An Integrated System for Polytonic Greek OCR

I. Generating the Data

Bruce Robertson, Dept. of Classics, Mount Allison University, New Brunswick, Canada
I. Generating the Data
   A. Reason
Why Ancient Greek OCR?

1. Rapid digitization of Greek texts not yet in digital libraries
Why Ancient Greek OCR?

1. Rapid digitization of Greek texts not yet in digital libraries
2. Study of textual variants and app. crit.
Why Ancient Greek OCR?

1. Rapid digitization of Greek texts not yet in digital libraries
2. Study of textual variants and app. crit.
3. Text reuse analysis
Why Ancient Greek OCR?

1. Rapid digitization of Greek texts not yet in digital libraries
2. Study of textual variants and app. crit.
3. Text reuse analysis
4. General-purpose OCR search, like Google Books
<table>
<thead>
<tr>
<th>Use</th>
<th>Manual Editing?</th>
<th>Automatic Spell-checking?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digitization</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Textual Variants</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Text Reuse</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>OCR Search</td>
<td>✗</td>
<td>✗ or ✓</td>
</tr>
<tr>
<td>Use</td>
<td>Manual Editing?</td>
<td>Automatic Spell-checking?</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Digitization</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Textual Variants</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Text Reuse</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>OCR Search</td>
<td>✗</td>
<td>✗  or ✓</td>
</tr>
</tbody>
</table>
I. Generating the Data

B. Challenge
Ἐν ἀρχῇ ἦν ὁ Λόγος,
καὶ ὁ Λόγος ἦν πρὸς τὸν Θεόν,
καὶ Θεὸς ἦν ὁ Λόγος.
Ἐν ἀρχῇ ἦν ὁ Λόγος,
καὶ ὁ Λόγος ἦν πρὸς τὸν Θεὸν,
καὶ Θεὸς ἦν ὁ Λόγος.
Smooth and Rough Breathing Marks

Ἐν ἀρχῇ ἦν ὁ Λόγος,
καὶ ὁ Λόγος ἦν πρὸς τὸν Θεόν,
καὶ Θεὸς ἦν ὁ Λόγος.
Ἐν ἀρχῇ ἦν ὁ Λόγος,
καὶ ὁ Λόγος ἦν πρὸς τὸν Θεὸν,
καὶ Θεὸς ἦν ὁ Λόγος.
Diversity of Greek Fonts in 19th C. archive.org Texts

κράτος τ' ἱσόψυχον ἐν γυναικῶν

καρδιόδηκτον ἐμοὶ κρατύνεις.

Περὶ δὲ δικαιοσύνης καὶ ἀδικίας

τε τυχάνων οὖσας πράξεις, καὶ ποίο

πιπλαττομένων ἢ μὴ

κ[α]λ]ῶς ἀλλως προς[τιθεμέ-
κότος ἀλλ' οὗ τῆς κομιζούσης φύσεως κτήμα
τες οὖν, ὡς ἐν ἔργον ἡμῖν ἐπιβάλλει μόνον
πρὸς ἀπολογιάν προφέρειν,
περιττὸν εἰλήφασιν. Ἡλλ'

εἰτ' έστιν Ἱοριανῆ θάλαττα λεγομένη.
Θεόπομπος ἀναγράφει δὲ ταύτης τὴν θέσιν,

ητίς ἐστίν: ἄριθμὸς, μονᾶς, δύναμις, χύδος, δυ
μοδύναμις, δυναμόχυδος, χυδόχυδος. Ταύτην
Recognizing Lines
Recognizing Lines
Accurate Binarization

cαι ἀδίκις σκέπτεον, περὶ
tυγχάνωσιν οὕτωι
Binarization Important to Results

B-Scores Varying By Image Darkness

Mogrify 'black-threshold' value (%)
I. Generating the Data
C. Resources
Contextless 'Greekness' Index

- Devised by Dr. Boschetti
- Based on dictionary and likely sequences of letters, etc.
- Named 'B-score' in these slides
Archive.org

- Provides:
  - Thousands of volumes rendered in high-resolution (400 ppi +) colour images
  - OCR results from ABBYY Finereader
    - Excellent Latin-script recognition
    - Poor Greek results
    - Top-quality line-segmentation
Open-source OCR Engines

- **Gamera**
  - Current focus of my team

- **Tesseract**
  - Nick White has worked extensively on this to generate good results

- **OCRopus**
  - Dr. Boschetti recently has been able to use Tesseract training sets for this engine
νυ νυν
δ'
ωρθωσας

ωρθωσας
I. Generating the Data

D. Method
Rigaudon Greek OCR Process

HOCR Output

Replace Latin-script output words with ones in same position from Archive's ABBYY output

Per-volume spellcheck table

Weighted Levenshtein Automatic OCR Spellchecker (x14 cores)

410,000 word dictionary from open Perseus Greek texts

Weighted Edit Table for Classifier

Replace spellchecked words

Reduction to unique Greek strings

Select highest-ranking binarization page

Boschetti scoring

Score table for binarization thresholds

HOCR Results at a range of binarization thresholds

Parallel Process x35 Cores

Classifiers for Teubner Sans, Teubner Serif, Oxford (Loeb, etc.)

Greek OCR For Gamera

JP2 Input Library

Page Segmentation Thru HOCR Input

HOCR Output

ABBY to HOCR Conversion

ABBY OCR Information file

Images from Archive.org

HOCR Latin / Greek Combining

Automatic OCR Spellcheck

HOCR "Blending"
Raw HOCR Production Using Gamera

- Plugin for Gamera OCR allows it to import high-quality line-segmentation information, compensating for Gamera's poor results in this critical function
- Plugin to output HOOCR
- Wrapper function generates a range of output pages based on binarization threshold (typically 10 - 20 per page)
This step aims to gather word-by-word the 'best' results from the range of results pages for each image.
Selects the highest-scoring result page overall.
Where a Greek word in this page is not in the dictionary and another page has a dictionary word in the exact same physical location, it replaces with dictionary word.
Automatic Spellcheck

- All pages in volume are reduced to a set of unique, decomposed Greek strings
- These are compared to dictionary using Levenshtein distances
- A 'weighting table', suitable for a given font, indicates which edits are preferable or allowed
- Result is 'light' correction, esp. of diacritics
Automatic Spellcheck
Weighting Table

['replace', ur'ϲϹ', ur'σςΣ', 1],  # for lunate fonts
['replace', ur'c', ur'σς', 1],  # for lunate fonts
['replace', ur'T', ur'Ττ', 1],
['replace', ur'Tr', ur'τ', 1],
['replace', ur'Uu', ur'u', 1],
['replace', ur'Y', ur'Υ', 1],
['replace', ur'E', ur'Ε', 1],
['replace', ur'Ε', ur'ε', 2],
['replace', ur'Z', ur'Ζ', 1],
['replace', ur'K', ur'κK', 1],
Optionally injecting Greek into Original Latin HOCR

- Don't want to try to get excellent Greek and Latin results, esp. when ABBYYY and others do better job with Latin
- In the case that archive.org provides Latin OCR:
  - If Rigaudon's output word is Greek, replace archive.org's ABBYYY output word with Rigaudon's
Reporting
I. Generating the Data

F. Results
Περὶ δὲ δικαιοσύνης καὶ ἀδικίας ...
tε τυγχάνουσιν οὕσαι πράξεις, καὶ ποία ...

κότος ἀλλ’ οὐ τῆς κομιζούσης φύσεως κτῆμα
tες οὖν, ὡς ἐν έργον ἡμῖν ἐπιβάλλει μόνον

κότος ἀλλ’ οὐ τῆς κομιζούσης φύσεως κτῆμα ...
tπ οὖν, ὡς ἐν έργον ἡμῖν ἐπιβάλλει μόνον ...

Results
πιπλαττομένων ἢ μὴ κ[αλ]ῶς ἄλλως προσ[τιθεμέ-]
κράτος τ' ἰσόψυχον ἐκ γυναικῶν καρδιόδηκτον ἐμοὶ κρατύνεις.
I. Generating the Data

G. Future
Multiple OCR Engines

- Take ABBYY data out of the process
  - With 'cleaning' Tesseract's line-segmentation is often as good
- Use Nick White's general-purpose polytonic classifier and ones specifically designed for a font
Resources

Output:
http://heml.mta.ca/rgaudon

Code:
https://github.com/brobertson/rgaudon

Further Topics

HPC Computing with Grid Engine
Python Flask Web Microframework
Making Book Images
An Integrated System for Generating and Correcting Polytonic Greek OCR

Bruce Robertson and Federico Boschetti

Part II

The Proof-reading Process

Federico Boschetti*

federico.boschetti@ilc.cnr.it

*ILC-CNR of Pisa

Digital Classicist Seminars – London, 19 July 2013
Manual corrections on OCR output may be performed by

- **Experts** Classicists devoted to proof-reading for a long-term project
- **Data Entry Firms** Professional proof-readers not skilled in the target language(s)
- **Crowd Sourcing** Students that are learning the target language(s)
- **Random Volunteers** People with heterogeneous education and skills
For this reason proof-reading tools focused on ancient languages should be

- centralized
- easy to use
- based on image / text comparison line by line
- optimized to catch attention on possible errors, distinguished by category
- efficiently providing the most probable correction
1. **Information Aggregation**
   - Enriched hocr files
   - Alignment with other editions
   - False negatives

2. **Proof-reader Web Application**

3. **False positives**
Enriched hocr files

OCR output formatted in hocr microformat
The hocr output produced by Rigaudon is postprocessed, in order to add information managed by the Proof-reading Web Application

Multiple sources
- Dictionaries with and without diacritics
- Multiple editions of the same work (if available)
- Syllabic repertory
In order to identify possible errors and provide good suggestions to correct them, the OCR output is spell-checked and the potential errors are processed step by step.

The spell-checker is based on dictionaries generated from Perseus’ text collection. An upper-case dictionary is used to evaluate if a character sequence is a word with a wrong accent or breathing mark.
Alignment with other editions

When another edition of the same work is available, the two editions are aligned word by word applying the Needleman-Wunsch algorithm.
False negatives and the risk of digital *contaminatio*

An example

- Rigaudon on the *Anecdota Graeca* edited by Cramer recognizes the word χόρος, which is rejected by the current spellchecker.
- The spell-checker suggests χορός as a correction.
- Also the alignment with Koster’s edition of the *Prolegomena de comoedia* suggests χορός.
- But the page image contains χόρος, a late form attested from Athenaeus to the Byzantine period.
Syllabication

- In order to prevent false negatives due to (rare) variants ignored by the dictionaries, the system distinguishes between random character sequences and well-formed syllabic sequences.
- Each potential error is divided in syllables and each syllable is evaluated according to its position.
- For example, χό-ρος is a well-formed syllabic sequence: χό- is a valid Greek initial syllable and -ρος is a valid final Greek syllable.
Overview

1. Information Aggregation

2. Proof-reader Web Application
   - The web interface
   - Cues
   - Self-corrections

3. False positives
Centralization

- The proof-reader is a web application inspired by the Mozilla hocr Editor interface but employs the WikiSource collaborative philosophy
- Texts are stored in a central XML native database
The Control Panel

Euclides, Opera1

133
208
344

The web interface
Cues
Self-corrections
Image / Text Pairs

 Generating and Correcting Polytonic Greek OCR
Cues

- **Wrong accents and breathing marks** Attention is focused on diacritics
- **Self-corrections** Special care is necessary to avoid the risk of *contaminatio*
- **Errors** Random errors
Example

Wrong accents and breathing marks
Self-corrections
Errors
Self-corrections and suggestions generated by alignment

In a self-correction, the reading has been substituted by the aligned word of another edition. Self corrections need three conditions:

- character sequence is refused by the spell-checker
- edit distance between the character sequence and the aligned edition is very close
- the character sequence is not a well-formed syllabic sequence
Example

The web interface
Cues
Self-corrections

Information Aggregation
Proof-reader Web Application
False positives
Dynamic Dictionaries

- Dictionaries used by the spell-checker are dynamically rebuilt when a milestone in proof-reading is reached.
- Enlarging the dictionaries, rare variants are acquired and used to spell-check the next works.
Overview

1. Information Aggregation
2. Proof-reader Web Application
3. False positives
False positives are deceitful

- By definition, false positives pass the spell-checking.
- Specially if they are graphically similar to the correct word, such as δ and ὅ in Greek or m and ni in Latin, they are difficult to be seen, in particular by proof-readers not skilled in the target language(s).
Example
Semantic Distance

- Semantic distance is calculated along the nodes of WordNet’s hierarchy, i.e. along the chain of hyponyms / hypernyms, in order to reach co-hyponyms.
- Different translations of the same concepts (e.g. *vis* in Latin and *efficacia* in Italian or *efficacy* in English) have semantic distance equal to zero.
- Semantically unrelated words (e.g. *vinum* in Latin and *efficacia* in Italian) have a large semantic distance.
AncientWordNet

- Synsets of AncientGreekWordNet and LatinWordNet have been extracted from bilingual dictionaries
- They are aligned to modern languages such as English, Italian, etc.
The proof-reading Web Application puts together the main features of individual and collaborative proof-reading tools currently available.

The entire work-flow is circular: Training OCR - Performing OCR - Spell-checking OCR - Correcting OCR - Enlarging dictionaries - Retraining OCR.
Thank you for your attention


C. Ringlstetter, K. Schulz, S. Mihov, K. Louka: The same is not the same - postcorrection of alphabet confusion errors in mixed-alphabet OCR recognition. 8th International Conference on Document Analysis and Recognition, 1, 406–410 (2005)

