Minoan linguistic resources: The Linear A digital Corpus

Tommaso Petrolito⊙⊕ Ruggero Petrolito⊙ Grégoire Winterstein⊕⊕
Francesco Perono Cacciafoco⊕⊕
⊙ Filologia Letteratura e Linguistica, University of Pisa, Italy
⊕ Linguistics and Modern Language Studies, The Hong Kong Institute of Education, Hong Kong
⊕ Linguistics and Multilingual Studies, Nanyang Technological University, Singapore

tommasouni@gmail.com, ruggero.petrolito@gmail.com,
gregoire@ied.edu.hk, fcacciafoco@ntu.edu.sg

30 July 2015
Introduction

- We’ll describe the Linear A/Minoan digital corpus and the approaches we applied to develop it
- Why we should develop a Linear A Corpus and the reasons for which we chose XML-TEI EpiDoc
- Available resources and developing process
- The Linear A Corpus as Cultural Heritage
Linear A and Minoan

- The Linear A script was used by the Minoan Civilization (Crete, 2500 – 1450 BC) and it still remains undeciphered.

- Many symbols are shared by both Linear A and Linear B and are assumed to have phonetic values. The others are probably logograms:
  
<table>
<thead>
<tr>
<th>Linear A/B</th>
<th>Linear A</th>
</tr>
</thead>
<tbody>
<tr>
<td>symbols</td>
<td>81</td>
</tr>
<tr>
<td>value</td>
<td>syllable</td>
</tr>
</tbody>
</table>

- Linear B has been deciphered (during the ’50s) and found to be used to write an Ancient Greek dialect, so many scholars are trying to decipher Linear A too.
Lack in digital resources

- After decades no deciphering attempts have been successful
- No heavy computational approaches have been attempted
- Only John G. Younger, in his website, provides a complete digital collection
  - Nevertheless, it is stored in two simple HTML pages with not strict structure and transcribed as transliterations
- A new digital corpus in a suitable format and well organized may be a useful resource
Available resources

- 1,427 Linear A documents containing 7,362-7,396 signs
- GORILA paper collection of inscriptions and transcriptions
- John G. Younger’s website
GORILA: Louis Godart and Jean-Pierre Olivier, *Recueil des inscriptions en Linéaire A*

- GORILA contains
  - a catalog of symbols/numeric codes
  - documents indexes with information about original place and type of support (these indexes were defined in the first place by Pope&Raison)
  - indexed documents descriptions including pictures, drawings and handmade transcriptions

- the GORILA information is the standard point of reference: even recent collections always refer to the GORILA volume and page
John G. Younger’s website

- [http://people.ku.edu/~jyounger/LinearA/](http://people.ku.edu/~jyounger/LinearA/)
- The website contains:
  - Two HTML pages, one for Haghia Triada’s documents, one for all the other places of origin
  - 1,077 transcriptions, with Linear B phonetics and GORILA code numbers (75.5% of the total amount of existing documents listed in GORILA)
  - A conversion table: GORILA code numbers to syllables
From Younger’s syllables to Unicode

<table>
<thead>
<tr>
<th>Unicode</th>
<th>GORILA</th>
<th>Syllable</th>
</tr>
</thead>
<tbody>
<tr>
<td>10600</td>
<td>AB01</td>
<td>DA</td>
</tr>
<tr>
<td>10601</td>
<td>AB02</td>
<td>RO</td>
</tr>
<tr>
<td>10602</td>
<td>AB03</td>
<td>PA</td>
</tr>
</tbody>
</table>

- The Unicode set of characters for Linear A was released in June 2014.
- The 1,077 documents represented on Younger’s website have been automatically converted:
  - from the syllable transcription (coexisting alongside GORILA code numbers for symbols not included in Linear B) to the full GORILA code numbers transcription.
  - from GORILA code numbers to Unicode.
Segmentation issues

- Separation is mainly indicated in two ways:
  - by isolating sign groups with numbers or logograms, thereby implying a separation
  - dots between sign groups, always used if there are long sign groups strings

- Example: This is a Linear A line: \( \text{\begin{figure}
\begin{center}
\includegraphics[width=\textwidth]{example.png}
\end{center}
\end{figure}} \)
  - \( \text{\begin{figure}
\begin{center}
\includegraphics[width=\textwidth]{example.png}
\end{center}
\end{figure}} \) is a number (it is assumed to be a number 5)
  - so \( \text{\begin{figure}
\begin{center}
\includegraphics[width=\textwidth]{example.png}
\end{center}
\end{figure}} \) and \( \text{\begin{figure}
\begin{center}
\includegraphics[width=\textwidth]{example.png}
\end{center}
\end{figure}} \) are assumed to be separated sign groups
Corpus data format

- **XML** provides important advantages
  - metadata on several levels of annotation
  - elements and entities for unsupported glyphs or symbols

- **EpiDoc** is a TEI DTD with customization for Epigraphy
  - TEI-using community can provide support
  - a wide range of best-practice examples are available online

- The "old" Leiden system annotation task, familiar to epigraphers, is quite similar to the XML TEI EpiDoc annotation process
Corpus data format example

```xml
<div lang="minoan"
   n="text"
   type="edition"
   part="N"
   sample="complete"
   org="uniform">
  <head lang="eng">Edition</head>
  <cb rend="front" n="HM 1673"/>
  <ab part="N">
    <lb n="1"/>
    <w part="N">𐀟𐀙</w>
    <space dim="horizontal"
       extent="1em"
       unit="character"/>
    <w part="N">𐀟𐀙</w>
    <lb n="2"/>
    <w part="N">𐀟𐀙</w>
    <g ref="#n5"/>
    <w part="N">𐀟𐀙</w>
    <lb n="3"/>
    <w part="N">𐀟𐀙</w>
    <g ref="#n12"/>
    <w part="N">𐀟𐀙</w>
  </ab>
</div>
```

Petrolito, Winterstein, Perono Cacciafoco

Linear A Corpus

NANYANG TECHNOLOGICAL UNIVERSITY
Unsupported glyphs handling

- Inside the EncodingDesc>CharDecl elements, glyph elements can be defined.
- g elements referring to glyphs can be used to represent unsupported symbols.

```xml
<glyph xml:id="n5">
  <glyphName>
    Number 5
  </glyphName>
  <mapping type="standardized">
    5
  </mapping>
</glyph>
```

```xml
<lb n="2"/>
  <w part="N">هة</w>
  <g ref="#n5"/>
  <w part="N"> italia</w>
```
Corpus size

- GORILA: 1,427 Linear A documents
- John G. Younger’s website: 1,077 Linear A transcriptions (75.5% of the total)
- Our corpus will contain up to 1,077 Linear A XML TEI EpiDoc documents
- The Unicode conversions of John G. Younger’s transcriptions have been converted in XML in an automatic way but the tagging has been only partially carried out
- The main remaining work (still in progress) is manually checking the data with the GORILA volumes
Before the release of Unicode 7.0, there was no way to visualize characters in the range 10600–1077F

The ’traditional’ Linear A font, LA.ttf, included wrong Unicode positions

We developed a new Linear A font, named after John Younger to show our appreciation for his work: John_Younger.ttf (available at http://openfontlibrary.org/en/font/john-younger)
The Linear A corpus is an important cultural monument, storing information about tradition, knowledge and lifestyle of Minoan people. Even without a full understanding of transcriptions some cultural features can be inferred:

- **Economics and commerce**: as some ideograms for basic commodities are similar to their Linear B counterparts, we can compare types and amounts of commodities.
- **Religion**: there are around thirty libation formulas transcribed on various supports.
Future work and Acknowledgements

- XSL style sheets in order to create suitable HTML pages
- A web interface to annotate and enrich the corpus information
- All the data will be freely available and published at the following URL: http://ling.ied.edu.HK/~gregoire/lineara

- This work was started when the 1st, 3rd and 4th authors were visitors at NTU, support by the Erasmus MULTI II exchange program.
- We thank John Younger for permission to use the data from his website.