Translation Arrays – alignment decisions

Version	1.1
Date released	28/02/2012
Author	Kevin Flanagan

Version history

Version	Date released	Author	Comments
1.1	28/02/2012	Kevin Flanagan	Overlap proposal changed, ‘null’ alignment described under ‘Missing alignment’, one-to-many and many-to-one added.
1.0	27/02/2012	Kevin Flanagan	First draft

Contents

Introduction

This document discusses the alignment of multiple-version-translation corpus data, and the decisions that are required in that regard prior to implementing any corpus management system that can generate and visualise the desired version-variation statistics. It follows on from some of the issues raised in Tom Cheesman's "Translation Sorting: Eddy and Viv in Translation Arrays" paper, and proposes an alignment system that addresses those issues, described in sufficient detail to allow software development to begin.

Prerequisites

Readers should be familiar with the details and terminology of the 'Translation Sorting' paper.

Terminology

'Translation Sorting' refers to comparing variation both at 'full text' and at 'text component' level, where the latter term is understood as meaning any section of the text, rather than the whole. In this document, the term 'segment' shall be used - in preference to 'text component' - to mean a section of the text with a defined start and end point, of arbitrary size but smaller than the full text. 'Corpus' shall refer to a collection of texts comprising a single source text (ST) in a source language (SL) and a number of target texts (TT) that are translations of the ST, all into the same target language (TL).

Scope

'Translation Sorting' is primarily concerned with comparing translations of Shakespeare, but the corpus management system should be designed to be suitable for other types of text.

Imaginary corpus

Examples in this document are based on an imaginary corpus where the ST consists of some arbitrary lines from Othello, divided into segments in various equally arbitrary ways. Lower case letters are used as ST-segment identifiers, but as will be seen, this is not meant to imply that segmentation must be applied to the ST by dividing it into a simple sequence of consecutive segments.

Purpose of alignment and example cases

Recall that, per 'Translation Sorting', a Viv value is calculated for each ST segment, and a custom user interface then displays the ST in such a way as to provide a visual indication of the varying Viv values for the different segments. The 'ideal case' for Viv value calculation can be defined as follows:

> **Viv value calculation 'ideal case'**: the ST segment has been aligned with a corresponding segment from each TT in the corpus. Eddy values are then calculated for each TT segment. The Viv value is calculated using those Eddy values.

A number of questions can immediately be asked about corpus alignment, including:

- How is the ST to be divided into segments? (Based on one or more grammatical definitions, such as one-segment-per-sentence?)

- Can ST segments overlap? (part of one segment shared with part of another segment)
- Can ST segments contain other segments? (all of one segment contained within another segment)
- How is the Viv value calculated for an ST segment if, for some of the TTs, there is no alignment to a TT segment?
- If ST segments can overlap or contain other segments, how are covered or partially-covered segments presented to the user or otherwise made available for inspection?

The further discussion and example cases below are used to illustrate some of these questions so as to propose answers. In the example cases, the division of the ST into segments is completely arbitrary and meant only for illustration purposes.

Simple case – corpus C_1

In this example, for a corpus {ST, TT_1, TT_2}, Viv values can be calculated per the 'ideal case' defined above, and other questions such as overlap etc. do not arise.

	ST	alignment	TT_1
a	Most potent, grave, and reverend signiors,		Nlhg klgvmg, tizev, zmw ivevivmw hrtmrlih,
b	My very noble and approved good masters,		Nb evib mlyov zmw zkkilevw tllw nzhgvih,
c	That I have ta'en away this old man's daughter,		Gszg R szev gz'vm zdzb gsrh low nzm'h wzftsgvi,
d	It is most true; true, I have married her		Rg rh nlhg gifv; gifv, R szev nziirvw svi

	ST	alignment	TT_2
a	Most potent, grave, and reverend signiors,		Atnq ojfalq, vqlri, itp vuncalco pailltdt,
b	My very noble and approved good masters,		Gr gvqc ffrjn ccq dhqrrsot tdtb gvltgbd,
c	That I have ta'en away this old man's daughter,		Joys O balo ic'io ncoi jcjo glo fwq'o dtgomogs,
d	It is most true; true, I have married her		Gc qn bdrn qrco; fbsi, S mstk okngcgh nrm

Missing alignment – corpus C_2

Consider corpus C_2 consisting of C_1 plus the following TT_3 and alignment:

	ST	alignment	TT_3
a	Most potent, grave, and reverend signiors,		Prvw srwhqw, judyh, dqg uhyhuhqg vljqlruv,
b	My very noble and approved good masters,		Pb yhub qreoh dqg dssuryhg jrrg pdvwhuv,
c	That I have ta'en away this old man's daughter,		
d	It is most true; true, I have married her		Lw lv prvw wuxh; wuxh, L kdyh pduulhg khu

In this corpus, there is no alignment between ST_b and any segment in TT_3. Does this affect how the Viv value should be calculated for that segment? Bearing in mind that the provisional formula for Viv is $\bar{x}$ Eddy / SN , the Viv value should not be overly skewed by the missing alignment – but the user should be aware that for the corpus under consideration, the Viv value being visualised does not represent all target texts in the corpus.

> **Proposal 1**: for 'missing' alignments, Viv values should be calculated in the same way, but the user interface should indicate to the user that one or more alignments were missing (e.g. some glyph next to the segment that can be clicked to show more information)

Note: in a case such as the above, a capability is required to add a 'null' alignment – to indicate that alignment has not been accidentally omitted, and against which annotations might be held – between ST_b and a particular point (as opposed to segment) in TT_3. (The user interface can by default prompt that the point immediately follows the TT segment aligned with ST_a.) This 'null' alignment is not used when calculating Viv values. The same capability is required to null-align segments of a TT that do not correspond to any ST segment.

What if there are no non-null alignments in the corpus at all between ST_b and any segment in any TT, or – depending on how segmentation is represented in the data – if there is ST text between segments, for which no segment has been defined, never mind aligned? Defining both of these as cases of 'unaligned' text, no Viv value should be calculated.

> **Proposal 2**: no Viv values should be calculated for 'unaligned' text, nor any visualisation formatting added for the text (though 'containment' cases should take account of unaligned text as described below)

Aggregation

'Translation Sorting' describes how Eddy results can be 'aggregated up' by combining the results from individual (say) sentences to create higher-level distinctiveness values for speeches, characters, scenes, etc. This description of aggregation implies two distinct but related capabilities for the corpus management system.

The user interface mock-up shown at the start of the paper implied a capability of switching from 'sentence-by-sentence view' to another view. We can imagine a 'sentence-by-sentence view' where

(segmentation and alignment permitting) each sentence is given separate formatting to reflect its own Viv value, as well as higher-level views such as 'speech' view, where the Viv value is calculated for the complete speech (and so the entire speech is formatted uniformly and in accordance with that value), or indeed lower-level views such as 'word' view, in cases where particular words of interest have been segmented and aligned. In this document, these views shall be described as providing different levels of 'display granularity' within the user interface. (It is quite feasible to allow users to make the display granularity be non-uniform, that is, have it set to 'sentence level' for most of the ST, but select a more coarse-grained display for a certain few lines or scene, so as to allow more alignments to be taken into account, such as the 'Containment' example below.)

Aggregation to a level of 'character' is, on the face of it, a separate capability. Rather than changing the display granularity, it is instead selecting which ST segments are of interest (those spoken by a particular character), and calculating a Viv value for all of them – so the level of granularity is the entire text, and all of the segments uttered by that character are formatted for visualisation in the same way. Parallel selections (so long as they are mutually exclusive) could so be visualised, so that separate Viv values are calculated for character X and character Y, with accordingly different visualisation formatting. Applying this kind of operation to ST display shall be described in this document as changing the 'selectivity' of the user interface.

These capabilities are orthogonal and can be used together. Changing the display granularity to 'scene level' and the selectivity to 'per character' would result in a display where, for each scene, Viv values are calculated for each character, and so within that scene, all the segments uttered by a given character have the same visualisation formatting.

Containment – corpus C_3

Consider corpus C_3 consisting of C_2 plus the following TT_4 and alignment:

	ST	alignment	TT_4
e	Most potent, grave, and reverend signiors, My very noble and approved good masters,		Suyz vuzktz, mxgbk, gtj xkbkxktj yomtouxy, Se bkxe tuhrk gtj gvvxubkj muuj sgyzkxy,
c	That I have ta'en away this old man's daughter,		Zngz O ngbk zg'kt gcge znoy urj sgt'y jgamnzkx,
d	It is most true; true, I have married her		Oz oy suyz zxak; zxak, O ngbk sgxxokj nkx

In this corpus, the translation of the ST text that above was segmented into ST_a and ST_b – and remains so aligned with TT_1, TT_2 and TT_3 in the corpus – corresponds to text in TT_4 that can't coherently be split into two parts, because of reordering or entangling of the thematic structure, etc. The segment ST_e aligned with TT_4 'contains' ST_a and ST_b. What does this mean for calculation and visualisation of Viv values? The answer must depend on the display granularity. If it is fine-grained enough to be showing separate Viv visualisation formatting for ST_a and ST_b, then the Eddy value for the alignment of ST_e above cannot be incorporated into the visualisation, and should be treated in a similar way to the 'missing' alignment case shown above. If display granularity is at a level where ST_a and ST_b are being aggregated into a unit which contains ST_e, then the ST_e alignment information can be used within the Viv value calculation.

Proposal 3: where alignments exist that are too coarse-grained to be used for Viv value calculation at the display granularity level in use, they should be omitted from Viv value calculation, and the user interface should indicate that an alignment had to be ignored (much as with Proposal 1 above)

Containment #2 – corpus C_4

Consider corpus C_4 consisting of C_3 plus the following TT_5 and alignment:

	ST	alignment	TT_5
a	Most potent, grave, and reverend signiors,		Xzde azepye, rclgp, lyo cpgpcpyo dtrytzcd,
b	My very		Xj gpcj yzmwp
	b2 noble		Lyo
	and approved good masters,		laaczgpo rzzo xldepcd,
c	That I have ta'en away this old man's daughter,		Esle T slgp el'py lhlj estd zwo xly'd olfrsepc,
d	It is most true; true, I have married her		Te td xzde ecfp; ecfp, T slgp xlcctpo spc

In this corpus – though this representation isn't ideal – the ST_b segment "My very noble and approved good masters" has been aligned with "Xj gpcj yzmwp lyo laaczgpo rzzo xldepcd". However, within that segment, an individual word alignment has also been added, between "noble" and "lyo". What does this mean for calculation and visualisation of Viv values? The answer again must depend on the display granularity. If it is at a level where the Viv value for ST_b is being visualised, then the value for ST_{b2} is redundant information. It might be possible to overlay the value for ST_{b2} in some version of the user interface, though that could risk being confusing, and is arguably unnecessary for a first version.

Proposal 4: when calculating the Viv value for a segment, any contained segments aligned to the same TT can be ignored.

Containment #3 – corpus C_5

Consider corpus C_5 consisting of C_4 plus the following TT_6 and alignment:

	ST	alignment	TT_6
a	Most potent, grave, and reverend signiors,		Nptu lpudos, mwfaa, boc ubzauboc thhmotst,
	My very noble and approved		Ow zatw
b2	good masters,		opama boc
			elmulyae imti lbtubut,
c	That I have ta'en away this old man's daughter,		Ufcy O fcza qd'cp bvdv siht pmb ocl't ceqhiudt,
d	It is most true; true, I have married her		Gy or nntz zsqd; usqd, K fcza lbxxkdc ibu

In this corpus, the ST text segmented above as ST_b has not been aligned – but part of it has been aligned as ST_{b2}. What does this mean for calculation and visualisation of Viv values? For instance, when the display granularity is set so that the Viv value for ST_b is to be visualised, how can the ST_{b2} alignment information be incorporated? One way would be to adjust the formula for Viv. With corpus C_4, the calculation could be written out as:

$$Viv(ST_b) = (Eddy(ST_b - TT_1)/7 + Eddy(ST_b - TT_2)/7 + Eddy(ST_b - TT_5)/7) / 3$$

(Note that '7' is the number of words in the ST segment, and that TT_3 and TT_4 aren't taken into account at this level of display granularity for the reasons explained above.)

To compute a value that might be used in C_5 as Eddy(ST_b –TT_6), we could assume that all the words in ST_b that are not in ST_{b2} have the most neutral tf value, that is, 1. That would allow us to write out the calculation for Eddy(ST_b –TT_6) like this:

$$Eddy(ST_b – TT_6) = (D/1) * 5 + Eddy(ST_{b2} – TT_6)$$

(where per 'Translation Sorting', D is the number of translations, divided by the presumed tf of 1, multiplied by 5 for the 5 unaligned words.)

In turn, for corpus C_5, the calculation for the Viv value for ST_b could be written out as:

$$Viv(ST_b) =$$
$$(Eddy(ST_b – TT_1)/7 + Eddy(ST_b – TT_2)/7 + Eddy(ST_b – TT_5)/7 + ((D/1) * 5 + Eddy(ST_{b2} – TT_6))/7) / 4$$

In other words, we could choose to use alignment information for a segment even when only part of the segment is aligned, rather than ignore the information altogether. As with 'missing' alignments, the user should be made aware that the 'neutral' presumption for the unaligned text has been made.

> **Proposal 5**: when calculating the Viv value for a segment, if there is no single aligment between that segment and a given TT, and if there are contained alignments that don't provide complete coverage for the segment, the Eddy value to be used for that TT should be calculated using the contained alignment Eddy values and a 'neutral' weighting for the rest of the segment

Overlap – corpus C_6

Consider corpus C_6 consisting of C_5 plus the following TT_7 and alignment:

	ST	alignment	TT_7
a	Most potent, grave, and reverend signiors,		Qswx tsxirx, kvezi, erh vizivirh wmkrmsvw,
b	My very noble and approved good masters,		Qc zivc rsfpi erh ettvszih kssh qewxivw,
f	That I have ta'en away this old man's daughter, It is most true;		Xlex M lezi xe'ir eaec xlmw sph qer'w heyklxiv, Mx mw qswx xvyi;
g	true, I have married her		xvyi, M lezi qevvmih liv

In this corpus, the alignment uses a different segmentation for the text that above is segmented as ST_c and ST_d. What does this mean for calculation and visualisation of Viv values? When the display granularity is such that the Viv values for ST_c and ST_d are being visualised, there's no clear way of making use of Eddy values for TTs that are aligned with ST_f and ST_g. In this case, the simplest approach is again to ignore those values and make the user aware that they have not been used. (Per the note about non-uniform display granularity under 'Aggregation above, it would be possible to allow the user to switch to visualising Viv values for ST_f and ST_g, so that Viv values for ST_c and ST_d are then ignored instead.) When display granularity is more coarse, so that (say) the single Viv value for the span of text encompassing ST_a - ST_d is being visualised (and therefore also ST_a and ST_g), the overlapping alignment is not an issue, and all the alignment information can be used.

It's possible that often an alignment like that shown above could be reformulated like this:

	ST	alignment	TT_7
a	Most potent, grave, and reverend signiors,		Qswx tsxirx, kvezi, erh vizivirh wmkrmsvw,
b	My very noble and approved good masters,		Qc zivc rsfpi erh ettvszih kssh qewxivw,
f	That I have ta'en away		Xlex M lezi xe'ir eaec
f1	this old man's daughter,		xlmw sph qer'w heyklxiv,
f2	It is most true;		Mx mw qswx xvyi;
g	true, I have married her		xvyi, M lezi qevvmih liv

In this case, ST_f and ST_{f1} are both contained by ST_c, while ST_{f2} and ST_g are both contained by ST_d, so the containment visualisation rules can be coherently applied. Where overlaps occur, it could be worth prompting users to try to use this kind of reformulated alignment.

> **Proposal 5**: when displaying Viv values for a segment, ignore Eddy values for overlapping (N.B. – as opposed to 'contained') ST segmentation, and encourage use of containment as a workaround where required.

What about overlapping TT segmentation? It may be considered questionable – we could certainly enforce a rule that disallows it – but it doesn't present any problem for Viv visualisation. However, when 'zooming in' on a particular TT, it could make Eddy value visualisation difficult.

One-to-many – corpus C_7

Consider corpus C_7 consisting of C_6 plus the following TT_8 and alignment:

	ST	alignment	TT_8
a	Most potent, grave, and reverend signiors,		Lnrt ontamt, cqgva, gmf qavaqamf rhcmhnqr,
b	My very noble and approved good masters,		Ly vaqy mndka gmf gooqnvaf cnnf lgrtaqr,
c	That I have ta'en away this old man's daughter,		Tegt H egva tg'am gwgy tehr nkf lgm'r fgucetaq,
d	It is most true; true, I have married her		Ht hr lnrt tqua; tqua, H egva lgqqhaf eaq

In this corpus, the same ST segmentation has been used to align it with TT_8 as was used to align it with TT_1, TT_2 and TT_3. However, the TT text that corresponds to ST_a is split, as is also the case for ST_b. It's possible that a more fine-grained ST segmentation could be used – that is, one that splits ST_a into two segments, each aligned with a single TT segment – and in such a case, the containment rules above would make that alignment just as useable for visualisation purposes. However, on some occasions that might not be possible for linguistic reasons, or the user simply may not have time to indicate what the more fine-grained ST segmentation should be. Implementing one-to-many alignments such as these covers both cases, and doesn't present any new issues for visualising ST segment Viv values.

Many-to-one – corpus C_8

Consider corpus C_8 consisting of C_7 plus the following TT_9 and alignment:

	ST	alignment	TT_9
h	Most potent, grave,		Cemp jephdp, olysh, ydr lhshlhdr muoduelm,
i	and reverend signiors,		
j	My very noble and		Cx shlx detbh ydr yjjleshr oeer cymphlm,
k	approved good masters,		
c	That I have ta'en away this old man's daughter,		Pfyp U fysh py'hd yvyx pfum ebr cyd'm ryqofphl,
d	It is most true; true, I have married her		Up um cemp plqh; plqh, U fysh cylluhr fh

In this corpus, a different ST segmentation has been used for alignment with TT_9, where the reverse technique has been used – multiple ST segments aligned to a single TT segment. What does this mean for calculation and visualisation of Viv values? When the display granularity is such that the Viv values for ST_a and ST_b are being visualised, can the alignment information for segments ST_h - ST_k against TT_9 be used? ST_h and ST_j will share a single Eddy value computed against TT_9, but if it is (say) an especially high value, there's no way of knowing whether that is because of a high level of variation in ST_h or in ST_j, or both.

> **Proposal 6**: when displaying Viv values for a segment, ignore Eddy values for many-to-one alignments that are not completely contained by the segment.

Coarse-grained selections and 'missing' alignments

Consider corpus C_2 as described above, consisting of {ST, TT_1, TT_2, TT_3}. When the display granularity is set so the separate Viv values for ST_a, ST_b, ST_c and ST_d are being visualised, it was proposed that missing alignment ST_b-TT_3 be simply ignored. What about more coarse-grained visualisation? Suppose a single Viv value for the four segments ST_a, ST_b, ST_c and ST_d were being visualised (the 'full text' in the case of this simple example corpus) – how should that Viv value be calculated, and what account, if any, should be taken of 'missing' alignments?

A Viv value for the full text can be obtained by calculating the average of the Viv values for the (contiguous, in this case) segments it contains. Since, as proposed above, the missing ST_b alignment is not considered to have skewed the calculation, neither does it skew the average Viv value for the full text. When calculating a Viv value for the full text (or other large section of text) it is obviously desirable not to exclude alignment information from any TT. Per the first 'containment' example above, when calculating/displaying the Viv value for relatively fine-grained segments, the Viv values for 'outer' segments (aligned to other TTs) that contain them have to be ignored. Conversely, per other 'containment' examples, when calculating/displaying the Viv values for relatively course-grained segments, the Viv values for 'contained' segments (aligned to other TTs) can be factored-in to the result, even if those finer-grained 'contained' segments do not provide complete coverage of the 'container' segment.

> **Proposal 6**: when visualising the Viv value for a section of text, the average Viv value for the segments therein should be used, regardless of 'missing' alignments. Calculation of the average should use the most coarse-grained segments (applying 'containment' rules to the others) so as not to exclude alignment information.

Unanswered questions

This document has been concerned primarily with corpus alignment and the implications for variation visualisation. Questions from 'Translation sorting' that it does not address – but which can be considered orthogonal to issues of alignment – include:

- How to represent other kinds of variation, e.g. speech assigned to a different character
- How to deal with 'noise' in the data such as spelling variations, inflections, etc.
- How to incorporate POS information to better represent variation
- How to adjust the Eddy formula when (say) 26 ST words are aligned to only 8 TT words

Nevertheless, if the alignment proposals in this document are considered acceptable, or as-and-when alternative proposals with sufficient clarity have been agreed, then implementation of a suitable corpus management system can begin.