The discussion about language support so far

This have not become minutes of the last meeting but a collection of the things that where discussed and that are relevant to continue discussing this subject.

Before march 5th

The discussion, about language support, started when issue 22 "Allow shape element usage" was discussed a while ago.

Use cases for issue 22 are written down here:
http://altoxml.github.io/alto_shape_use_cases/ALTO_shape_usecases.html
A number of cases are described where polygons are needed to describe strings, text lines and text blocks.
The use cases "Skewed text 1" & "Skewed text 2" where the text is rotated on the page raise a question about the meaning of the attribute "rotation". Does rotation describe the rotation of the block itself or the text within the block? Do we have a rotated rectangle or a polygon with rotated text?

Frederick suggested that there could be a relation between this issue and language support issues that we might need for Asian languages. It's also interesting because the ALTO Board is still searching more Asian language people in digitization and projects they are working on.
At the end of this meeting we agreed on looking at use cases for other languages where text direction might be an issue and to look at other standards.

Discussion of March 5

Input for the discussion is a document with examples from Siang Hock.

Jean Philippe explains a little about the what he found looking at different standards that have language support. Some details:

- In EPUB, since it is a complete publication, has functionality for reading order which is the order in which the different text blocks must be read. In ALTO we regard reading order as something that is recorded on document level in for example METS.
- CSS has writing modes that addresses text direction characteristics. OCR can probably not detect text direction by itself so it must be given a text direction up front, per page or per textblock if there is variation in text direction on a page.

Jean Philippe says the BnF is now creating ePub from ALTO XML files. It's difficult because the accuracy of the OCR often requires lots of manual corrections. Also because ALTO XML files are page-based; ePub is not page-based.

Discussion of April 8

Input for the discussion is a table in which you can compare text direction characteristics (block, link, string level) and the way these can be described in these standards. It's meant to explore these characteristics in the context of ALTO.

In this document the need for language support is described as follows:

"It's necessary to record enough text characteristics of a physical page

1. to be able to reproduce a representation of the physical page from the ALTO.

2. and to be able to process the text from the ALTO to create an index or transform the text to a file with another format, for example EPUB. How is for example hebrew recorded in ALTO? Is a word that must be red right to left be recorded as "hebrew" or "werbeh" in ALTO? In the last case ("werbeh") creating the representation is not a problem. Is that different from the EPUB?"

(https://github.com/altoxml/schema/blob/master/v3/Comparison%20of%20text%20direction%20elements.pdf)

Jean Philippe; The text direction element in CSS is language oriented and not layout oriented. When you describe text direction as layout there might more variation, for example bottom to top which the language has not.

Joachim questions the relevance of adding extra language support in ALTO, he explains that when ALTO is produced with segmentation and OCR software, this software is given the language and is using the characteristics of this language like text direction to interpret the page. Therefore things like text direction are already implicitly available in the ALTO in the attribute LANG. The LANG attribute is available for the elements String, TextLine and TextBlock.

Joachim worries that by adding more text direction logic, the ALTO will become too complex and that the extra functionality will never be used because it's usually too expensive to collect this information. 99% of the ALTO is generated automatically with rules and not by hand.

The whole reading sequence can not be done in ALTO, it's about the document and not just a page. Reading sequence is not layout based. TEI tries to combine physical (layout) and logical structure in one and because of that, the files become really big.

As the purpose of ALTO is to describe the layout and content of a page the elements LINE and SPACE have a function beyond layout and content. The purpose of LINE and SPACE is also to be able to extract the text from the ALTO easily in the right sequence. SPACE only has a meaning if there is a known reading order in the LINE. Without LINE the block would just contain a list of words/strings with no reading sequence.

Reading order was also discussed in an issue that came from Impact. In that issue reading order ment the sequence of the blocks and it was decided not to support this in ALTO. Here we are talking about reading order in the block.

Jean Philippe suggest that we might not need to change the ALTO if it can be used for Asian languages. Maybe we just need to make guidelines for handling non western language material.

Jean Philippe describes the use case for which text direction and reading order are relevant. When you want to use the ALTO to produce an EPUB or text to speech functionality, you need to be able to export the text in the correct order. Pages with mixed languages and/or mixed text directions are the most complex.
One language can have different text directions on a page. So if you have a page with different languages like Arabic and English do you need to record text direction on a more detailed level in order to get a useful ALTO?

Joachim explains the work that they have done for a project where material in Hebrew was digitized. In the project he describes, the direction of the Hebrew in ALTO is left to right (on a string and/or line level?) where it was right to left on the actual page. This does not cause problems or misinterpretations in this project. He promises to provide some examples.

Joachim describes possibilities for recording text direction where the text with a different reading sequence must be in a separate block. For me (Evelien) it was not possible to reproduce what was suggested.

Discussion of may 6

Jean Philippe has written a document which hold many examples of pages with variation in text direction. https://github.com/altoxml/schema/blob/master/v3/discussion%20of%20ALTO%20language%20support.pdf.

The discussion was postponed to the next meeting.