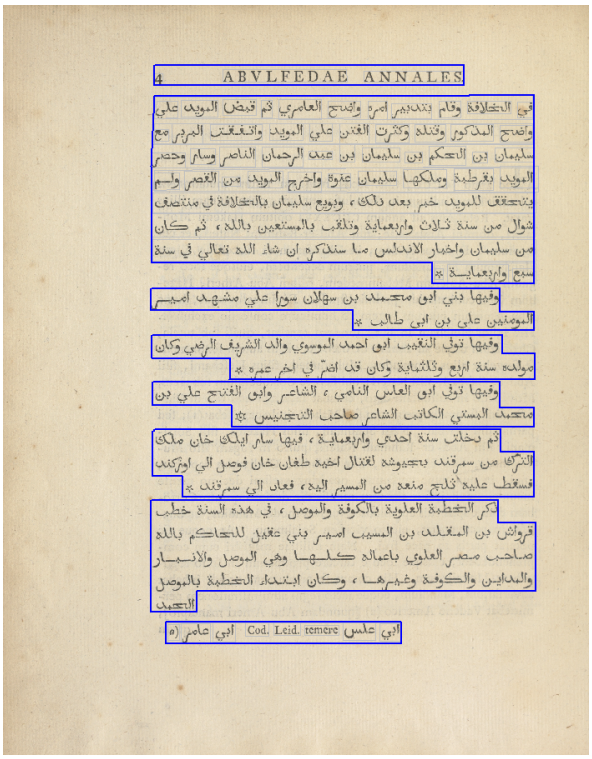
ALTO Languages support

1. Arabic sample (with latin secondary language)

http://gallica.bnf.fr/ark:/12148/bpt6k62219939/f14.image

1789-1794, monography



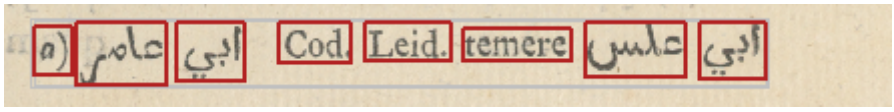
Language and script

Should be defined at the top level (ALTO root) or at element (block, line, word) level.

Language: arab
Script: naskh

Notes:

1. Most of the time, setting the language at the top level is enough. But some languages use mixed layout modes (Chinese, Japanese), and some documents are multilingual. Eg: this foot note has arab and latin words.



2. Language and script are 2 different things. Eg. Turkish language can be written with latin or arabic scripts.

Use cases: OCR correction, automatic language processing, accessibility, etc.

Block flow direction

Block flow direction is defined by the logical blocks reading order. This can be implemented in ALTO with:

1. order of XML elements within the file,
2. IDNEXT attribute set on Block element,
3. a top level attribute could also document the block flow direction. Here, top-to-bottom (horizontal lines of text).

Note : for use and reuse of ALTO contents, 1. and 2. are easier to handle than 3. alone (3. implies a topological analysis of the ALTO elements).

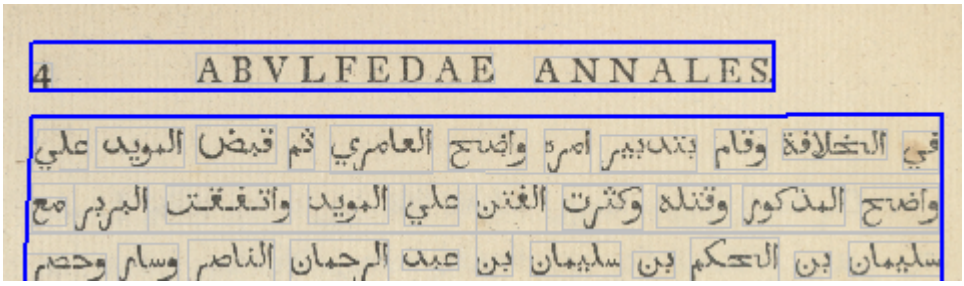
Reading direction

Reading direction within a line of text can be known :

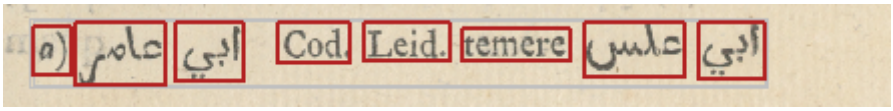
1. with XML elements order within the ALTO file
2. with an IDNEXT attribute on String
3. with a dedicated attribute (ltr, rtl) on Line element
4. with the language information (but there are exceptions).

Use cases and software may have various heuristics to handle this, but 1. and 2. will always be easier to implement than (3) or a language based assumption system (4).

In the sample, first block: LANG=”latin” (ltr). Second block: LANG=”arab” (rtl)

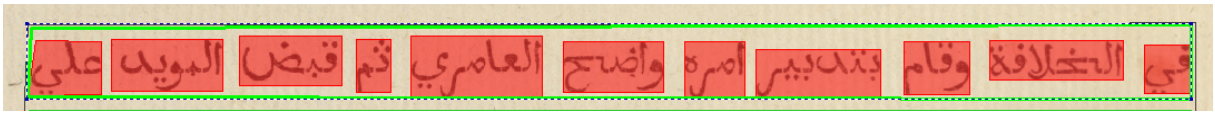


In some cases, reading direction can’t be guessed from language. The foot note below, in a mainly arabic page, has a western reading direction (ltr).



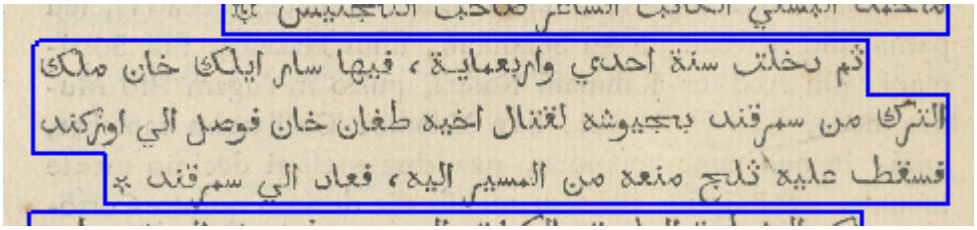
ALTO Use cases:

- **OCR edition/correction** tool with a “Next” button to edit words in a line (or lines in a block). This functionality needs the “reading direction” information.



- **Ebook conversion**: the reading direction information is needed to correctly typeset the text. First indent of the paragraph, text alignment, etc.

  Arabic sample:



Text rendition for arab: right alignment, first indent __

xxx xxx xxx xxxxx xxxx xxxx xxxx TSRIF__

xx xxx xxxx xxxx xxxx xxxxx xx xxxxx xxxxx xxx

TSAL xxxx xxxx xxxx xx xxxx xxx xxx

Default rendition (western writing mode) without reading direction information:

__xxx xxx xxx xxxxx xxxx xxxx xxxx TSRIF

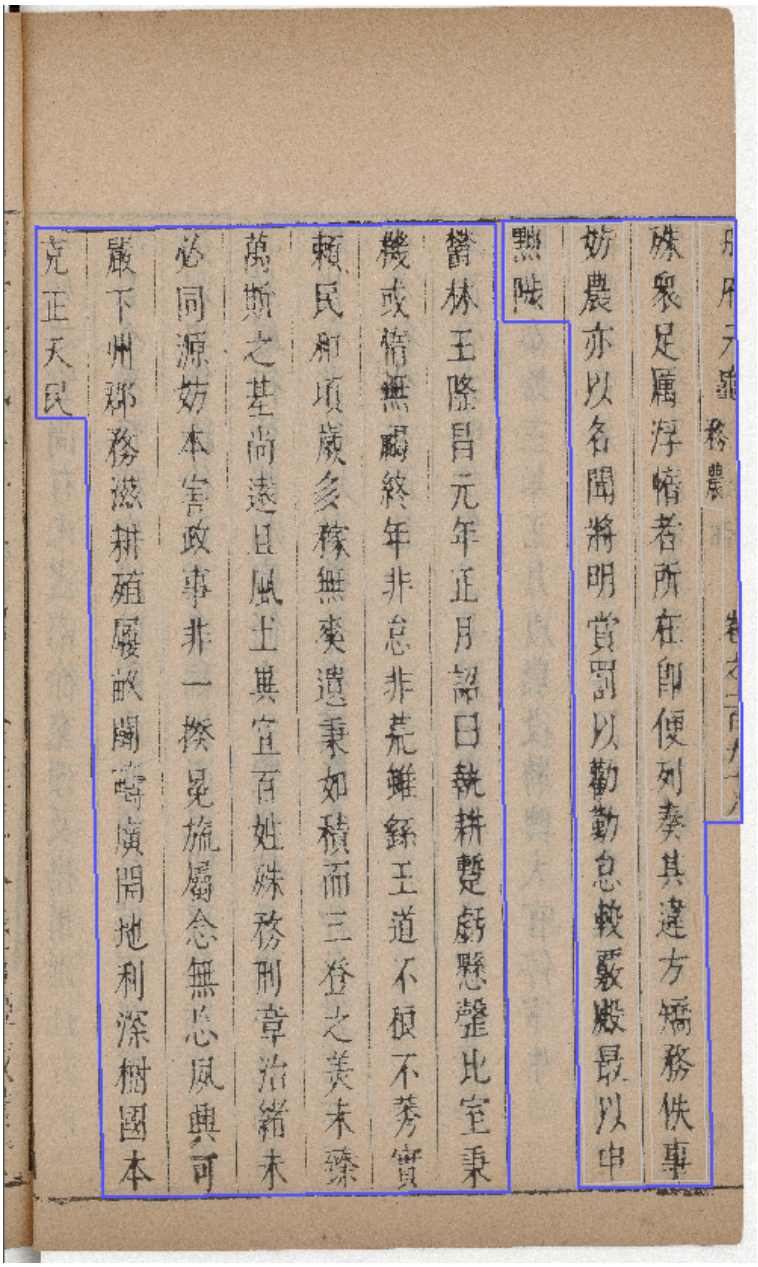
xxx xxxx xxxx xxxx xxxxx xx xxxxx xxxxx xxx xx

TSAL xxxx xxxx xxxx xx xxxx xxx xxx

- **Speech synthesis**: order of words in a line must be known

2. Chinese samples

1. Manuscript



Language and script

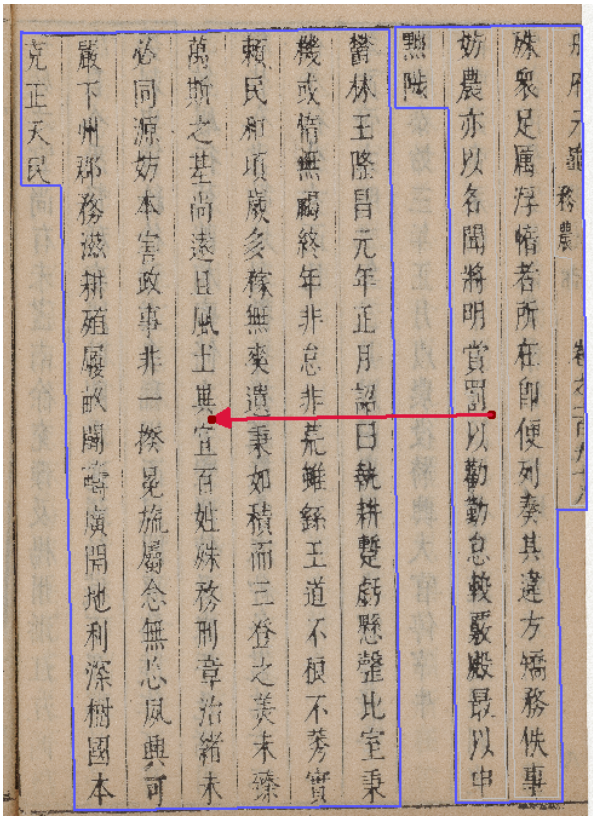
Could be defined at the top level (ALTO root) level (it's a monolingual document).

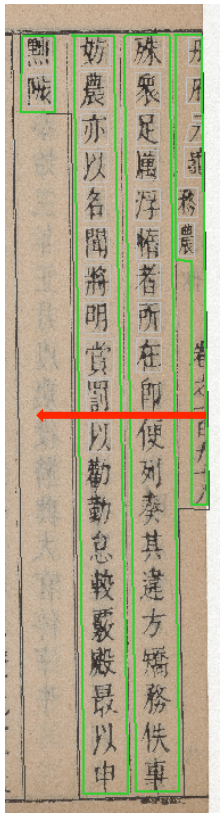
Language: Chinese
Script: Traditional chinese

Block flow direction

Block flow direction is defined by the blocks reading order. A top level attribute could also say: (vertical)right-to-left.

This information also rules the way software must process lines of text within a block. Same as above: IDNEXT attribute on Line element, XML elements order, language based guessing.

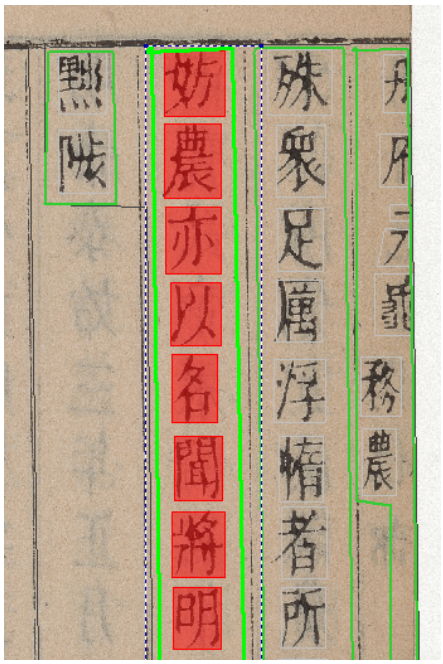




Reading direction

Here, reading direction may be set on the top level (the document has an uniform layout).

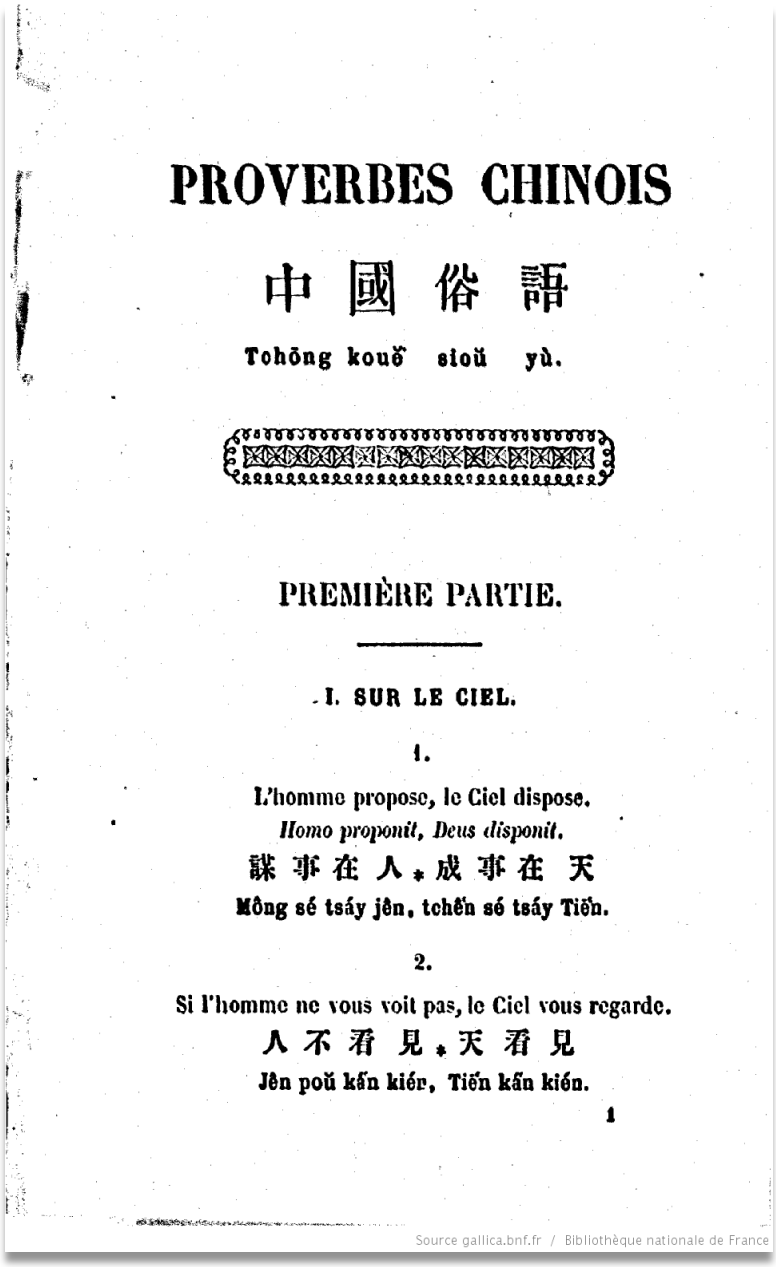
As Chinese can have 2 principal layouts, reading direction must be set to “top to bottom”.



ALTO Use cases: same as Arabic sample

2. **Mixed languages monography**

http://gallica.bnf.fr/ark:/12148/bpt6k57497215/f12.image



PROVERBES CHINOIS

中 國 俗 語

Tchōng kouŏ sioŭ yù.

PREMIÈRE PARTIE.

I. SUR LE CIEL.

1.

L'homme propose, le Ciel dispose.
Homo proponit, Deus disponit.
謀 事 在 人，成 事 在 天
Mŏng sé tsáy jên, tchĕn sé tsáy Tiĕn.

2.

Si l'homme ne vous voit pas, le Ciel vous regarde.
人 不 看 見，天 看 見
Jên poŭ kán kién, Tiĕn kán kién.

1

Language and script

Western layout with French and Latin texts, Chinese text and its phonetic transcription (pinyin):

- Top level : LANG="fr", reading direction: ltr
- Latin texts: LANG="lat"
- Chinese ideograms lines : LANG="chi"
- Phonetic transcription (pinyin): LANG="pinyin"

Reading direction

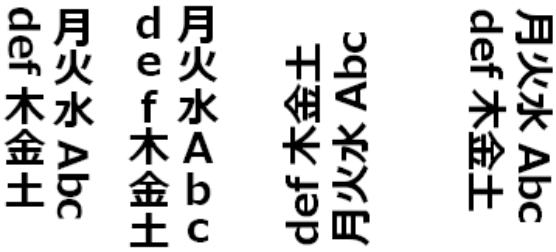
- Top level : reading direction= ltr
- or on Chinese texts only reading direction= ltr

Here, reading direction can't be deduced from language (Chinese can be left-to-right or top-to-bottom).

Other layout features

Orientation of text

Specify the orientation of text within a (vertical) line.



mixed upright sideways-left sideways-right

ALTO already has a Rotation attribute defined on blocks, but use cases are different: a content has a different orientation (relatively to the principal orientation in page or document):

- a large table or an illustration with a sideways layout:


12 MAGASIN PITTORESQUE.

tres. Ainsi, par exemple, il n'est pas rare de voir des cavernes parcourues par des cours d'eau réaliser sous nos yeux la supposition que nous avons faite pour les grottes d'Arcy. La Laibach, en Carinthie, s'engouffre dans la grotte d'Adelsberg, puis reparaît, pour disparaître de nouveau et se perdre enfin dans la caverne de Reifnitz, près de la ville de Laibach. Aux portes de Trieste il existe un cours d'eau souterrain que l'on a cherché à utiliser pour la ville. Dans le département du Jura, la Croisance sort des grottes de Planches-sur-Arbois; la Seille, de celle de Baume-les-Messieurs. Dans celui de l'Isère, la Sassenage s'échappe des grottes du même nom, et la grotte de Balme est parcourue par un ruisseau. On ne peut pénétrer qu'en bateau dans la caverne de Frédéric, en Wurtemberg; et dans celle de Dunobl (Lancashire), en Angleterre, une cascade tombe du plafond et en forme d'autres avant de sortir de la grotte.

On le voit, les cavernes en général, et celles d'Arcy en particulier, sont des canaux souterrains qui ne sont plus parcourus par les eaux qui les ont creusés; et il serait facile de montrer qu'on trouve tous les passages, toutes les nuances entre une simple cavité creusée par une rivière dans les roches qui la bordent, et les systèmes de grottes et de cavernes les plus compliqués. L'action est la même; elle est lente, insensible, mais tous les faits géologiques sont d'accord pour nous prouver ce que peuvent les agents les plus faibles lorsque leur action se continue pendant les milliers de siècles qui correspondent aux âges géologiques de notre planète. En effet, c'est bien avant l'époque historique que les grottes d'Arcy formaient un canal souterrain à la Cure. Il est aisé de le démontrer. Le plafond et le sol sont couverts de stalactites et de stalagmites énormes qui se sont formées avec une extrême lenteur, car elles sont l'œuvre des gouttes d'eau qui suintent de la voûte et s'évaporent en déposant la faible proportion de carbonate de chaux qu'elles tenaient en dissolution. La grosseur et la hauteur de ces stalactites dénotent donc une action prolongée pendant des centaines de siècles; or il ne se forme pas de stalactites dans un canal traversé par un cours d'eau, et, en effet, le pavé de stalagmite recouvre partout le limon et les cailloux roulés. Il faut donc se reporter bien au delà des temps historiques pour arriver à la période où les grottes d'Arcy étaient remplies par une rivière souterraine. Mais si l'on se demande à quelle époque ses eaux ont commencé à dissoudre et à désagréger lentement la pierre calcaire, l'imagination trouve encore des centaines, peut-être des milliers de siècles, entre le moment où la rivière attaquait le rocher et celui où elle remplissait les vastes cavités qu'elle a délaissées depuis.

PRIÈRE D'UNE FEMME ARABE
AU TOMBEAU DE SON ÉPOUX.

(Voy., sur les Funérailles des musulmans, la Table des dix premières années.)

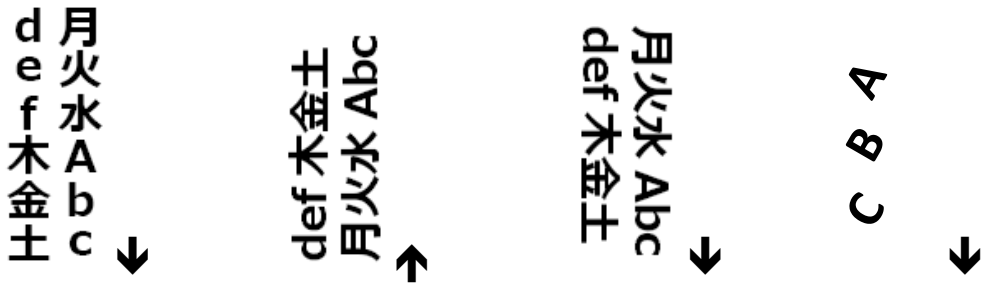
Les Arabes récitent, devant les tombeaux, des prières consacrées par d'anciennes traditions; mais ils expriment aussi leurs souhaits pour les êtres qu'ils ont perdus dans des improvisations dont le caractère varie suivant leur sensibilité ou leur imagination. Rarement ils laissent éclater leur douleur; ils semblent plutôt s'étudier à la contenir: le sentiment qui domine dans ces épanchements de leur âme est une confiance absolue en la volonté divine. Voici, comme exemple, quelques passages d'une prière que l'on a entendu prononcer à une jeune femme.

« O Dieu puissant qui as créé la terre, les montagnes qui lui servent d'appui, et les sept cieux qui la couvrent; Dieu éternel qui as placé au firmament l'astre du jour et le flambeau de la nuit, qui as posé entre les deux océans d'eau douce et d'eau amère des barrières insurmontables; Dieu miséricordieux qui as créé l'homme avec l'eau, et qui, pour sa nourriture, fais couler la pluie des nuages, verdir l'herbe, germer le grain, croître la vigne et le palmier, mûrir la figue, l'olive et la grenade, prends pitié de ma douleur, ne permets pas que je blasphème. Louange à toi, Dieu unique et infini. Tu avais facilité à celui que je pleure le chemin qui conduit à la vie; tu lui avais donné une forme agréable, une taille fine, un corps délié, le recueillement de l'esprit et la sobriété de la parole. Tu lui avais donné l'ouïe et la vue, et, bien qu'il vécût au milieu des pervers, la doctrine divine ne l'a trouvé ni aveugle ni incrédule. Il a goûté la parole du prophète et les dogmes du Coran, merveilleux écrit sur la table gardée. Fidèle musulman, il n'a pas vécu avec faste au milieu de sa famille, il n'a pas transgressé le divin précepte qui défend le meurtre et l'infidélité; croyant vertueux, il n'a pas nié la résurrection et détourné ses regards de la vie future; serviteur du Miséricordieux, il suivait les inspirations de l'esprit, et résistait aux séductions d'Ebhis. Prosterné le matin, le soir et durant les nuits, il récitait dévotement les versets les plus saints de l'Évidence, dont la lecture procure l'indulgence et les faveurs du Seigneur. Il a désiré des enfants qui lui inspirassent la crainte de Dieu; il a secouru ses proches; il a protégé l'orphelin, répandu l'aumône sur le voyageur et sur le pauvre; il s'est interdit les délassements défendus durant les mois sacrés; il a observé l'abstinence pendant le jeûne du Ramadan; il a visité les saints lieux; il a mérité la récompense de sa persévérance et l'accomplissement des promesses de l'Éternel.

» O Dieu, tu as fait passer le juste de la vie à la mort; que la paix soit avec lui. Rends-lui si frais et si doux le tombeau où tu lui as commandé de descendre, qu'au jour de la sépa-

- text within a block with a specific orientation :

Text orientation defines the direction of glyphs relatively to the reading direction (clockwise angles):



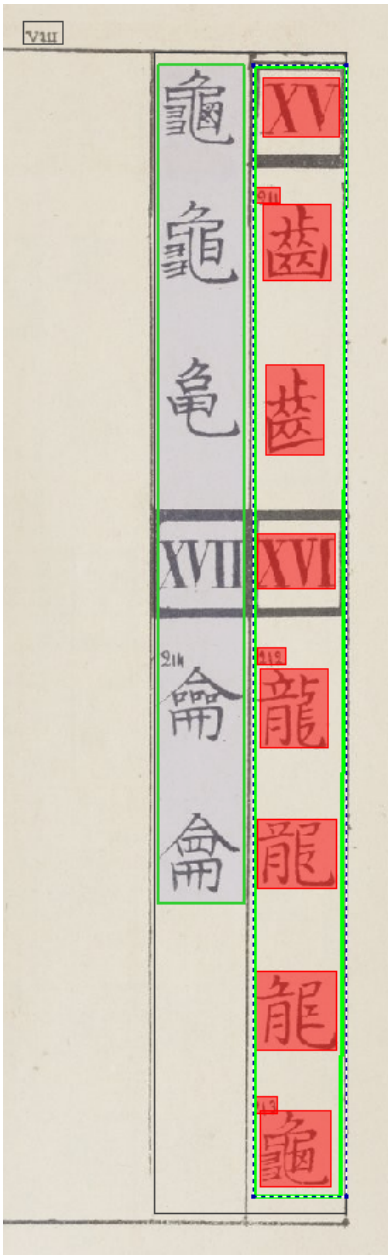
+180° -90° -90° +135°

ALTO Use cases: text orientation is needed to produce accurate facsimiles.

Text combine

Monography with combination of multiple characters into the space of a single character. In East Asian documents, the 'text-combine: horizontal' effect is often used to display Latin-based strings.

http://gallica.bnf.fr/ark:/12148/bpt6k6580409q/f14.image



ALTO Use cases: not clear…

3. Conclusion

For core OCR purposes, ALTO functionalities are enough: blocks order, LANG attribute, UNICODE characters.

For advanced use cases, ALTO lacks a complete writing modes model. Data needed to feed this model can be set:

- Before OCR (to help it) – Eg: this block is detected (by human or image analysis system) as horizontal in a Chinese newspaper (mainly vertical).
- During OCR – Eg: this block has been tagged as latin.
- After OCR – Eg: enrichments from other processes are input for "enhanced" ALTO: text, text and layout attributes. Example: ALTO ➔ EPUB ➔ ALTO

ALTO proposals

- Language and Script:
  - attributes at all levels: page, block, line, word (optional attributes)
- Block flow direction:
  - attribute at the page level (optional): (horizontal)top-to-bottom, (vertical)right-to-left, (vertical)left-to-right, mixed (eg for Chinese and Japanese newspapers).
    NB: bottom-to-top writing mode doesn't exist (in the real world)
  - IDNEXT at block and line levels
- Reading direction:
  - attribute at block and line levels: top-to-bottom, bottom-to-top, right-to-left, left-to-right
  - IDNEXT at word level
- Text orientation:
  - Specific new attribute (angle values) at line and word levels

+ guidelines for languages support with ALTO