WHAT’S IN A NAME? ISSUES IN NAMED ENTITY RECOGNITION

GOALS FOR PRESENTATION

• Introduce OSU’s *Herodotos Project* as it pertains to onomastics

• Discuss issues project has faced that are of a general, somewhat theoretical, nature for onomastic research
THE HERODOTOS PROJECT

Basic idea of this “digital humanities” project: compiling all known information about ancient peoples (working from Greek/Latin view of the ancient world)

→ Thus, the “Herodotos Project” (based on Herodotos being the “father of history”)
THE HERODOTOS PROJECT

• An NEH-funded portion of the project is aiming first to develop as complete a catalogue as possible of ethnonyms (representing “ethnic” groups, with all the inherent pitfalls in mind of such a designation) recognized by the ancients in classical sources.
THE HERODOTOS PROJECT

• We then plan to use the catalogue of group-names to create a virtual “museum” of informational “pages” populated with available material on these groups (regarding location, origins, language, art, genetics, customs, mythologies, religion, sources on them, etc.)
The big obstacle:

How to develop such a catalogue?
**First attempt:** comb through ancient (Classical Greek and Latin) sources manually to pick out group names

- easy, right? Wrong!
--time-consuming if done this way
--so, how to proceed efficiently?
Computational Linguistics to the Rescue!

• this is actually a “Named Entity Recognition” (NER) task, an issue which computational linguistics has tackled

• NER = a machine learning technique for automatically identifying “named entities” in a text or discourse, where NAMED means a referential entity using a name and ENTITY here simply means something that occurs in a text
Computational Linguistics to the Rescue!

- NER requires training on large amounts of text in which humans have manually, accurately, and consistently identified the named entities of interest.

- then, after training on the annotated text, the computational algorithm takes over and picks out the named entities of interest in non-annotated text
How we annotate:

Personal Names /PRS
Group Names /GRP
Geographic Locations /GEO
Null /0

*BIO style annotation (Beginning, Inside, Outside)
ONOMASTIC ISSUES:
What IS a name?

- Consistency or Context
- Ambiguity
- Morphology
I. Ultimately, every noun can be construed as a name.

ex. forum/0
    forum/GEOB Romanum/GEOI

    mons/0
    mons/GEOB Vesuvius/GEOI
II. Many group names refer or relate to place names; others, to personal names
Three cases of ambiguity in name-identification (see Appendix for original passage in Herodotos).

A) Delphians
B) Lydians
C) The Pythia
A) Delphians

• Same word used for the geographic place and the inhabitants of that place

Liddell & Scott, *Greek-English Dictionary*
Sometimes, Preposition Determines Annotation

• **Agency = GRP**

  ὑπὸ Δελφῶν καλέεται Γυγάδας (Hdt.1.14)

  (it) is called “Gygian” by the Delphians.

• **Location = GEO**

  ἐν δὲ Δελφοῖσι ὡς ἐσῆλθον τάχιστα ἐς τὸ μέγαρον
  (Hdt.1.47)

  But at Delphi, as soon as they went into the hall...

• **Ambiguous = ??**

  ὁ δὲ ὡς τὸ ἐκ Δελφῶν ἥκουσε (Hdt.1.48)

  But when he heard (this) from the Delphians.
B) The Lydians

- Singular (ὁ Λυδός) is commonly used to refer to Croesus “The Lydian”
- Plural (οἱ Λυδοὶ) is not multiple Croesuses, but is used to refer to “The Lydians” as a group.

ex. ὡς ἐσῆλθον τάχιστα ἐς τὸ μέγαρον οἱ Λυδοὶ χρησόμενοι τῷ θεῷ καἰ ἐπειρώτων τὸ ἐντεταλμένον
οἱ Λυδοὶ
C) The Pythia

- ἡ Πυθίη = “The Pythia” = the oracle
- Like Croesus “The Lydian”?
- No. It is (as far as we can see) only ever used to refer to the oracle at Delphi.
“The Pythia”

ἡ Πυθίη ἐν ἑξαμέτρῳ τόνω λέγει τάδε. (Hdt. 1.47)

ταῦτα οἱ Λυδοὶ θεσπισάσης τῆς Πυθίης συγγραψάμενοι οἴχοντο ἀπιόντες ἐς τὰς Σάρδις. (Hdt. 1.48)
Authorial Intention and Tone?

Pliny the Elder *Natural History* 2.16.6

“cum singuli quoque ex semet ipsis totidem deos faciant *lunones* Geniosque adoptando sibi, gentes vero quaedam animalia et aliqua etiam obscura pro dis habeant ac multa dictu magis pudenda”
lunones
BUT….. another complexity:

Pliny the Elder *Natural History* 7.47.1

“Auspiciatus enecta parente gignuntur, sicut *Scipio Africanus* prior natus primusque Caesarum a caeso matris utero dictus, qua de causa et Caesones appellati.”
Scipio Africanus

- We are not creating a prosopography
- Mining names for group names
Take-aways….

- Context drives decisions
- Robustness of NER
- Value of human readers
Thank you!
ὁ τι μέν νυν τὰ λοιπὰ τῶν χρηστηρίων ἐθέσπισε, οὓς λέγεται πρὸς οὐδαμῶν: ἐν δὲ Δελφοῖς ώς ἐσῆλθον τάχιστα ἐς τὸ μέγαρον οἱ Λυδοὶ χρησόμενοι τῷ θεῷ καὶ ἐπειρώτων τὸ ἐντεταλμένον, ἡ Πυθίη ἐν ἑξαμέτρῳ τόνω λέγει τάδε. [3]

“οἶδα δ’ ἐγὼ ψάμμου τ’ ἀριθμὸν καὶ μέτρα θαλάσσης, καὶ κωφοῦ συνίημι, καὶ οὐ φωνεύντος ἀκούω. οὐδεὶς μ’ ἐς φρένας ἦλθε κραταιρίνοιο χελώνης ἐφομένης ἐν χαλκῷ ἀμ’ ἀρνείσθη κρέεσσιν, ἢ χαλκός μὲν υπέστρωται, χαλκόν δ’ ἐπιέσται.