

'Measuring Similarities and Visualising Patterns in a Text Tradition using Pairwise Sequence Alignment and t-Distributed Stochastic Neighbour Embedding'

Estara J Arrant

Jeroen Verrijssen

Appendix: Misalignment Typology

Figure 1: Mathematically Insignificant Misalignment due to Small Word or Letter Insertions

<DirEntry 't1610i_Ex15.16.txt'> <DirEntry 't1652i_Ex15.16.txt'>

תפ-לעליהו לא ימתאדמותא ותברא בתקופא דרעגברו-תא יתקו לן כא
 |||||
 תפ-לעליהו לא ימתא-ות-ברא בתקופא דרעגבר-ורתד-ישתקו לן כא

בנ-יא: עדזמך דיעברו עמך יי יתנחזלי ארנו נא-עדזמך דיעברו לן
 |||||
 בנ-יא: עדזמך דיעברו עמך יי יתנחזלי ארנו נא: עדזמך דיעברו לן

מדהא לי יודקנ יתא יתמג יזתי הדי בקא
 |||||
 מדהא לי יזתי-ית-מג יזתי הדי בקא

90

0.6923076923076923

Here, an alignment between t1610i and t1652i at Ex. 15:16, the misalignments occur at ות and at ית because of inserted words. If they were corrected, the score would be unchanged because the number of matches to mismatches would be unchanged.

Figure 2: A Mathematically Significant Alignment

<DirEntry 't1629a_Ex15.15.txt'> <DirEntry 't1681a_Ex15.15.txt'>

בכ-יא יתבה לי ורברבנ יאדומא יתקי פי ימואבא-יאחדתנו לרת יתא: א
 |||||
 בכ-וא-יתבה לי ורברבנ יאדומא יתקי פי-אחדת ית-תהו לרת יתא: א

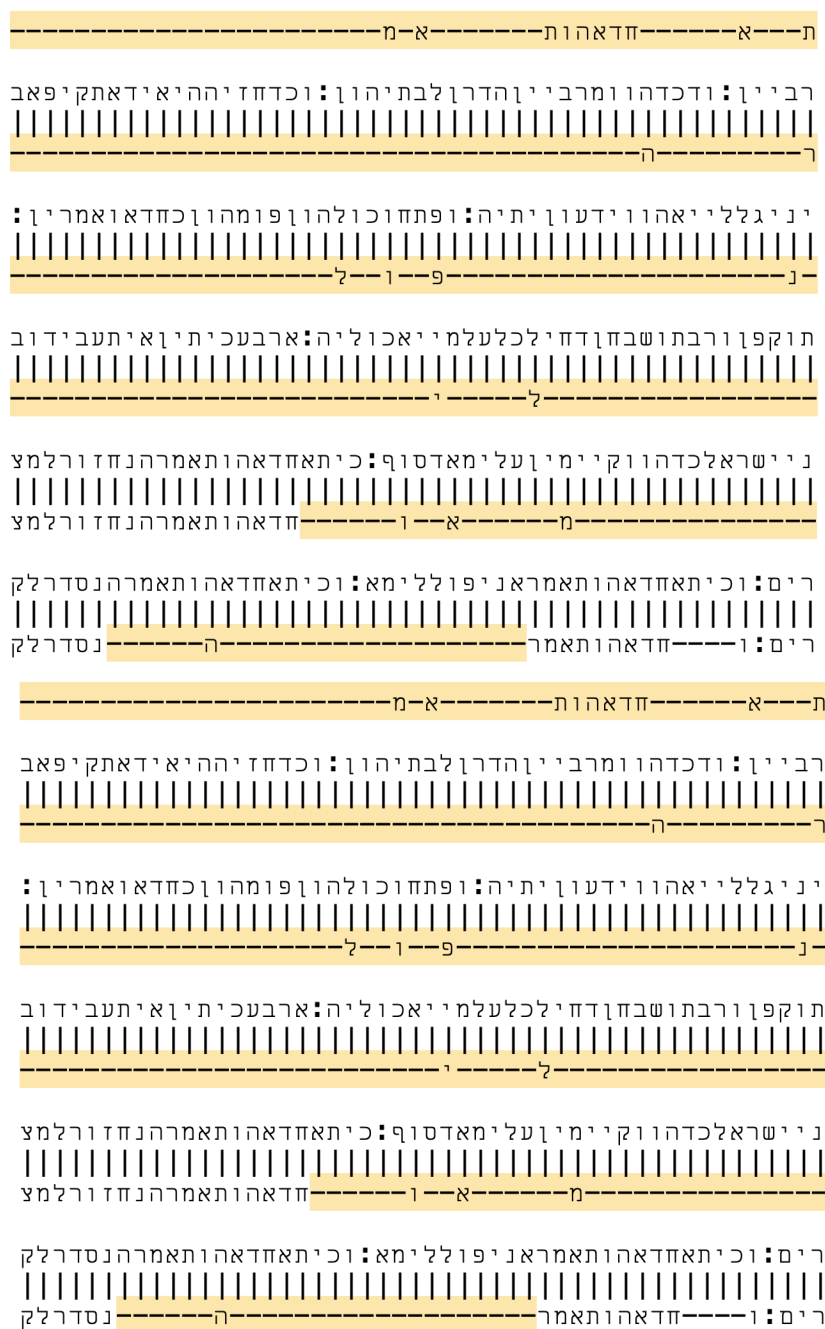
יתברו לי בהו לבג-ופיהו לן-כלדהו ויתבי לך ייריא רעאדכנען
 |||||
 יתברו ל-בהו-לבג-יהו ללכלד-י-י-ר-יא רעאדכנען

45

0.4639175257731959

Here, the misalignment occurs in t1681a's word אחדת, which has been aligned (with false mismatches) to ימואבא and not to the correct correspondent, אחדתנוך. These false mismatches artificially lower the alignment score.

Figures 3 and 4: Passage Insertions



Between t1610i's *Arba' Kitin* passage occurs in a slightly different location than t1647i's, due to shorter *tosefta* material. The algorithm could not find the correct portion of the text in t1647i, which occurs at the end of the second to last line here.

Figure 5: Gapping to Create Matches

שראליתמצראימייתייןרמאיןעלכיקימא: וחמון ישראליתגבורתיד
 שראליתמצראימייתיין-----עלכיקימא:
 אתקיפאדעבדהיניסייןבמצרים: ודחילעמאמןקדסייןוהאמינו
 הכ-----ד-----י-----ן
 בשוםמיראדייובנביאותיהדמשהעבדיה: האבכןשבחמשהובנייש
 ש-----ב-----חמשהובנייש

t1685a's הכדון does not occur in t1652 (top sequence). The algorithm has sought to find matches, incorrectly, for this word.

Figure 6: A Noisy yet Valid Alignment

שראליתשבחשירתאהד-אקדסייןואמרייןלמימרנודהונושבחקדסיין
 שראליתשבחשירתא-דהאקדסייידאמרו-למימרנודהונושבחקדסיין
 רמאדמתגאי-עלביותננאומתנטל-עלמנטל-יא: ו-כלמ-ןדמתגאה
 דהו-ארסעלרמ----יאומתגוניהעלגיותנניאאריכלמאןדמיגאה
 קדמיההואבמימריהמתפרעמיניהסוסווןורוכביהוןעלדא-תגאו
 קדמוהי-במימריה-פרעמיניהמצראיסוסות-הוןעלדאיתגאו

Between t1691a and Hebr313 at this location (green) there are legitimate differences in wording which cause some false matches. However, the negatives (gaps and mismatches) outweigh these false matches and the alignment reflects the divergence between the texts.