Disambiguation and Morphological Analysis of Linear B Sign Sequences

The DĀMOS corpus

- Only online, (not yet) annotated corpus of Mycenaean Greek
- Still unlicensed, will be Open Access
- All 5800 known Linear B documents are available
- Currently annotation is done manually
  - Query database for potentially interesting words
  - Look up rules for sign sequence expansion
  - Disambiguate
  - Verify using dictionary
  - Requires knowledge of ancient Greek, English, and Spanish

The DĀMOS corpus

- Write an annotation support tool
- Let a machine do all the heavy lifting
- Automatize at much as possible of the pipeline
- Luckily no syntactic analysis needed (not much syntax in the corpus)

A Short History of Linear B

- Script used to write earliest dialect of Greek, Mycenaean
- Used in Crete and the mainland between 1400 and 1200 BCE
- Extant texts (mainly on clay tablets) are solely for temporary administrative records
- Syllabic, i.e. each sign corresponds to a single syllable
- Some ideograms for objects and units of measure
- Overall slightly less than 200 unique signs

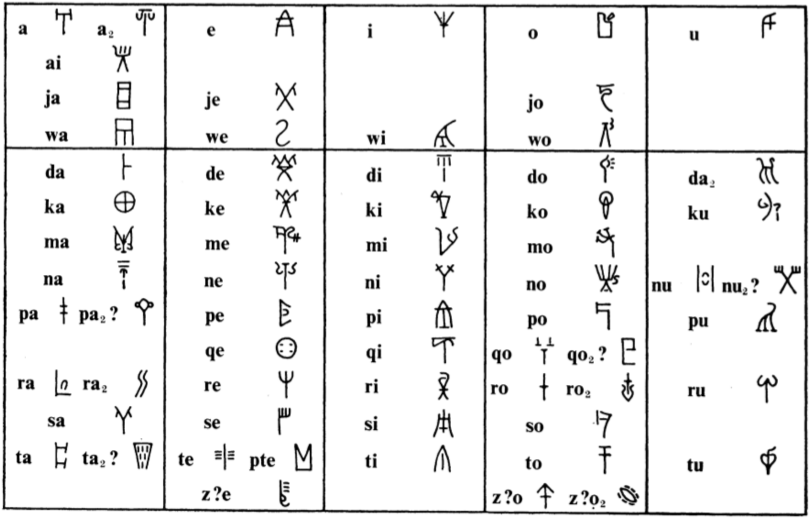
A Short History of Linear B

- Descends from the Linear A script
  - Predecessor remains undeciphered
- Current consensus that Linear A was used to write a language unrelated to Greek
- Decipherment took more than 70 years
- Groundwork by Alice Kober, full decipherment by Michael Ventris and John Chadwick [Ventris53]

A Short History of Linear B

- Construction of combinatoric sign grids
- Assumption that inflection usually happens in the last syllable and won't change consonant
- Creation of an internally consistent grammar
- Guessing a single word and checking if everything else falls into place (a-mi-ni-so → Amnissos)
- Understanding is still not complete

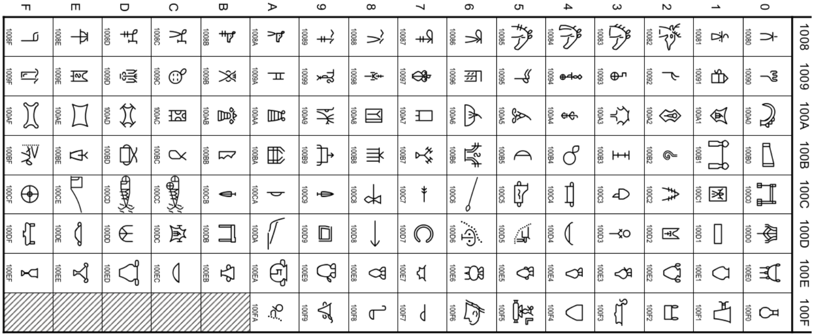
A Short History of Linear B



a a$_2$	e	i	o	u
ai				
ja	je		jo	
wa	we	wi	wo	
da	de	di	do	da$_2$
ka	ke	ki	ko	ku
ma	me	mi	mo	
na	ne	ni	no	nu nu$_2$?
pa pa$_2$?	pe	pi	po	pu
	qe	qi	qo qo$_2$?	
ra ra$_2$	re	ri	ro ro$_2$	ru
sa	se	si	so	
ta ta$_2$?	te pte	ti	to	tu
	z?e		z?o z?o$_2$	

Figure: Syllabary [Ventris53]

A Short History of Linear B



	1008	1009	100A	100B	100C	100D	100E	100F
0	𐂀 10080	𐂐 10090	𐂠 100A0	𐂰 100B0	𐃀 100C0	𐃐 100D0	𐃠 100E0	𐃰 100F0
1	𐂁 10081	𐂑 10091	𐂡 100A1	𐂱 100B1	𐃁 100C1	𐃑 100D1	𐃡 100E1	𐃱 100F1
2	𐂂 10082	𐂒 10092	𐂢 100A2	𐂲 100B2	𐃂 100C2	𐃒 100D2	𐃢 100E2	𐃲 100F2
3	𐂃 10083	𐂓 10093	𐂣 100A3	𐂳 100B3	𐃃 100C3	𐃓 100D3	𐃣 100E3	𐃳 100F3
4	𐂄 10084	𐂔 10094	𐂤 100A4	𐂴 100B4	𐃄 100C4	𐃔 100D4	𐃤 100E4	𐃴 100F4
5	𐂅 10085	𐂕 10095	𐂥 100A5	𐂵 100B5	𐃅 100C5	𐃕 100D5	𐃥 100E5	𐃵 100F5
6	𐂆 10086	𐂖 10096	𐂦 100A6	𐂶 100B6	𐃆 100C6	𐃖 100D6	𐃦 100E6	𐃶 100F6
7	𐂇 10087	𐂗 10097	𐂧 100A7	𐂷 100B7	𐃇 100C7	𐃗 100D7	𐃧 100E7	𐃷 100F7
8	𐂈 10088	𐂘 10098	𐂨 100A8	𐂸 100B8	𐃈 100C8	𐃘 100D8	𐃨 100E8	𐃸 100F8
9	𐂉 10089	𐂙 10099	𐂩 100A9	𐂹 100B9	𐃉 100C9	𐃙 100D9	𐃩 100E9	𐃹 100F9
A	𐂊 1008A	𐂚 1009A	𐂪 100AA	𐂺 100BA	𐃊 100CA	𐃚 100DA	𐃪 100EA	𐃺 100FA
B	𐂋 1008B	𐂛 1009B	𐂫 100AB	𐂻 100BB	𐃋 100CB	𐃛 100DB	𐃫 100EB	
C	𐂌 1008C	𐂜 1009C	𐂬 100AC	𐂼 100BC	𐃌 100CC	𐃜 100DC	𐃬 100EC	
D	𐂍 1008D	𐂝 1009D	𐂭 100AD	𐂽 100BD	𐃍 100CD	𐃝 100DD	𐃭 100ED	
E	𐂎 1008E	𐂞 1009E	𐂮 100AE	𐂾 100BE	𐃎 100CE	𐃞 100DE	𐃮 100EE	
F	𐂏 1008F	𐂟 1009F	𐂯 100AF	𐂿 100BF	𐃏 100CF	𐃟 100DF	𐃯 100EF	

Figure: Ideograms [Unicode13]

Methodology

- Syllabary for an unrelated language
- Sign sequences are phonetic approximations of Mycenaean words
- Not able to express all phonetic features of Greek
  - Consonant clusters
  - Ending consonants
  - Diphthongs
- Scribes created various hacks to circumvent script limitations
- Possibly weaknesses in the assigned phonetic values
- Possibly two very similar dialects

Methodology - Ambiguity and Irregularities

- po-me - ποιμήν - shepherd
    - Dative singular: po-me-ne - ποιμήνέι
    - Nominative singular: po-me-ne - ποιμήνές
- i-qo → ίππος
- i-ko → ιχos or ικos or ισκos
- No syntax to aid reconstruction of intended declension

Methodology - Preprocessing

- Almost no preprocessing necessary
- Scribes were helpful
    - Comma between words
    - Signs on neatly drawn lines
    - No hyphenation
    - Medium encouraged fixing mistakes
- Considerable effort invested into correctly identifying signs
- Lack of syntax → One to one correspondence of lines to sentence-ish constructs

Methodology - First approaches

- Approximate matching using n-grams
- Create dictionary of Mycenaean words
- Run spell checker on sign sequences
- Produces nonsensical suggestions
  - Phonetic gap too large

Methodology

- Stochastic language models won't work
- Average text length 14 signs
- Roughly 150 documents exceed 50 signs (not words!)
- Longest text around 600 signs

Methodology

- Tried-and-True rule based analyzers
- Good track records with other Greek dialects [Crane91]
- Two step process is reduced to single problem
- Easily adapted to state-of-the-art knowledge
    - ...and limited by it

Methodology

- Drawback: Have to assemble rule set
    - Even worse: there are multiple to choose from
    - Worst of all: all are written in different languages
    - Collaborator agreed to assist

Methodology - Basic Script Rules

- p → p^h, b, p
- k → k^h, g, k
- r → r, l
- q → k^w, g^w
- Syllables with initial consonant clusters are written with signs matching the vowel of the syllable
- Ending consonants are omitted (ti-ri-po-de → tripodes)
- Much more complex ones possible (some site specific [Hooker80])

Methodology - Measurement

- Measurement of success/failure quite complex
- No training set to verify against
- Did build Ad-Hoc references from literature for ngram similarity test
- Have to rely on subjective impressions from collaborators

Outlook

- Assemble rule set
- Convert it into a sensible format
    - Using endless chains of regular expressions is not satisfactory
    - In an ideal world rule set and source code is kept apart
- Functional Morphology sounds appropriate

Outlook

- Plan to have robust results until finals
- Maybe incorporate the ngram similarity algorithm as a postprocessing step
- Try to extend tool's application to more than annotation assistance
  - If output is stable across known similar texts look for notably worse results
- Unfortunately, irreducibly complex

Propaganda

- Bring civili...computer science to the barbarians
- It hurts to see how they do things
- Find/Show appropriate tools for working with small and barely understood information systems
- Hopefully, won't need constant funding (and maintenance) to be of use

References

- Evidence for Greek Dialect in the Mycenaean Archives, Ventris and Chadwick, The Journal of Hellenic Studies, Vol. 73 (1953), pp. 84-103
- Generating and parsing classical Greek, Crane, Literary and Linguistic Computing 6.4 (1991) pp. 243-245
- The Unicode Standard 6.3
- Linear B: An Introduction, Hooker, Bristol Classical Press, 1980