not involve
152 the insertion or removal of any sign in particular.

153 3 Methodology

154 As we have described above, the alignment of
155 cuneiform fragments is challenging. After carry-
156 ing out the survey of previous work in the previ-
157 ous section, we determined that the best solution
158 would be to use sequence alignment. In order to
159 account for the sort of variations that are common
160 in cuneiform script, and for the sort of gaps that
161 occur in cuneiform fragments, we needed a string
162 alignment library in which the scoring could be
163 defined programmatically instead of specifying the
164 whole scoring matrix (due to a large number of
165 cuneiform signs, such a matrix would be unpracti-
166 cally large). For that reason, we choose the library
167 *python-alignment*⁵. We chose the library because
168 it is pure Python, works with arbitrary vocabulary
169 and can perform local (Smith and Waterman, 1981)
170 and global alignments (Needleman and Wunsch,
171 1970).

⁴<https://www.ancientlives.org/>

⁵<https://github.com/eseraygun/python-alignment>

The sign mapping problem Cuneiform signs are polyvalent, i.e., the same sign (e.g. UD) can have multiple phonetic readings (e.g. *tam*, *tú*, *par*, *ut*, *hiš*), and the same phonetic readings can be expressed with several discrete signs: for instance, the syllable /*tu*/ can be written with the sign TU (the reading is then $tu_{(1)}$), UD (= tu_2), DU (= tu_3), TUM (= tu_4). The complexity of the system is further compounded by the fact that the repertoire of signs changed over time: signs that originally had distinct shapes coalesced into the same signs (e.g. the sign forms KU, TUG₂, and EŠ₂, originally independent, coalesced into the same sign, which had *ku*, *tug₂*, and *eš₂* as possible readings). This fact means that the sign mapping adopted for the alignment should be geared towards the specific period of the texts involved. For this reason, the cuneiform Unicode block (U+12000–U+123FF, U+12400–U+1247F, U+12480–U+1254F) cannot be used for the alignment, because its sign repertoire amalgamates several different periods (Studt, 2007). The repertoire that fits best manuscripts from the first millennium BCE is the list from Assyrisch-Babylonische Zeichenliste (henceforth ABZ, = Borger 1988). In total, our sign database contains:

- 2490 signs (1936 composite and 554 simple) with 9194 readings.

- 826 signs with ABZ, which can be mapped to 694 ABZ numbers.

Readings, logograms, numbers, and compound graphemes in ATF format are converted to a sequence of ABZ numbers by getting a sign with matching reading or name from the sign list and using the ABZ value of that sign. If the sign does not have an ABZ number, the sign name is used instead. Compound graphemes and numbers can be arbitrary, so if they do not match a sign or reading from the sign list, they are used as is. For example 2 is a reading of MIN (ABZ570), but 9 is not a reading. "2" would become "ABZ570" but "9" would stay as "9".

Some signs are formed by combining several signs (e.g. the sign ŠAB consists of PA + IB): to account for this phenomenon, we treat juxtaposed signs outside of groupings as individual signs (e.g. |PA.IB| from ŠAB will be become PA IB, but |GA₂×(ME.EN)| would remain unchanged.).

Scoring Since no reference alignments exist for cuneiform texts, we determined parameters in preliminary experimentation on aligning fragments

to the manuscripts. We ran local and global alignments of fragments presented in table 2 in appendix A against all the manuscripts, starting with default values: 2 for a match, -1 for substitution, and -2 for a gap.

Line breaks are essential in literary texts which have an established line structure. Therefore we assigned a high score for line break match and a significant penalty for substitution.

Some signs are transliterated as variants when the correct sign is unclear. Substituting a variant to one of those signs should not be penalized, and we use the maximum score of all possible combinations for the score of the alignment substitution.

Some substitutions are common, and some signs have a similar meaning. Therefore, these substitutions should not be penalized, and to determine common pairs, we performed local alignments, filtered out alignments with similarity less than 80%, and calculated the frequency of each pair. From the most frequent pairs, we selected those which made sense (Table 3 in appendix C). These common substitutions have a positive but smaller score than the exact match.

The textual lacunae are problematic for the alignment, and we extended the python-alignment library with an affine gap penalty (Gotoh, 1982) has a significant penalty, but the extension is very cheap to allow the gap to cover the missing signs. Such a low penalty usually is not recommended (Smith and Waterman, 1981), but it is needed to get good results in our case. In the case of a lacuna, we are not dealing with actual addition or removal but missing data in the other sequence. When a gap is cheaper than a substitution, there are two ways to arrange the gaps:

```

a a - - - - a a
- - b b b b - -

```

If there are many such gaps, the number of possible alignments with an equal score grows exponentially. To overcome the issue, we modified the global backtracking algorithm to return only a single alignment. The final scoring is presented in Table 1. To filter out uninteresting results, we ignore alignments scoring less than or equal to 100.

We used global alignment for the final results because local alignment resulted in too many irrelevant matches. Furthermore, we discovered that global alignment was better even in the case of fragments that contain excerpts from different texts and that therefore match different texts in the corpus: in

Type	Score
match	16
substitution	-5
common substitution	7
break match	6
break substitution	-10
gap start	-5
gap extension	-1
break gap extension	-10
variant	max of all combinations

Table 1: Scoring

these cases, the score of the global alignment, even with the penalty for the non-matching sections, was higher than that of the local alignments. An example alignment of the fragment K.17700 is shown in appendix D.

4 Evaluation

To evaluate the results, we picked 100 manuscripts with at most 20 lines from ‘Poem of Creation’ (*Enūma eliš*) (Appendix B). The size of less than 20 lines corresponds to the average size of a fragment in the collection of fragments put at our disposal by the eBL project.⁶ We ran a global alignment against all the manuscripts. A chapter matches if any of the manuscripts in the chapter has a score larger than 100. The method was able to assign the correct chapter to most of the inputs (recall 0.92), but produced many false positives (precision 0.39). The F_1 -measure is 0.55 and F_2 -measure is 0.73. The F_2 -measure is more suited to our use case because the goal is to identify as many fragments as possible, and we can tolerate false positives but must avoid false negatives. The identifications have to be validated by an expert, and they can filter out the false negatives.

The traditional approach to sign identification was slow and laborious: it often took weeks to identify a single fragment, and decades to catalogue collections thereof (see Section 1). Even so, the success was rather limited, and ca. 90% of the pieces could not be identified. The method proposed here speeds up the process dramatically, improves the success rate, and renders it possible to compare masses of transliterated fragments with large text corpora. As a first step we implemented the sign

⁶<https://www.ebl.lmu.de/fragmentarium>
The average size is 259,641 lines / 19,053 fragments = 13.62 lines.

mapping with ABZ numbers described above (3) and a simple transliteration search allowing us to query signs in consecutive lines. We made that system available to the eBL Team, who was able to identify hundreds of pieces that had escaped undetected with the traditional method (see Jiménez et al. 2018 and Jiménez et al. 2019). While that system was successful, it still required the user to construct queries and would find only exact matches, so it was not fully automatic. A simple possibility to achieve full automatization is querying systematically with the whole fragment, but that procedure does not yield interesting results. To find matching texts, domain knowledge is required to construct a clever query that would match divergent texts. A simple improvement could be to replace one sign at a time with a wildcard matching any sign in the target sequence, but the approach is unfeasible for large fragments as a query needs to be performed for each sign in the fragment, and what can match is also limited. Our string alignment approach allows for more drastic changes in the sequences, which do occur in practice, and, importantly, the whole process of searching for matches is fully automated.

5 Conclusion

We identify fragments of clay tablets written with cuneiform script. Mesopotamian literature can be reconstructed on the basis of fragments that match partially other known fragments. The detection of these partial matches is paramount for progress in knowledge. The traditional method for detecting these matches was slow and inefficient. In order to account for the ambiguities of cuneiform script, in particular for the polyvalence of the sign and for signs composed of other signs, we implemented a sign mapping that transforms a transliteration into a sequence of signs that correspond to the specific period in which the fragment was written. Then we implemented a string alignment algorithm in order to align those signs, and fine-tuned the scoring to adapt it to our needs, which included adding a series of common substitutions that should not be penalized. We used an affine gap penalty to account for gaps, an exceedingly common feature of fragments. The method described was able to assign 92% of the fragments to a correct context. The method will speed up considerably the task of assigning fragments to texts, and therefore accelerate the reconstruction of Mesopotamian literature.

358

References359
360

Rykle Borger. 1988. *Assyrisch-babylonische Zeichenliste*. Neukirchen-Vluyn.

361
362
363
364
365

Rykle Borger. 1991. Ein brief sin-idinnams von larsa an den sonnengott sowie bemerkungen über joins und das joinen. *Nachrichten der Akademie der Wissenschaften in Göttingen. I. Philologisch-historische Klasse*, (1991/2):39–84.

366
367

James Brusuelas. 2016. *Engaging Greek: Ancient Lives*, pages 187–204. Ubiquity Press.

368
369
370

Osamu Gotoh. 1982. [An improved algorithm for matching biological sequences](#). *Journal of Molecular Biology*, 162.

371
372
373
374
375

Enrique Jiménez, A. Heinrich, A. Häntinen, Zs. Földi, T. Mitto, J. Peterson, and U. Gabbay. 2019. From the electronic babylonian literature lab 8–15. *KASKAL. Rivista di storia, ambienti e culture del Vicino Oriente Antico*, (17):231–279.

376
377
378
379

Enrique Jiménez, A. Häntinen, Zs. Földi, A. Heinrich, and T. Mitto. 2018. From the electronic babylonian literature lab 1–7. *KASKAL. Rivista di storia, ambienti e culture del Vicino Oriente Antico*, (16):75–94.

380
381
382

Theodore Kwasman. 1998. A new join to the epic of gilgameš tablet i. *Nouvelles Assyriologiques Brèves et Utilitaires*, (1998/99).

383
384
385
386

Wilfred G. Lambert. 1992. *Catalogue of the Cuneiform Tablets in the Kouyunjik Collection of the British Museum: Third Supplement*. British Museum Press, London.

387
388
389

Johann-Mattis List. 2011. [Multiple sequence alignment in historical linguistics. a sound class based approach](#). pages 241–260.

390
391
392
393

Mozhgan Momtaz, Kayvan Bijari, Mostafa Salehi, and Hadi Veisi. 2016. Graph-based approach to text alignment for plagiarism detection in persian documents. volume 1737, pages 176–179. CEUR-WS.

394
395
396
397

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.

398
399
400
401

A. L. Oppenheim, E. Reiner, and M. Roth. 1956-2011. *The Assyrian Dictionary of the Oriental Institute of the University of Chicago*. Oriental Institute, Chicago.

402
403

Jelena Prokić, Martijn Wieling, and John Nerbonne. 2009. [Multiple sequence alignments in linguistics](#).

404
405
406
407
408

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. volume 1180, pages 1004–1011. CEUR-WS.

Avi Shmidman, Moshe Koppel, and Ely Porat. 2016. [Identification of parallel passages across a large hebrew/aramaic corpus](#). *CoRR*, abs/1602.08715. 409
410
411

T. F. Smith and M. S. Waterman. 1981. [Identification of common molecular subsequences](#). *Journal of Molecular Biology*, 147. 412
413
414

Margret Studt. 2007. [A practical and critical guide to the unicode blocks »cuneiform« and »cuneiform numbers« of unicode standard version 5.0](#). 415
416
417

Alex C. Williams. 2015. *Computationally Accelerated Papyrology*. Ph.D. thesis, Middle Tennessee State University, Murfreesboro, Tennessee. 418
419
420

Alex C. Williams, Hyrum D. Carroll, John F. Wallin, James Brusuelas, Lucy Fortson, Anne-Francoise Lamblin, and Haoyu Yu. 2014. [Identification of ancient greek papyrus fragments using genetic sequence alignment algorithms](#). In *2014 IEEE 10th International Conference on e-Science*, volume 2, pages 5–10. 421
422
423
424
425
426
427

428
429
430
431
432
433
434

A Development Fragments

Our development fragments in Table 2 were chosen because they are known to match one chapter (“fragment”) or more than one (“school”). The “fragments” contain significant gaps of the type that is common in cuneiform fragments. The non-matching fragments were chosen at random.

Museum number	Type
K.19352	Fragment
K.17700	Fragment
BM.36681	School
BM.36688	School
BM.99811	School
BM.101558	School
K.20949	School
K.17591	Fragment
K.18617	Fragment
K.19604	Fragment
K.20637	Fragment
K.21209	Fragment
Rm.468	Fragment
K.20074	not in Corpus, random
BM.110295	not in Corpus, random
BM.82855	not in Corpus, random
K.20703	not in Corpus, random, colophon
K.15836	not in Corpus, random, colophon

Table 2: Development fragments

B Test Fragments

To evaluate the results, we picked 100 manuscripts with at most 20 lines from ‘Poem of Creation’ (*Enūma eliš*) and compared it against all existing texts in a corpus of cuneiform literary texts. The size of less than 20 lines corresponds to the average size of a cuneiform fragment. The manuscripts chosen are:

- SB VII NinNAQuo2b
- SB I BabaNBQuo1
- SB V NinNACom1
- SB I BabaNBSch14
- SB I AššNA8
- SB I BabaNB9
- SB V NinNACom2
- SB VI NinNACom5
- SB V BabaNBSch2
- SB III BabaNBQuo1
- SB I SipNBSch1
- SB IV NinNA1b
- SB VI NinNA5
- SB VII SipNBQuo2
- SB II KalNA1
- SB VII BabaLBSch1
- SB I BabaNBSch12
- SB II NinNA1
- SB I AššNA4
- SB VI BabaNBCom2
- SB II BabaLBSch7
- SB I BabaNBSch4
- SB VII BabaNB1
- SB I BabaNBSch22
- SB VI BabaNBSch1
- SB II NinNA4

- SB IV BabaLBQuo3 470
- SB VII AššNACom1a 471
- SB VII NinNAQuo2a 472
- SB I BabaNBQuo3 473
- SB VI AššNA3 474
- SB I BabaNBSch7 475
- SB VI NinNACom3 476
- SB I BabaNBSch13 477
- SB I BabaNBSch5 478
- SB III HuzNAQuo1 479
- SB V BabaLBQuo1 480
- SB I NinNA1c 481
- SB VII NinNACom3 482
- SB VI BabaNBCom3 483
- SB V BabaLBQuo3 484
- SB IV BabaNBSch2 485
- SB IV BabaNBSch4 486
- SB VI BabaNBSch3 487
- SB V BabaNBCom3 488
- SB IV BabaNBSch1 489
- SB V NinNAQuo2 490
- SB III NinNACom1 491
- SB VI NinNAQuo2 492
- SB II BabaLBSch6 493
- SB IV HuzNA4a 494
- SB I BabaNBSch24 495
- SB I BabaNBSch15 496
- SB VII BabaNBQuo2 497
- SB IV HuzNA2 498
- SB I NinNA6b 499
- SB VII BabaNBSch4 500
- SB II BabaNBQuo2 501

502	• SB V BabaLBQuo2	• SB III BabaLBSch1	534
503	• SB VI BabaNBSch2	• SB VII BabaNBQuo1	535
504	• SB VII NinNA4c	• SB VII BabaLBQuo4	536
505	• SB V BorNBQuo1	• SB V AššNA1	537
506	• SB III AššNAQuo1	• SB VII NinNAQuo3	538
507	• SB VII BabaNBCom3	• SB IV NinNACom1	539
508	• SB I BabaNBSch8	• SB III BabaNBSch3	540
509	• SB V BabaNBSch6	• SB III BabaNBSch2	541
510	• SB I BabaNBSch11	• SB VII AššNAQuo2	542
511	• SB I BabaNBSch1	• SB VII HuzNA1c	543
512	• SB V NinNA6a		
513	• SB VII BabaLBQuo3		
514	• SB VII NinNAQuo4		
515	• SB VI SipLBSch1		
516	• SB I BabaNB7		
517	• SB II BabaNBSch1		
518	• SB IV AššMA1c		
519	• SB V NinNA3		
520	• SB I BabaNBSch6		
521	• SB III BabaNBSch1		
522	• SB I NinNA1d		
523	• SB I NinNACom1		
524	• SB VII NinNACom4		
525	• SB I AššNASch1		
526	• SB I AššNA7		
527	• SB VI KalNA1		
528	• SB VII AššNAQuo1		
529	• SB III BabaNBSch5		
530	• SB III BabaLBQuo1		
531	• SB IV BabaNBCom2		
532	• SB VI BabaLBQuo2		
533	• SB II BabaLBSch3		

544
545
546
547
548
549
550
551
552
553
554

C Substitutions

Although there are usually more than two ways of writing a word in cuneiform, some sign substitutions are particularly common (e.g. TU = ABZ58 vs TI = ABZ73 in *le-mut-tu* vs *le-mut-ti*). In order to determine the most common substitutions, we performed local alignments, filtered out alignments with similarity less than 80%, and calculated the frequency of each pair. These common substitutions have a positive but smaller score than the exact match. The list is shown in table 3.

ABZ58	ABZ139
ABZ75	ABZ231
ABZ142	ABZ579
ABZ70	ABZ75
ABZ55	ABZ59
ABZ308	ABZ142
ABZ537	ABZ55
ABZ73	ABZ207
ABZ75	ABZ312
ABZ367	ABZ449
ABZ214	ABZ371
ABZ84	ABZ586
ABZ73	ABZ139
ABZ86	ABZ328
ABZ352	ABZ138
ABZ597	ABZ353
ABZ104	ABZ7
ABZ376	ABZ73
ABZ61	ABZ427
ABZ207	ABZ139
ABZ532	ABZ427
ABZ5	ABZ371
ABZ318	ABZ411
ABZ545	ABZ354
ABZ58	ABZ73
ABZ381	ABZ207
ABZ461	ABZ536
ABZ68	ABZ86
ABZ342	ABZ61
ABZ318	ABZ142
ABZ597	ABZ545
ABZ68	ABZ328

Table 3: Common substitutions

D Example Alignment

The global alignment of the fragment K.17700 against the manuscript SB I NinNA3 from 'Poem of Creation' (*Enūma eliš*) has significant gaps due to parts broken away from the fragment. An affine gap penalty with a low cost for gap extension allows for significant gaps. 88 % of the signs in the preserved regions are identical. Line-break (#) matches have a high score so that the lines will stay aligned.

```
K.17700      # - - - - - - - #
NinNA3       # ABZ308 ABZ69 ABZ61 ABZ449 ABZ61 ABZ295 ABZ152 #

K.17700      - - - - - - -
NinNA3       ABZ13 ABZ437 ABZ381 ABZ99 ABZ231 ABZ532 ABZ461

K.17700      - - - - # - - -
NinNA3       ABZ13 ABZ374 ABZ139 ABZ565 # ABZ308 ABZ69 ABZ61

K.17700      - - ABZ295 - # - - -
NinNA3       ABZ449 ABZ61 ABZ295 ABZ152 # ABZ461 ABZ342 ABZ381

K.17700      - - ABZ396 ABZ308 - - - #
NinNA3       ABZ427 ABZ532 ABZ396 ABZ308 ABZ70 ABZ61 ABZ318 #

K.17700      - - - - - - - -
NinNA3       ABZ449 ABZ537 ABZ461 ABZ142 ABZ342 ABZ231 ABZ73 ABZ367

K.17700      ABZ86 ABZ308 ABZ73 - - # - - -
NinNA3       ABZ86 ABZ308 ABZ73 ABZ332 ABZ545 # ABZ367 ABZ78 ABZ6

K.17700      - - - - ABZ401 ABZ579
NinNA3       ABZ211 ABZ545 ABZ55 ABZ342 ABZ401 ABZ579
```