Assessing Quality of Digital Objects Created in Large Scale Digitization

Emily Campbell
Special Projects Librarian
University of Michigan-Ann Arbor

Background

"U of M will engage in ongoing review (through sampling) of the resulting digital files, and shall inform Google of files that do not meet benchmarking guidelines or do not comply with the agreed upon format" – *UM-Google-Agreement Section 2.4 Digitizing the Selected Content*

Process

- Two file formats: bitonal TIFF's and continuous JPEG2000 files
- Samples are pulled off the Google server
- Once the samples have been moved to storage a random consecutive 20 page sample is created from every volume received and a .csv file is created

Process, cont.

- A number of volumes is uploaded to the database each day
- A script then combines the individual .csv files for each of the volumes into one large volume that is, in turn, uploaded into the QR database.
- A worksheet is generated for each volume and is accessed via our web interface.

Critical vs Non-Critical

Critical: *Page content is incomprehensible due to conditions not inherent to the physical book*

Critical

qu'on éprouve à faire concorder les résultats de l reconstruction interne et ceux de la reconstructio comparative. Le système le plus largement utilis comporte trois laryngales, c'est-à-dire qu'il repos essentiellement sur la reconstruction interne ; l critère principal est l'*effet de coloration*, *infra*, p. 1

« Quasi-sonantes » de la reconstruction interne	Correspondants anatoliens	Reconstructio
**E* (pas d'effet de coloration)	*h*	**H*$_1$
**A* (coloration *a*)	*hh* (en hittite seulement)	**H*$_2$
**O* (coloration *o*)	*h*	**H*$_3$

Mais ce système est insuffisant pour rend compte des correspondances où *h* manque en ana tolien sans qu'on puisse invoquer une chute cond tionnée par l'environnement, et inversement de l anatolien ne correspondant pas à une quasi-sonant En outre, d'autres correspondances paraissent étu blir l'existence d'une laryngale labio-vélaire **H* et d'une laryngale palatale **E*^{y}, *infra*, p. 16.

D) *Évolution des laryngales*

a) Les laryngales entre voyelles. — Elles ne se conserve qu'en anatolien, hitt. *mehur* « temps » (ailleurs : **mē*-), et di paraissent dans les autres langues indo-européennes, prou quant des contractions vocaliques.

b) Les laryngales devant voyelle. — Elles se conservent anatolien et « colorent » éventuellement la voyelle (1) e

(1) Selon certains, comme J. Kuryłowicz, **H*$_1$ colore en *a* aus bien **o* que **e* ; selon d'autres, comme Beekes, *Sprache* 18, 197 p. 117-131, **H*$_1$ ne colore que **e* : il existe donc, selon lui, u alternance **a* : **o*.

Non-Critical

Non-Critical: *Page content is comprehensible though certain adverse conditions do exist*

Non-Critical

›f power through antiquity, through a thematics that ·efers to a more traditional conception of the state and ı cunning deliberation on the part of the statesman, Odysseus?

Müller: In my version of the play, the Trojan war is just ı sign or image for the Socialist revolution reaching the stage where it ends up in stagnation, in a stalemate si-:uation. Odysseus didn't want to enter the war; he was :ompelled to do so. Nobody really wanted it but now they ıre all in it and the only way out is to go deeper into ıt in order to put it to an end. There is no ideology any-nore but you can't end the war without destroying the enemy.

Lotringer: How does this relate precisely to the Socialist ·evolution?

Müller: A student from Göttingen was writing his disser-tation on my version of *Philoctetes*. He came to East Ber-lin to ask me a few questions. He sat down and he took ı small piece of paper out of his socks. Then he read his questions. One of them was why the rocks on Lemnos where Philoctetes was isolated, were red. He had found out - I didn't know it - that Trotsky's first exile was spent on an island near Turkey known for its red rocks. That was why he had hidden the paper in his socks. He was afraid that mentioning Trotsky would be a problem at the border. He read the play as a reference to a situation where Stalin needs Trotsky's help again and tries to persuade him to come back. But Trotsky doesn't want to return, he has grown beyond that. The student didn't know that I wrote the play about the stalemate situation of Soviet Socialism, and more generally about the Russian revolution in the context of world revolution. Lenin's idea that the German revolution was near because revolution was bound to happen first in industrial countries didn't prove true. The German revolution failed and he had to give up on the idea of revolution or implement it in one country only. And since there was no other object, it meant colonizing your own population.

Error Types

Characteristic	Critical	Non-Critical
Thick	Some letters are heavy, thick or dark compared to the rest of the text, resulting in an inability to distinguish individual letters and words	Some letters are heavy, thick or dark compared to the rest of the text, but letters and words are legible and there is no loss of information
Broken	Letters are substantially broken, resulting in unreadable text	Some letters are broken, but the text is readable
Blurred	Text is blurry and unreadable, image content cannot be distinguished	Text and/or image is slightly blurred, not "crisp", but all information is still legible
Cleaning	Portions of the text or image are erased from page, resulting in loss of information content	There are missing portions of the page, but there is no resulting loss of information. Or physical objects not associated with the text are visible, but do not result in a loss of information
Warp	Page is noticeably curved with portion of text/image either missing or illegible	Page is noticeably curved, but text is still readable and there is no resulting loss of information
Crop	Over-cropping has cut off a portion of an image or has caused enough text loss that information is missing or unreadable	Text/image edges have been cut off, but there is no resulting loss of information or image includes area beyond margin of the page e.g. scanning cradle is visible (under cropping)
Obscured	Text or images are covered in some way, but not erased and there is a loss of information	N/A**when part of a page is obscured but there is no loss of information, this is classified as a cleaning error*
Cleaning	N/A	Black and white text, captured as color

Outliers and other error types

Error Not Noted	Description
Black and White Textual JPEG200's	Pages that have been captured as JPEG200's even though they contain only text, they do not have ANY residual color to them
Upside down pages	Pages that appear on the screen upside down, but there is no loss of information from this error, it is seen very rarely
Skew	When a page is titled at an angle, but the page is not cropped off and there is not resulting loss of information
Pagination issues	When blank pages have been inserted due to a mistake in the metadata, there will be what appear to be "missing pages " in the text
Marginalia	Handwritten comments, underlining or drawings found within the volume that are not inherent to the volume

“Oddballs”

- Any error that has not previously been noted by reviewers
  - Information collected on a monthly basis
- One of the most important parts of the process because it allows us to provide specific examples to Google

Reporting

- Monthly Reports are sent to Google
  - Allows for a solid dialogue based on numbers
- Collecting longitudinal data since May of 2007 and we continue to do so on a monthly basis

Trends

	May 2006-April 2007	May 2007-April 2008	May 2008-April 2009	May 2009-April 2010
# of critical Thick Volumes	189	70	19	144
# of critical broken volumes	518	121	76	64
# of critical blurred volumes	252	40	10	54
# of critical cleaning volumes	208	214	1256	439
# of critical warp volumes	47	37	14	22
# of critical crop volumes	424	246	100	67
# of critical obscured volumes	57	35	21	8
# of noncritical colorization volumes	3250	272	35	19
# of volumes reviewed	33,047	36,981	29,677	17,850

Special Projects

- C-Size
  - Oversized books were not a problem
- Special Collections
  - Pre-scan sorting process to control for quality
  - Pre and post scan condition reviews
  - Where to go next?

Conclusions

- Our data has shown that the quality of digital images has greatly improved over the past several years
- Quality will continue to improve as Google refines its process
- As technology improves the original scans will become more and more viable as replacements or surrogates for the physical objects

M Library