LANGUAGE MODELLING FOR AUTHORSHIP ATTRIBUTION IN HOMERIC TEXTS

FASOI MARIA
POSTGRADUATE STUDENT

MSc DIGITAL METHODS FOR THE HUMANITIES
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SUPERVISOR: J. PAVLOPOULOS

CO-SUPERVISOR : M. KONSTANTINIDOU

NOVEMBER 2020

INTRODUCTION

- **Question**
  - Authorship Attribution
- **Where?**
  - Homeric Texts
    - Iliad
    - Odyssey
    - 4/33 Homeric hymns
- **How?**
  - Statistical Language Models – SLM
  - Long Short-Term Memory – LSTM
- **Why?**
  - Linguistic affinity of rhapsodies & hymns with Iliad/Odyssey
  - Classification of Odyssey/Iliad excerpts from:
    - Language models
    - Human annotators via questionnaire

QUESTIONS

1. Are there rhapsodies in the Iliad and the Odyssey respectively that show more linguistic affinity with the whole of the respective epic?
2. Are there rhapsodies in the Iliad and the Odyssey that deviate from the linguistic style of the Homeric epics?
3. How linguistically similar are the Homeric hymns: "To Apollo", "To Aphrodite", "To Demeter" and "To Hermes" in Homeric epics?
4. Can artificial language models categorize excerpts from the Iliad and the Odyssey into the respective epic more successfully than the human interpretation?

HOMERIC EPICS

> *«Dealing with the Homeric question since the time of Friedrich August Wolf can be described as the most controversial chapter of literary research.» Albin Lesky*

HOMER?

ILIAD

ΙΛΙΑΣ

Μῆνιν ἄειδε, θεὰ, Πηληιάδεω Ἀχιλῆος
οὐλομένην, ἣ μυρί' Ἀχαιοῖς ἄλγε' ἔθηκε,
πολλὰς δ' ἰφθίμους ψυχὰς Ἄϊδι προΐαψεν
ἡρώων, αὐτοὺς δὲ ἑλώρια τεῦχε κύνεσσιν
οἰωνοῖσί τε πᾶσι· Διὸς δ' ἐτελείετο βουλή·
ἐξ οὗ δὴ τὰ πρῶτα διαστήτην ἐρίσαντε
Ἀτρεΐδης τε ἄναξ ἀνδρῶν καὶ δῖος Ἀχιλλεύς.

ODYSSEY

ΟΔΥΣΣΕΙΑ

Ἄνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον, ὃς μάλα πολλὰ
πλάγχθη, ἐπεὶ Τροίης ἱερὸν πτολίεθρον ἔπερσε·
πολλῶν δ' ἀνθρώπων ἴδεν ἄστεα καὶ νόον ἔγνω,
πολλὰ δ' ὅ γ' ἐν πόντῳ πάθεν ἄλγεα ὃν κατὰ θυμόν,
ἀρνύμενος ἥν τε ψυχὴν καὶ νόστον ἑταίρων.
ἀλλ' οὐδ' ὣς ἑτάρους ἐρρύσατο, ἱέμενός περ·
αὐτῶν γὰρ σφετέρῃσιν ἀτασθαλίῃσιν ὄλοντο,
νήπιοι, οἳ κατὰ βοῦς Ὑπερίονος Ἠελίοιο
ἤσθιον· αὐτὰρ ὁ τοῖσιν ἀφείλετο νόστιμον ἦμαρ.

Why Homeric epics?

- Object of deep reflection since antiquity.
- Homeric question (19th c.)
  - Existence of the poet Homer and the authorship of the epics (Latacz, 2000).
  - Composition of epics: performed by one or more composers (Latacz, 2000).
  - In the 20th c. Many great works on Homer and new translations of Homeric epics were published.
  - The Homeric question has not been resolved to date.

HOMERIC HYMNS

HOMER?

Homeric Hymns

- Why Homeric hymns?
  - In antiquity many works are attributed to Homer including Homeric hymns(Latacz, 2000).
  - Alexandrian philologists seem to have removed the collection from the poet's overall work (Morris & Powell, 1997).

HOMERIC EPICS AND HOMERIC HYMNS

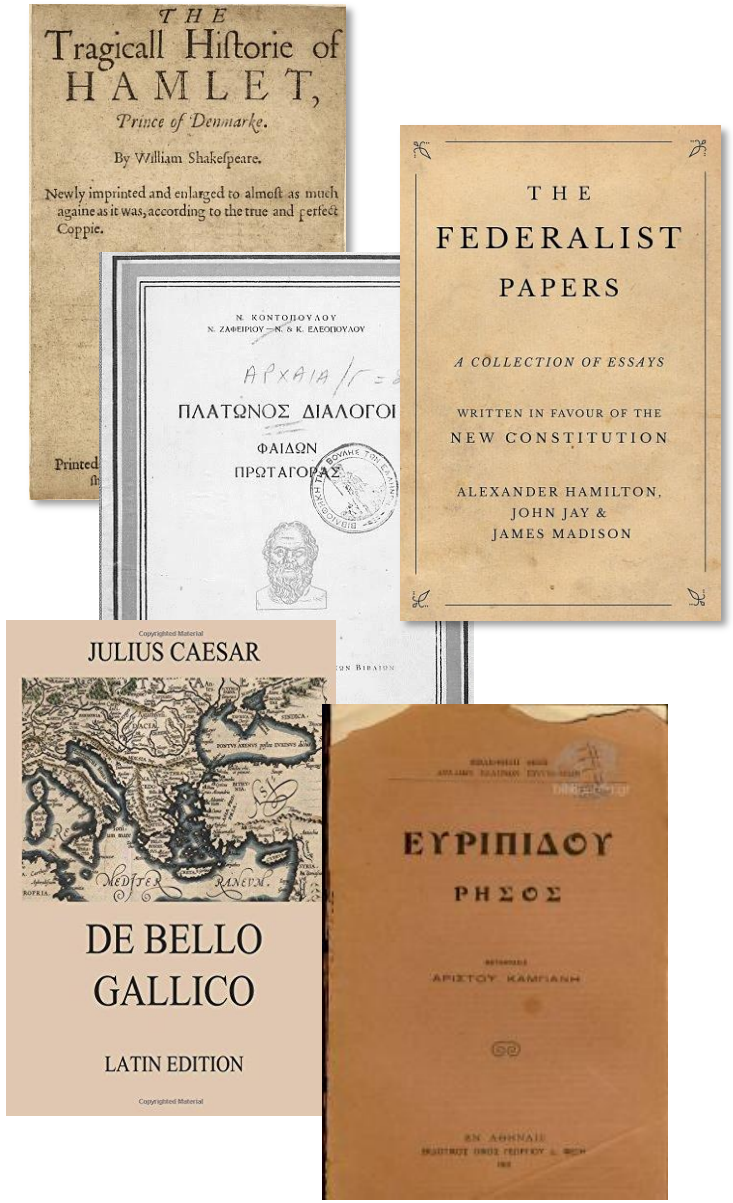
- Metric poems in dactylic hexameter
- Around the 8th c. BC, the composition of the Iliad
- Later with some time interval the composition of the Odyssey
- Most Homeric hymns were composed during the Archaic period (6th-7th c. BC)
- Some Homeric hymns are considered works of the Hellenistic period (323-30 BC)

-UU / -UU / -UU / -UU / -UU / --

- = a long syllable | U = a short syllable

AUTHORSHIP ATTRIBUTION

- What is?
  - Issue of recognition of the author of an anonymous text or text whose paternity is disputed (Love, 2002)
- History flashback
  - 18th c. William Shakespeare
  - 19th c. Platonic dialogues
  - 20th c. Federalistic Papers
- Researches of 21st c.
  - «*Commentarii de Bello Gallico*» Julius Ceasar (Kestemont et al., 2016)
  - «*Rhesus*» Euripides (Manousakis & Stamatatos, 2018)

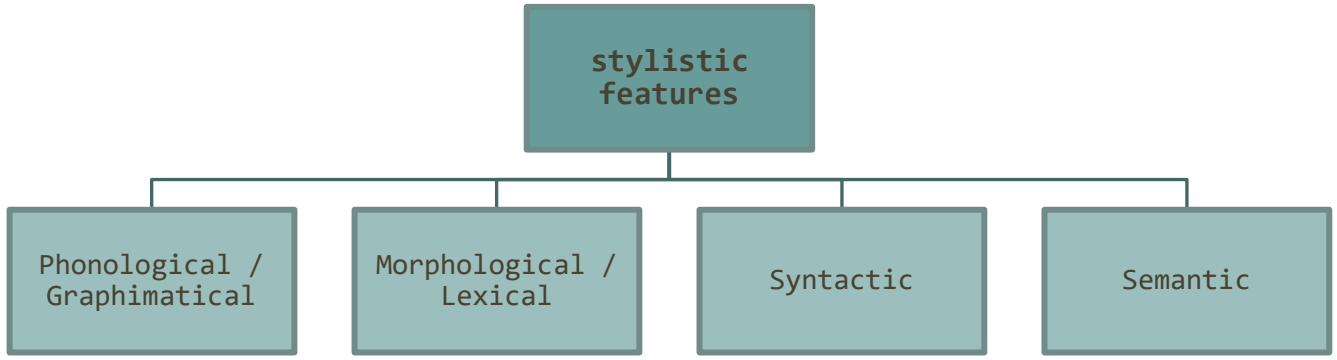


AUTHORSHIP ATTRIBUTION

- Fields of application
  - Plagiarism detection (Kimler, 2003)
  - Forensic Investigation (Chaski, 2005)
  - Phishing (Gollub et al., 2013)

- How is it determined?
  - «stylistic features»

    (Stamatatos, 2009)



STATISTICAL LANGUAGE MODELS (SLM)

- Statistical Language Models (SLM)
  - distribution of probabilities in word sequences
    - When?
      - When the context is known (Jurafsky & Martin, 2000)

"The color of lavender is purpl___":

P("e" | "The color of lavender is purpl__") = ;

  - character n-grams (Peng et al., 2003)
    - sequences of tokens [token = word, character, etc.]
    - n = number of elements
      - n=1,2,3...

Μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην

n=3

[μῆν], [ῆνι], [νιν], [ιν_], [ν_ἄ], [_ἄε], [ἄει], [ειδ], [ιδε], [δε_], [ε_θ], [_θε], [θεὰ], [εὰ_], [ὰ_π], [_πη], [πηλ], [ηλη], [ληϊ], [ηϊά], [ϊάδ], [άδε], [δεω], [εω_], [ω_ἀ], [_ἀχ], [ἀχι], [χιλ], [ιλῆ], [λῆο], [ῆος], [ος_], [ς_ο], [_οὐ], [οὐλ], [ὐλο], [λομ], [ομέ], [μέν], [ένη], [νην]

NEURAL LANGUAGE MODELS
LONG SHORT-TERM MEMORY (LSTM)

- Long Short-Term Memory (LSTM)
  - Recurrent Neural Network (RNN)
  - long duration of memory

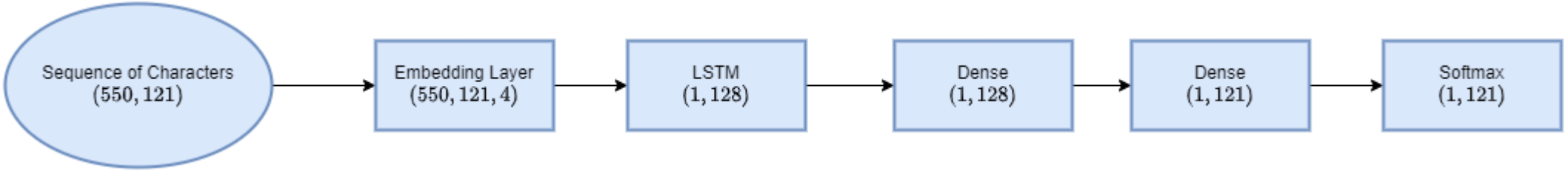


Figure 1. LSTM architecture implemented

QUESTIONNAIRE FOR THE CLASSIFICATION

- 23 Human annotators – Philologists
  - 5 undergraduate students of Philology
  - 3 graduates of Philology
  - 10 postgraduate students with undergraduate studies of Philology
  - 5 Teachers of Philology in secondary education
- 20 excerpts
  - 6 Iliad
  - 6 Odyssey
  - 2 To Apollo
  - 2 To Aphrodite
  - 2 To Demeter
  - 2 To Hermes

They were not classified by language models

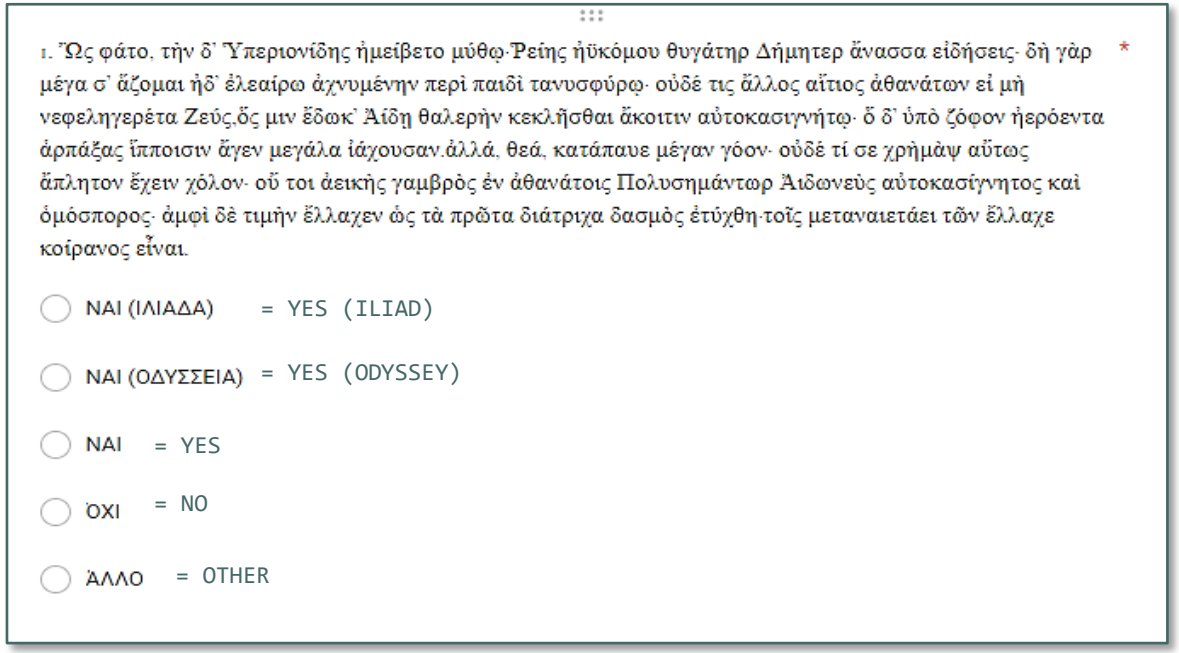
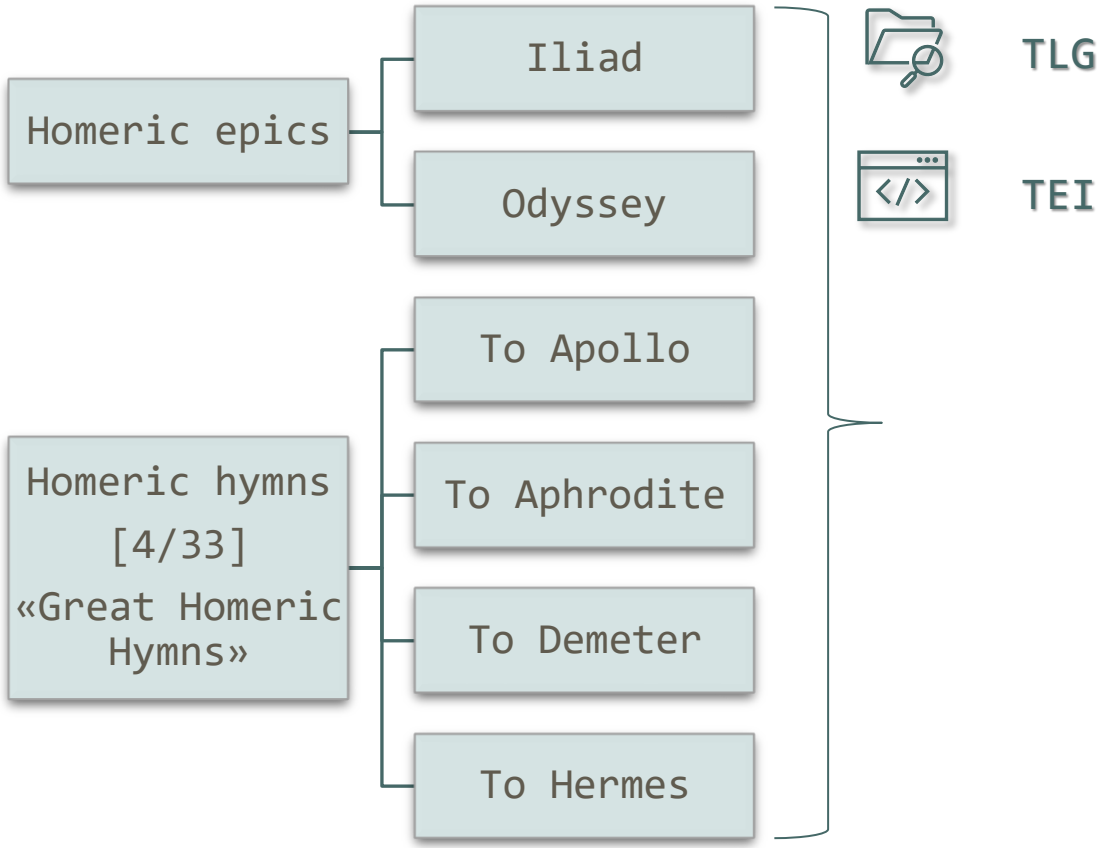


Image 1. Indicative example of an excerpt from the Homeric hymn "To Demeter"

✓ *The questionnaire which created for this Master thesis can be found at the following link : https://forms.gle/DA11AMQq4iRh2bx99*

DATA PREPROCESSING



The accents and diacritics?

1. Part of the spelling
2. An integral part of the metric rules

1. Subtraction: punctuation
   editorial symbols
2. Conversion : lowercase

121 characters
(120 + space)

ϊ ά έ ή ί ϋ α β γ δ ε ζ η θ ι κ λ μ ν ξ ο π ρ ς σ τ υ
φ χ ψ ω ΐ ΰ ό ύ ώ ὰ ά ᾶ ὰ ἄ ἅ ἆ ὲ έ ὲ ὲ ἔ ἕ ὴ ή ὴ ὴ ἤ
ἥ ἦ ἧ ὶ ί ῖ ὶ ἴ ἵ ἶ ἷ ὸ ό ὸ ὄ ὅ ὺ ύ ὺ ὺ ὕ ὖ ὗ ὼ ώ ὼ ὼ
ὤ ὥ ὦ ὧ ὰ ὲ ὴ ὶ ὸ ὺ ὼ ᾐ ᾑ ᾒ ᾓ ᾔ ᾕ ᾠ ᾡ ᾢ ᾣ ᾰ ᾱ ᾲ ᾳ ᾴ ᾗ
ῆ ῃ ΐ ῖ ῗ ῦ ῤ ῥ ῦ ῳ ῴ ῶ ῷ

M.L. WEST, HOMERI ILIAS [BIBLIOTHECA SCRIPTORUM GRAECORUM ET ROMANORUM TEUBNERIANA]. STUTTGART-LEIPZIG: TEUBNER.
P. VON DER MÜHLL, HOMERI ODYSSEA, BASEL: HELBING & LICHTENHAHN, 1962: 1-456.
T.W. ALLEN, W.R. HALLIDAY, AND E.E. SIKES, THE HOMERIC HYMNS, 2ND EDN., OXFORD: CLARENDON PRESS, 1936.

APPROACHES

	1	2
○ Experiments	**Authorship attribution with SLM**	**Classification of Homeric texts with SLM & LSTM**
○ Categories of Language Models	I. Cyclical SLM's II. Cross SLM's	I. SLM-based classification II. LSTM-based classification III. Comparison of the category I. & II. with excerpts from the questionnaire, given to human-annotators
○ Evaluation measurements	• Perplexity	• F1-score

AUTHORSHIP ATTRIBUTION WITH CYCLICAL SLM'S (CROSS VALIDATION)

- **24 IliadSLM & 24 OdysseySLM** (character-lvl)
  - Each SLM was trained in a different subset of rhapsodies crowd number of 23.
  - Each SLM was scored with the remaining one 24th rhapsody of the respective epic.
- TRAINING SET
  - 9.600 characters < beginning of each rhapsody
- RATED TEXTS
  - 20 random samples × 600 characters [*Bootstrapping*]

1. **Perplexity**
   - The most common evaluation measurement of a language model
2. **Confidence Intervals**
   - Safer estimate of a parameter of a population based on a random sample of that population

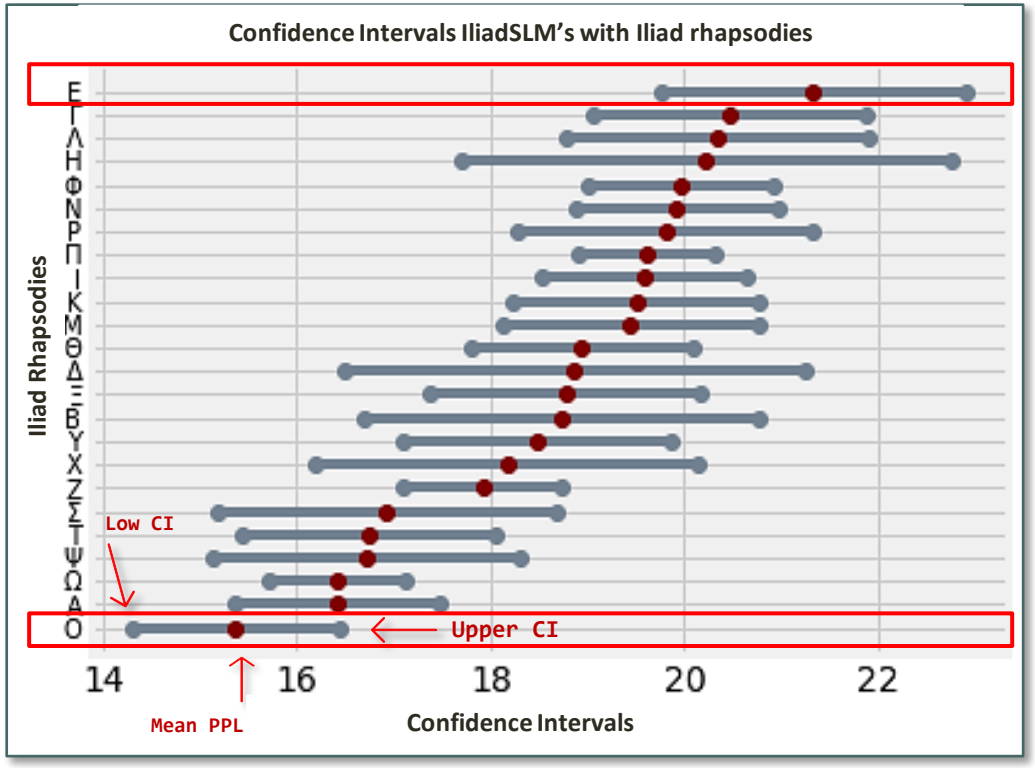
Algorithm 1: SLM application in Iliad and Odyssey (circular)

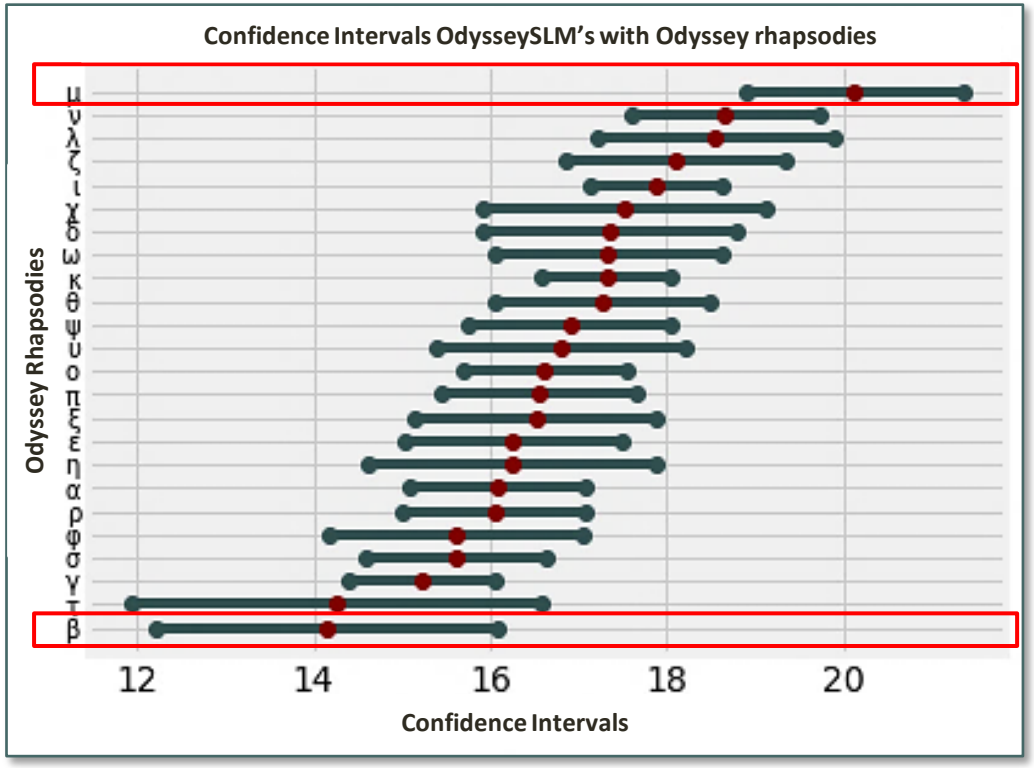
The algorithm can be generalized to any language model, whether statistical or neural.

```
Initialize Iliad_models as a list of 24 n-gram models where n = 3
Initialize Odyssey_models as a list of 24 n-gram models where n = 3

Initialize Iliad_train as a list of all rhapsodies
Initialize Odyssey_train as a list of all rhapsodies

1. repeat from m=0 to m= 24, step=1
2.     repeat from r=0 to r= 24, step=1
3.         if m is equal to r:
4.             continue
5.         Train Iliad_models[m] using samples from Iliad_train[r]
6.         Train Odyssey_models[m] using samples from Odyssey_train[r]
```

AUTHORSHIP ATTRIBUTION WITH CYCLICAL SLM'S (CROSS VALIDATION)

Homeric epics





AUTHORSHIP ATTRIBUTION WITH SLM
CROSS MODELS

- **1 IliadSLM & 1 OdysseySLM** (character-lvl)
  - The IliadSLM trained in the 24 rhapsodies of the Iliad and was graded with each rhapsody of the Odyssey.
  - The OdysseySLM trained in the 24 rhapsodies of the Odyssey and graded with each rhapsody of the Iliad.
- TRAINING SET
  - 9.600 characters < beginning of each rhapsody
- RATED TEXTS
  - 20 random samples × 600 characters [*Bootstrapping*]

1. **Perplexity**
2. **Confidence Intervals**

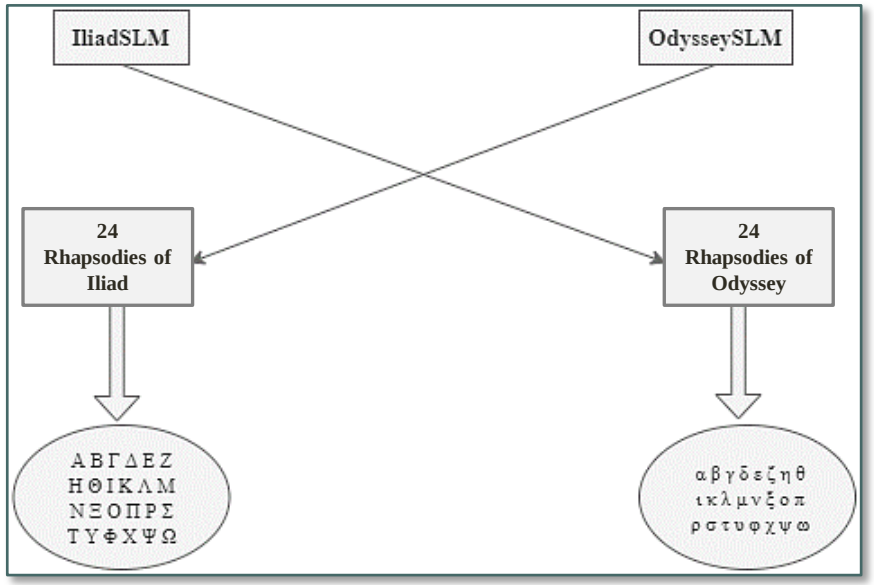
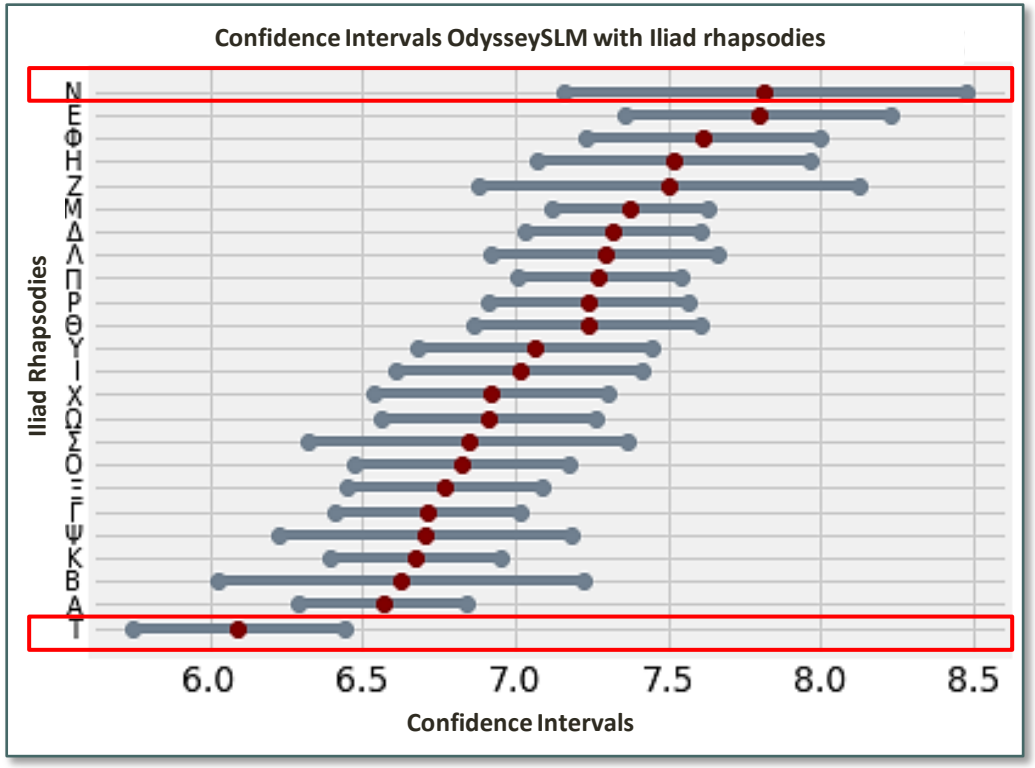
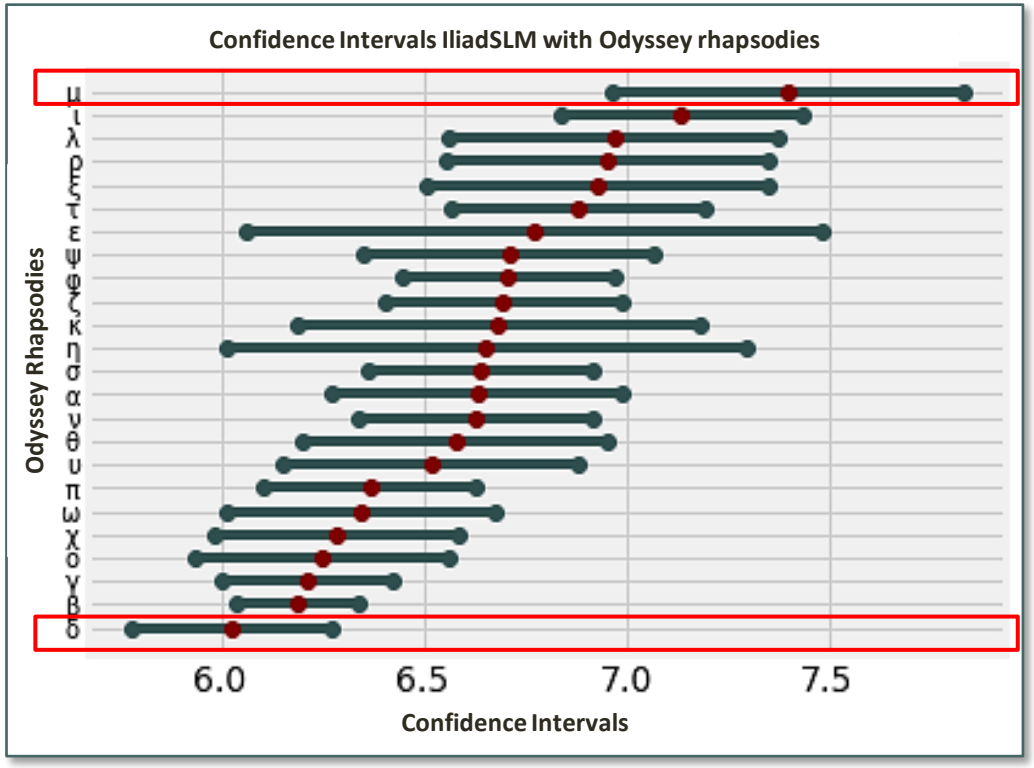


Figure 2. Cross evaluation

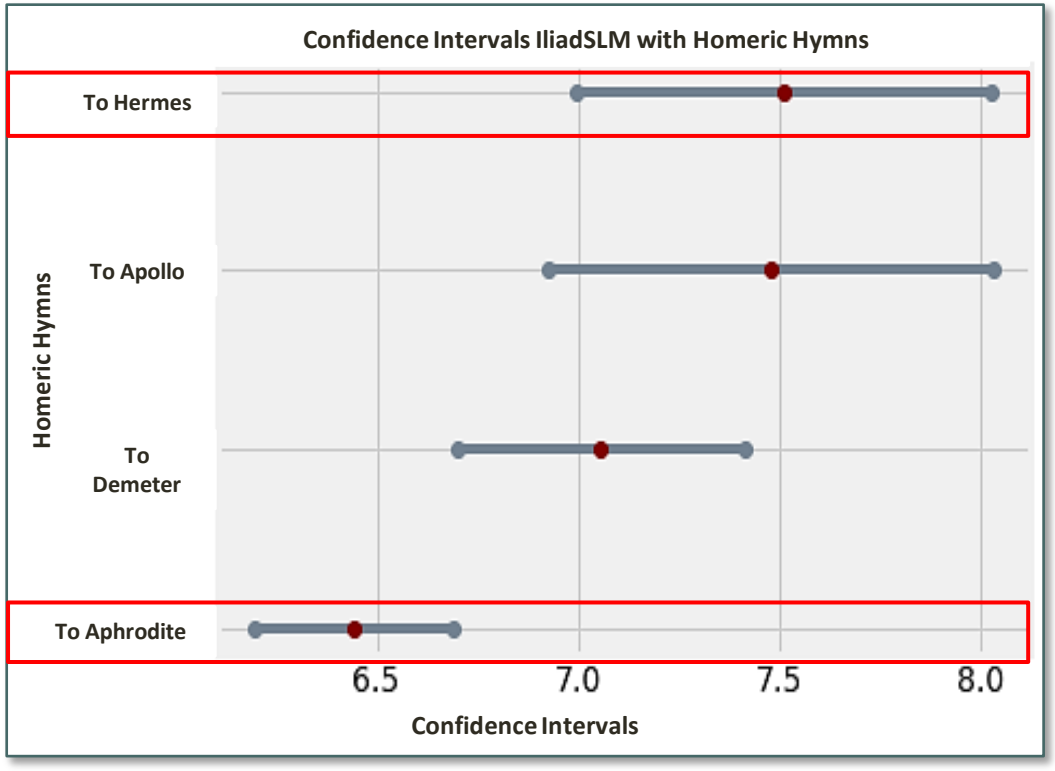
AUTHORSHIP ATTRIBUTION WITH SLM
CROSS MODELS

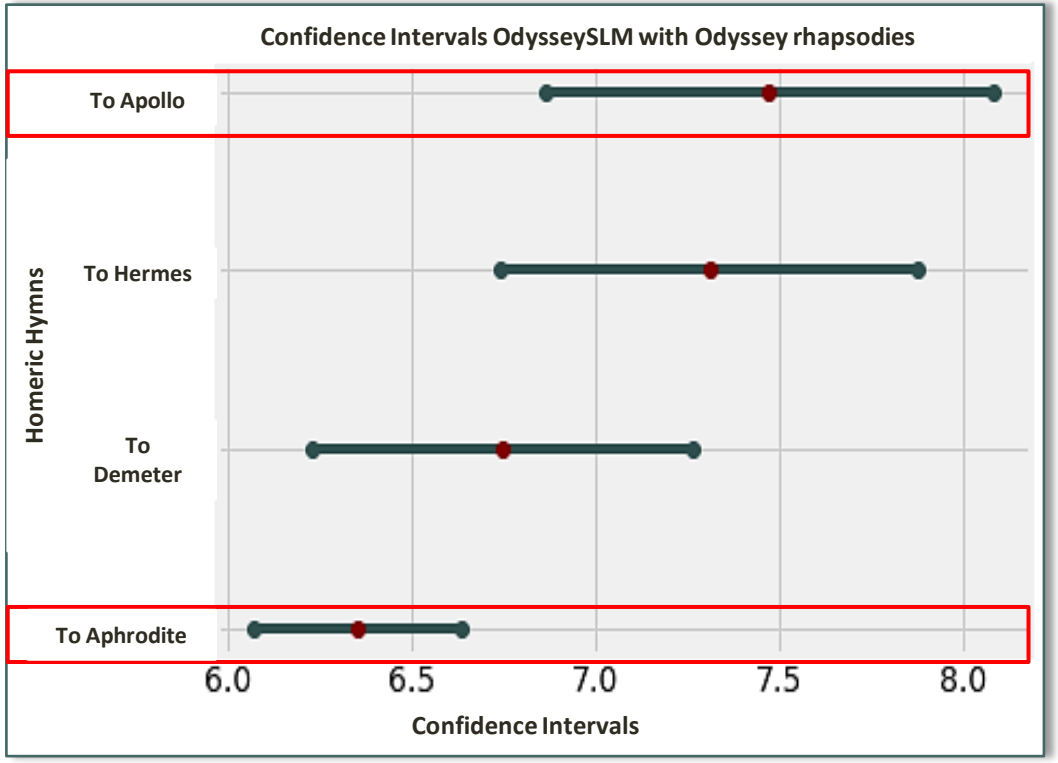




AUTHORSHIP ATTRIBUTION WITH SLM

Homeric Hymns

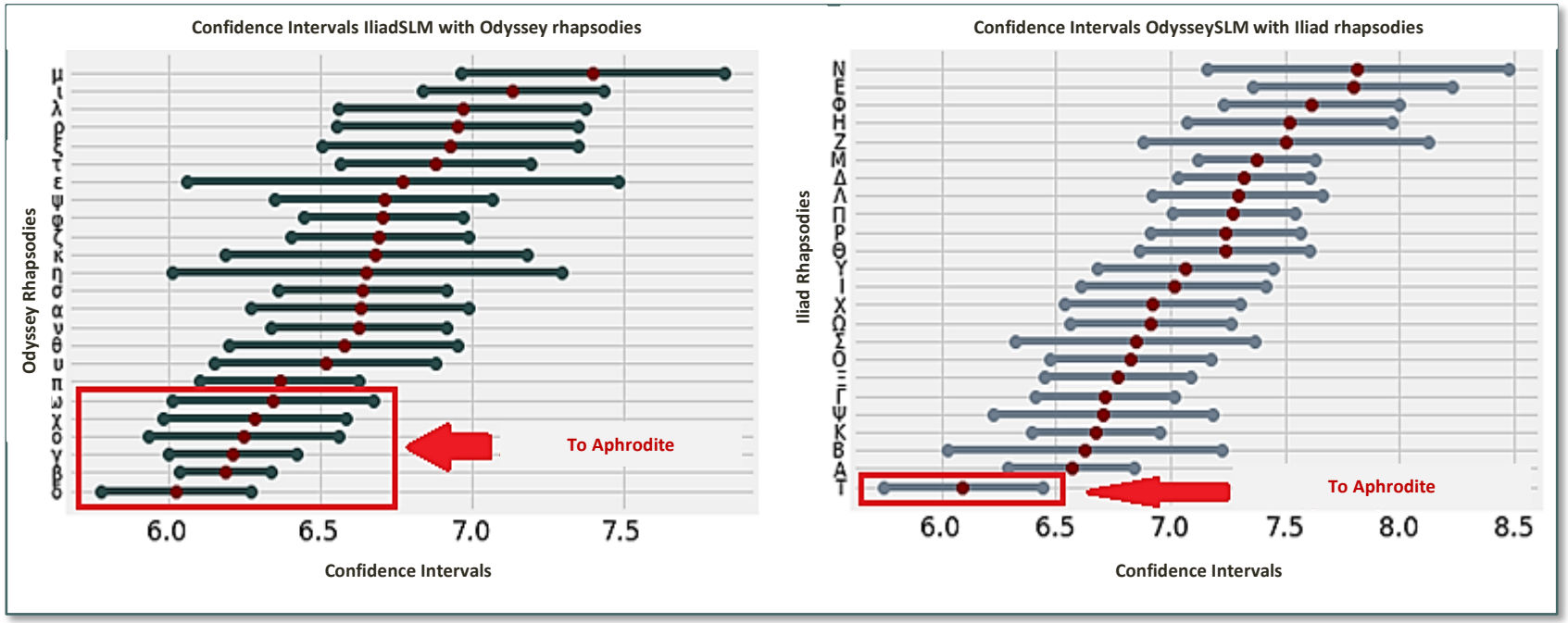




DISCUSSION ON AUTHORSHIP ATTRIBUTION WITH SLM

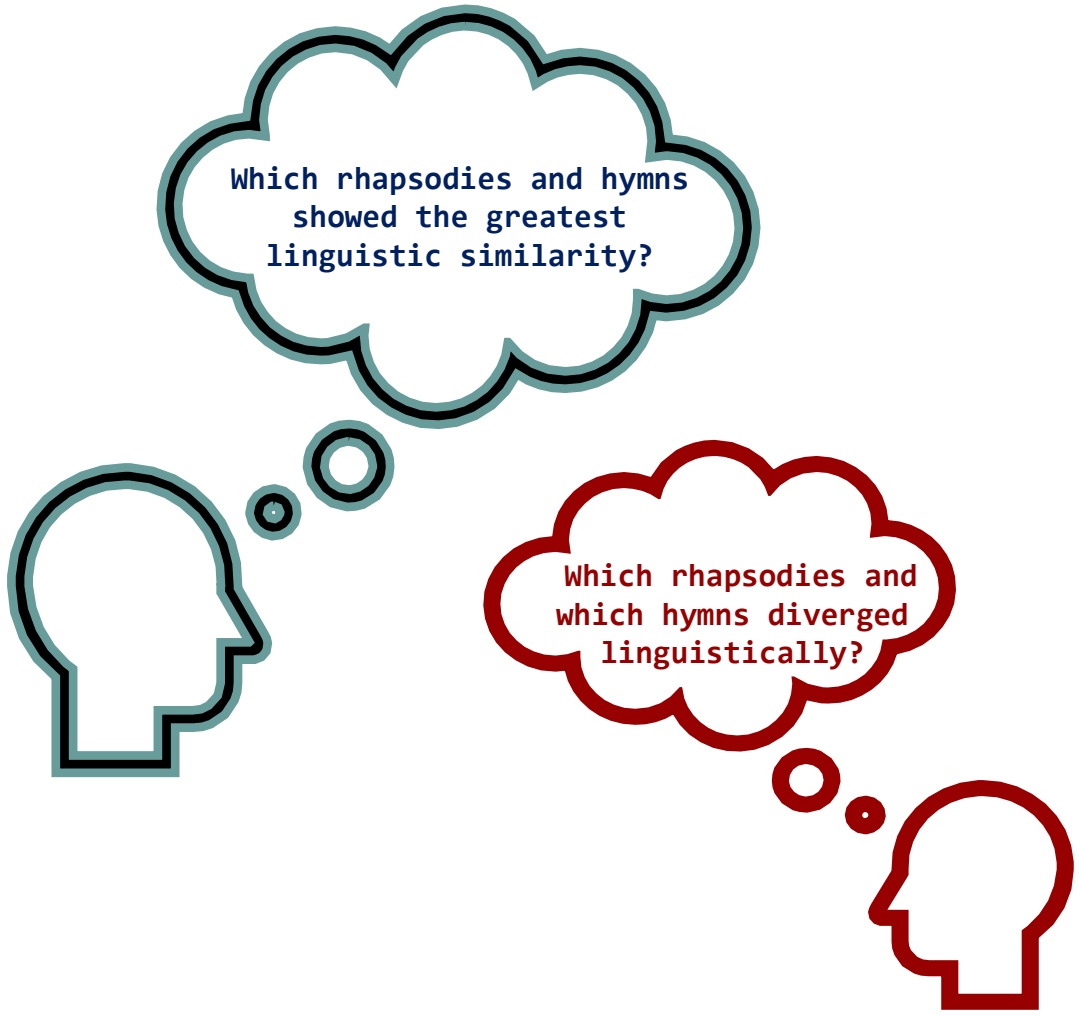
The hymn "To Aphrodite" is often considered by scholars to be the most Homeric of all the other hymns, because it is close to the Homeric epics in terms of poetic language, style and theme.

(Morris και Powell, 1997)



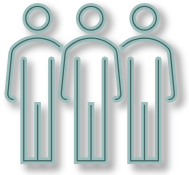
DISCUSSION ON AUTHORSHIP ATTRIBUTION WITH SLM

- **Cyclical IliadSLM's**
  - Rhapsody **"O" of Iliad**
  - Deviated **"E" of Iliad**
- **Cyclical OdysseySLM's**
  - Rhapsody **"β" of Odyssey**
  - Deviated **"μ" of Odyssey**
- **Iliad Cross model**
  - Rhapsody **"δ" of Odyssey**
  - Deviated **"μ" of Odyssey**
  - Homeric Hymn **"To Aphrodite"**
  - Deviated **"To Hermes"**
- **Odyssey cross model**
  - Rhapsody **"T" of Iliad**
  - Deviated **"N" of Iliad**
  - Homeric Hymn **"To Aphrodite"**
  - Deviated **"To Apollo"**



CLASSIFICATION OF HOMERIC TEXTS WITH SLM AND LSTM
COMPARISON OF LANGUAGE MODELS WITH HUMAN-ANNOTATORS

vs.



Language systems

SLM & LSTM

Human-annotators

CLASSIFICATION OF HOMERIC TEXTS WITH SLM AND LSTM

- **2 SLM & 2 LSTM** (character-lvl)
  - 1 IliadSLM & 1 OdysseySLM
  - 1 IliadLSTM & 1 OdysseyLSTM

- TRAINING SET
  - 9.600 characters < beginning of each rhapsody
  - removal of rhapsody of the questionnaire
    - 6 rhapsodies from Iliad [“Γ”, “Θ”, “Κ”, “Λ”, “Ο” & “Φ”]
    - 6 rhapsodies from Odyssey [“γ”, “ζ”, “κ”, “ν”, “σ” & “φ”]

- EVALUATED TEXTS
  - 6 Iliad excerpts
  - 6 Odyssey excerpts

  } Questionnaire

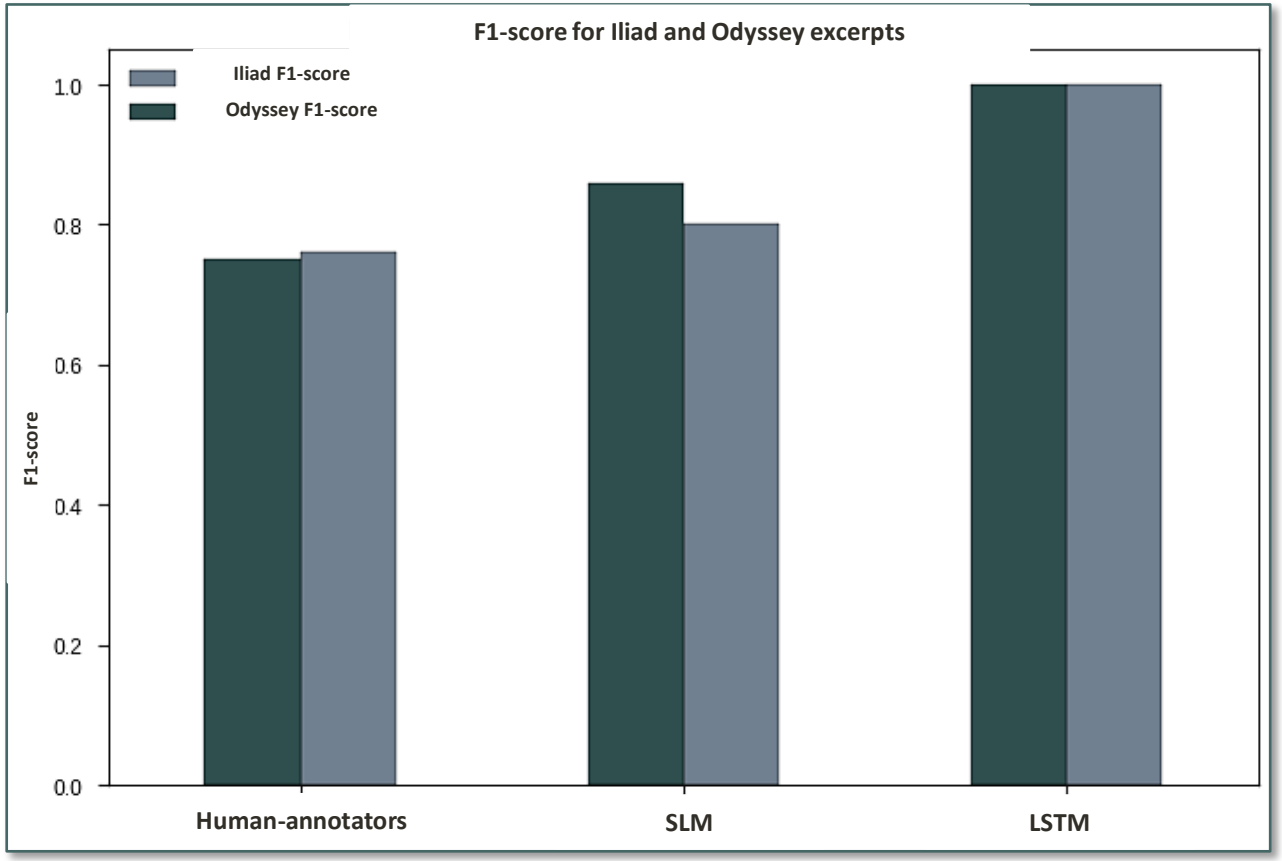
Algorithm 2: Binary classifier

This function returns a tag of 0 or 1, depending on the class predicted to belong to the given quote.

```
1. Function classify(Iliad_model, Odyssey_model, text):
2.    Set PPL_O equal to Perplexity(Odyssey_model, text)
3.    Set PPL_I equal to Perplexity(Iliad_model, text)
4.    if PPL_O is less than PPL_I:
5.        return “0” ⟵ Odyssey
6.     else
7.        return “1” ⟵ Iliad
```

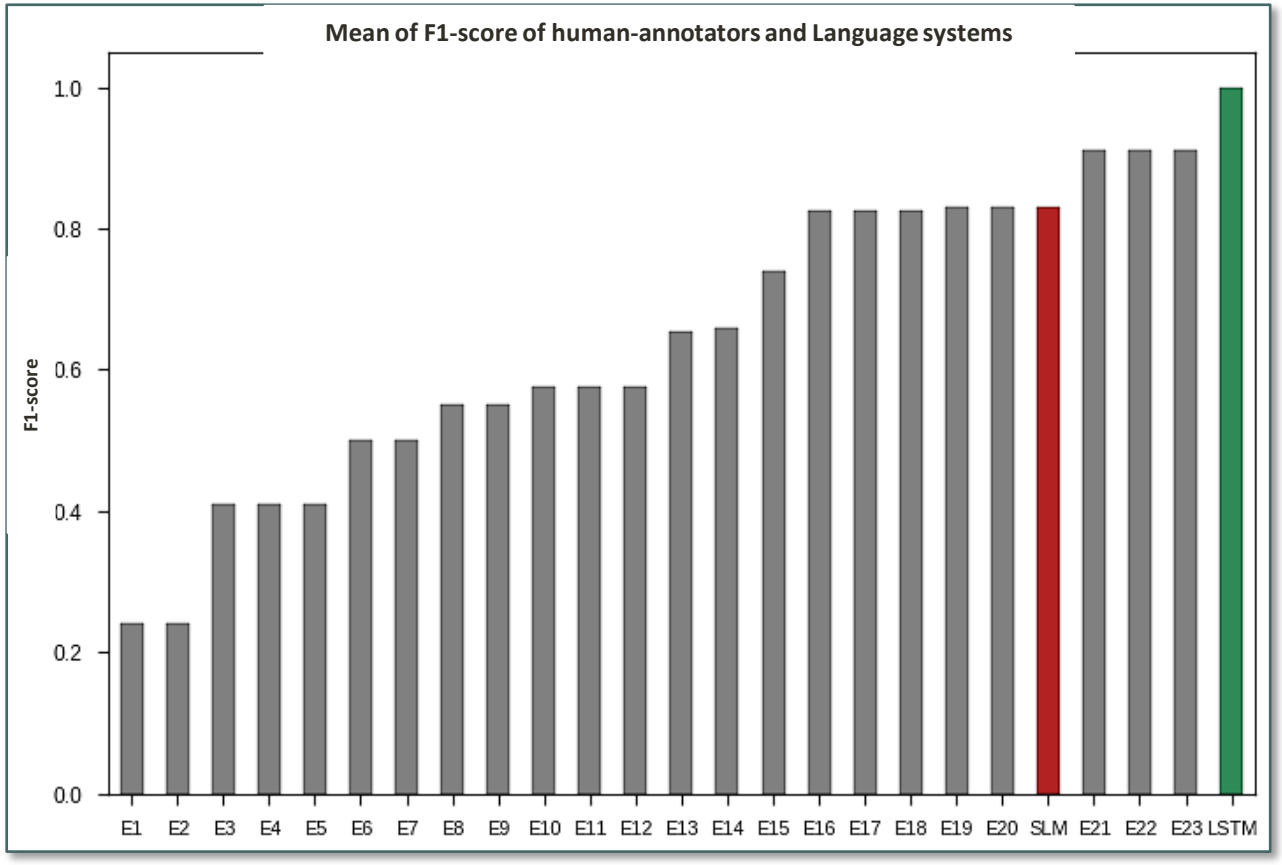
- **F1-SCORE** $\in [0,1]$
- 1 if all excerpts are classified correctly

CLASSIFICATION OF HOMERIC TEXTS WITH SLM AND LSTM COMPARISON WITH HUMAN-ANNOTATORS



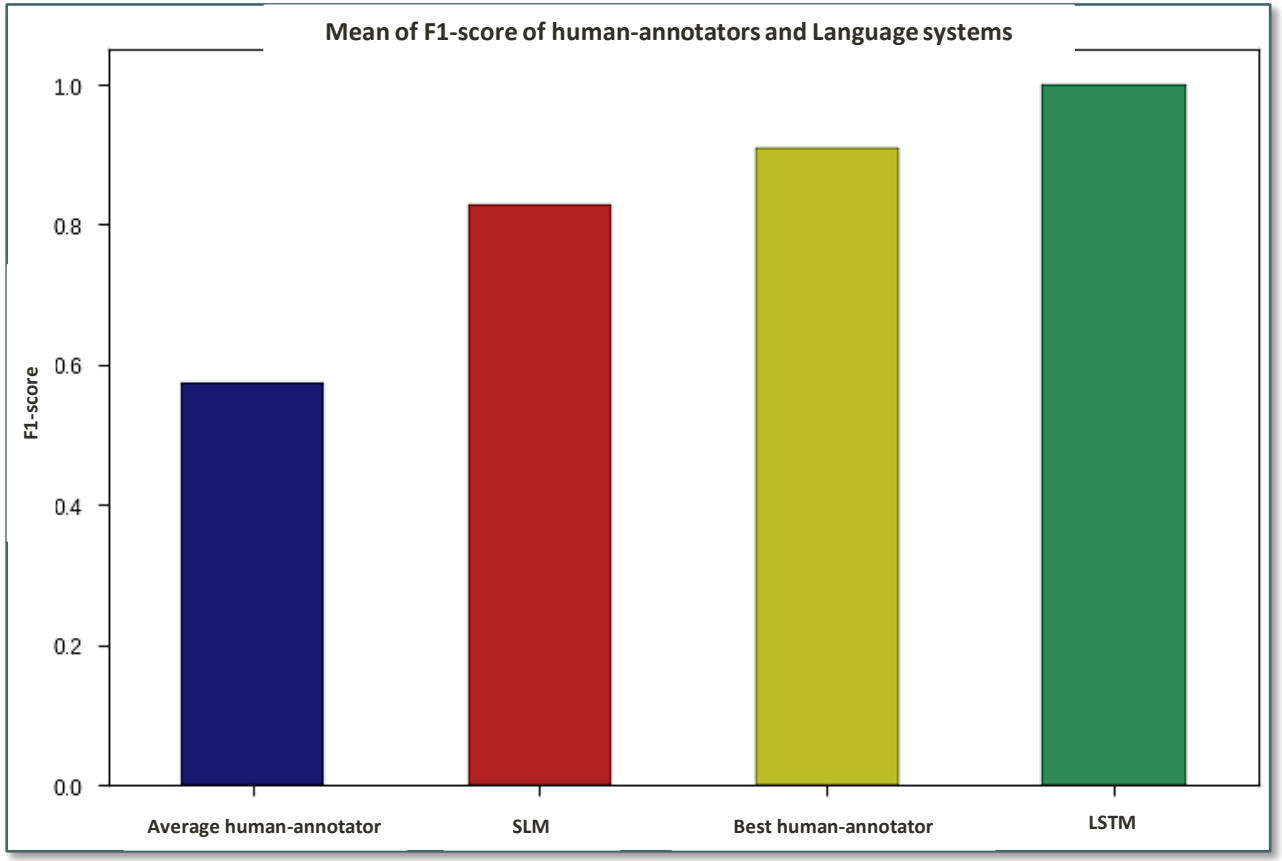
	Iliad F1-score	Odyssey F1-score
LSTM	1.00	1.00
SLM	0.80	0.86
Human-annotators	0.76	0.75

CLASSIFICATION OF HOMERIC TEXTS WITH SLM AND LSTM COMPARISON WITH HUMAN-ANNOTATORS



Overall performance of human-annotators F1-score, SLM and LSTM for the Iliad and Odyssey excerpts of the questionnaire

CLASSIFICATION OF HOMERIC TEXTS WITH SLM AND LSTM COMPARISON WITH HUMAN-ANNOTATORS



Overall performance F1-score of average human-annotator, best human-annotator as well as language systems, SLM and LSTM, for the Iliad and Odyssey excerpts of the questionnaire

DISCUSSION ON THE CLASSIFICATION OF HOMERIC TEXTS WITH SLM AND LSTM

Language systems		Human interpretation
LSMT	SLM	Human-annotators
1.00	0.83	0.755

Neural language models, such as the LSTM, perform remarkably well in the classification between the Iliad and the Odyssey, both from traditional statistical language models and from human-annotators who are somewhat familiar with Homeric texts.

CONCLUSION

1. Indeed, there are rhapsodies in both the Iliad and the Odyssey that show a greater linguistic affinity than others with the entire epic.
2. The language models seem to distinguish some rhapsodies that have a greater deviation from the linguistic style of the epics. This gives rise to further research to see if the discrepancies are significant enough to indicate different paternity.
3. The Homeric hymn “To Aphrodite" shows the greatest linguistic affinity with the whole of the Iliad and the Odyssey than the other hymns.
4. Artificial language models can more successfully categorize Iliad and Odyssey passages into their respective subordinate work than the human interpretation.

FUTURE WORK

- Enrichment of the questionnaire with more excerpts
- Exploring the Homeric question with other categories of Neural language models
- Classification of Homeric passages among other ancient writers (e.g., Hesiod)

BIBLIOGRAPHY

Chaski, C. E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International journal of digital evidence*, 4(1), 1-13.

Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., and Stein, B. (2013). Recent trends in digital text forensics and its evaluation plagiarism detection, author identification and author profiling. In *Proceedings of Conference and Labs of the Evaluation Forum*, CLEF, pages 282–302, Valencia, Spain.

Kimler, M. (2003). *Using style markers for detecting plagiarism in natural language documents*. Institutionen för datavetenskap.

Love, H. (2002). *Attributing authorship: An introduction*. Cambridge University Press.

Morris, I., & Powell, B. B. (1997). *A new companion to Homer*. Brill.

Thank you very much!

For any question, contact the following e-mails:

- ***Fasoi Maria:*** *maria.fassoi95@gmail.com*
- ***John Pavlopoulos:*** *annis.pavlo@gmail.com*
- ***Maria Konstantinidou:*** *konstantinidou.maria5@gmail.com*