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:
-

	Linear A/B	**Linear A**
symbols	81	260
value	syllable	logogram

- 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

(about 2 A4 pages of text at 11pt)

- GORILA paper collection of inscriptions and transcriptions
- John G. Younger's website

GORILA

- **GORILA**: Louis **G**odart and Jean-Pierre **O**livier, ***R**ecueil des **i**nscriptions en **L**inéaire **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

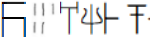
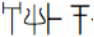
- `http://people.ku.edu/~jyounger/LinearA/`
- the website contains
  - two HTML pages, one for Haghia Triada's documents, one for all the other places of origin
  - 1,077 transcriptions, with Linear B phonetics and GORILA code numbers (75.5% of the total amount of existing documents listed in GORILA)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

From Younger's syllables to Unicode

Unicode	GORILA	Syllable
10600	AB01	DA
10601	AB02	RO
10602	AB03	PA

- 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

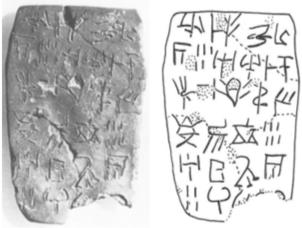
- Separation is mainly indicated in two ways:
  - by isolating sign groups with numbers or logograms, thereby implying a separation
  - dots between sign groups, always used if there are long sign groups strings
- Example: This is a Linear A line: 
  - is a number (it is assumed to be a number 5)
  - so and  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

```
<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">𐙂𐘝</w>
    <space dim="horizontal"
           extent="1em"
           unit="character"/>
    <w part="N">𐘀𐘲</w>
```



```
<lb n="2"/>
   <w part="N">𐘢</w>
   <g ref="#n5"/>
   <w part="N">𐘠𐙂𐘃</w>
<lb n="3"/>
   <w part="N">𐘸</w>
   <g ref="#n12"/>
   <w part="N">𐙁𐘘𐘤</w>
```

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

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

```
<lb n="2"/>
   <w part="N">𐘢</w>
   <g ref="#n5"/>
   <w part="N">𐘆𐘚𐘳 𐘃</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 remaing work (still in progress) is manually checking the data with the GORILA volumes

John Younger ttf

- 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)

From Linear A to Minoan culture

- 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.