

# HoLM: Analyzing Linguistic Unexpectedness in Homeric Poetry

John Pavlopoulos<sup>1,2</sup>, Ryan Sandell<sup>3</sup>, Maria Konstantinidou<sup>4</sup>,  
Chiara Bozzone<sup>3</sup>

<sup>1</sup> Department of Informatics, Athens University of Economics and Business

<sup>2</sup> Archimedes/Athena RC, Greece

<sup>3</sup> Ludwig-Maximilians-Universität München, Germany

<sup>4</sup> LaPPED, Department of Greek Philology,  
Democritus University of Thrace, Greece



May 2024

LREC-CoLING 2024

Slides available at: [www.ryan-sandell.com/research/](http://www.ryan-sandell.com/research/)

1. What is the **authorship** of the Homeric epics, *Iliad* and *Odyssey* (~ 8th c. BCE)?

# Research Questions and Goals

1. What is the **authorship** of the Homeric epics, *Iliad* and *Odyssey* (~ 8th c. BCE)?
2. Can the composition of these works be explored at the level of *verses* (sequences of ~5–8 words) by identifying **linguistically surprising verses**?

# Research Questions and Goals

1. What is the **authorship** of the Homeric epics, *Iliad* and *Odyssey* (~ 8th c. BCE)?
2. Can the composition of these works be explored at the level of *verses* (sequences of ~5–8 words) by identifying **linguistically surprising verses**?
3. Are  $n$ -gram language models a potentially profitable approach to 2.?

# Philological Background

1. Debate on Homeric authorship goes back (at least) to the 3rd c. BCE.
  - ▶ Ancient scholarship at Alexandria; some surviving papyri from the late 1st Millennium BCE with textual variants.
  - ▶ These epics were oral and traditional, and though they are composed in a natural language, they are remarkably *formulaic* (Parry 1971, Lord 1960, Bozzone 2014, 2024).

# Philological Background

1. Debate on Homeric authorship goes back (at least) to the 3rd c. BCE.
  - ▶ Ancient scholarship at Alexandria; some surviving papyri from the late 1st Millennium BCE with textual variants.
  - ▶ These epics were oral and traditional, and though they are composed in a natural language, they are remarkably *formulaic* (Parry 1971, Lord 1960, Bozzone 2014, 2024).
2. Exactly when, where, and how these works were first put into writing is unknown (Turner 2011).
  - ▶ A single, writing poet (unlikely)?
  - ▶ Multiple poets working with a scribe(s)?

# Philological Background

1. Debate on Homeric authorship goes back (at least) to the 3rd c. BCE.
  - ▶ Ancient scholarship at Alexandria; some surviving papyri from the late 1st Millennium BCE with textual variants.
  - ▶ These epics were oral and traditional, and though they are composed in a natural language, they are remarkably *formulaic* (Parry 1971, Lord 1960, Bozzone 2014, 2024).
2. Exactly when, where, and how these works were first put into writing is unknown (Turner 2011).
  - ▶ A single, writing poet (unlikely)?
  - ▶ Multiple poets working with a scribe(s)?
3. Certain: the poems also contain later **interpolations**.
  - ▶ Book 10 of the *Iliad* is likely a very large interpolation (Danek 1988, Danek 2012).
  - ▶ Most potential interpolations are perhaps the size of a single verse.

# Approaches to Authorship Analysis of the Homeric Epics

1. Known-author attribution is impossible in this case, while unknown-author verification is very challenging (Juola 2006, Stamatatos 2009, Kabala 2020).
  - ▶ Bozzone and Sandell (2022) use book-level frequencies (word bigrams, character trigrams) with clustering techniques to explore likely groupings of books.
  - ▶ Fasoï et al. (2021) and Pavlopoulos and Konstantinidou (2023) employ character-level language modeling at the book level and in excerpts of several hundred lines to try to identify linguistically unexpected sections.
2. Frequent identical or near-identical verses (*formulas*) create challenges.
  - ▶ Could some repeated formulas be interpolations taken out of their original context?

# Formularity in Homer

	<b>Verse</b>	<b>Line numbers</b>
I1	killan te zatheēn tenedoio te iphi anasseis	(38, 452)
I1	aideisthai th' hierēa kai aglaa dechthai apoina	(23, 377)
I1	hōs ephat' euchomenos, tou d' eklue Phoibos Apollōn	(43, 457)
I2	tō d' hama tessarakonta melainai nēes heponto	(534, 545, 556)
I2	<i>idem</i>	(630, 644, 710 )
I2	<i>idem</i>	(737, 759)
O1	tēn d' au Tēlemachos pepnumenos antion ēuda	(213, 230, 306, 345)
O1	ton d' au Tēlemachos pepnumenos antion ēuda	(388, 412)
O1	ton d' ēmeibet' epeita thea glaukōpis Athēnē	(44, 80, 314)
O2	ton d' au Tēlemachos pepnumenos antion ēuda	(129, 208, 309)
O2	keklyte dē nyn meu, Ithakēsioi, hotti ken eipō	(25, 161, 229)

Examples of identical verses in Books 1 and 2 of the *Iliad* and *Odyssey*.

# Present Work

1. Statistical testing from Pavlopoulos and Konstantinidou 2023, applied to single verses (not entire books or lengthy excerpts).
2. Variant of **Perplexity** estimates the linguistic unexpectedness of each verse.

1. Statistical testing from Pavlopoulos and Konstantinidou 2023, applied to single verses (not entire books or lengthy excerpts).
2. Variant of **Perplexity** estimates the linguistic unexpectedness of each verse.
3. Correlation analysis:
  - ▶ Perplexity and frequency of named entities.
  - ▶ Perplexity and frequency of character 5-grams.
  - ▶ Perplexity and (inverse) frequency of words.

1. Statistical testing from Pavlopoulos and Konstantinidou 2023, applied to single verses (not entire books or lengthy excerpts).
  2. Variant of **Perplexity** estimates the linguistic unexpectedness of each verse.
  3. Correlation analysis:
    - ▶ Perplexity and frequency of named entities.
    - ▶ Perplexity and frequency of character 5-grams.
    - ▶ Perplexity and (inverse) frequency of words.
- ⇒ High Perplexity robustly correlates with *hapax legomena*.
- ▶ Bias may be partly overcome by combining perplexity and word frequency into a single measure.

# Estimation of Unexpectedness: PPL Scores

1. Given the verses  $V^b$  from book  $b$  of a poem (*Iliad* or *Odyssey*), we train a statistical character-level language model  $m_b$  on all the verses of the remaining books.

# Estimation of Unexpectedness: PPL Scores

1. Given the verses  $V^b$  from book  $b$  of a poem (*Iliad* or *Odyssey*), we train a statistical character-level language model  $m_b$  on all the verses of the remaining books.
2. We compute the negative log-likelihood of  $m_b$  for each character of  $v \in V^b$ , and then we average this score across the verse's characters, which we call the average bits per character (BPC) score (Hwang and Sung 2017).

# Estimation of Unexpectedness: PPL Scores

1. Given the verses  $V^b$  from book  $b$  of a poem (*Iliad* or *Odyssey*), we train a statistical character-level language model  $m_b$  on all the verses of the remaining books.
2. We compute the negative log-likelihood of  $m_b$  for each character of  $v \in V^b$ , and then we average this score across the verse's characters, which we call the average bits per character (BPC) score (Hwang and Sung 2017).
3. Equivalent Perplexity variant is defined as follows (Graves 2013, Dror et al. 2020):

$$PPL(v, m^b) = 2^{|\bar{w}| * BPC(v, m^b)} \quad (1)$$

# Estimation of Unexpectedness: PPL Scores

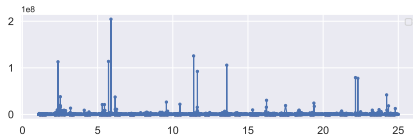
1. Given the verses  $V^b$  from book  $b$  of a poem (*Iliad* or *Odyssey*), we train a statistical character-level language model  $m_b$  on all the verses of the remaining books.
2. We compute the negative log-likelihood of  $m_b$  for each character of  $v \in V^b$ , and then we average this score across the verse's characters, which we call the average bits per character (BPC) score (Hwang and Sung 2017).
3. Equivalent Perplexity variant is defined as follows (Graves 2013, Dror et al. 2020):

$$PPL(v, m^b) = 2^{|\bar{w}| * BPC(v, m^b)} \quad (1)$$

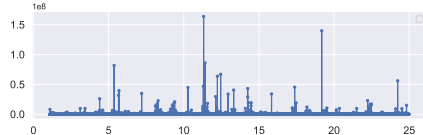
4. We obtain 48 LMs in total (24 for each poem).

# The HoLm Dataset

The HoLM dataset consists of one *PPL* score per Homeric verse.



(a) Iliad



(b) Odyssey

*PPL* per verse across the books (horizontally) of the two Homeric poems.

# Data Preparation and Preprocessing

- ▶ Underlying text: digital editions of Monro and Allen 1920 (*Iliad*) and Murray 1919 (*Odyssey*).
- ▶ Texts were divided into individual hexametric lines (= verses; 12107 for *Odyssey*, 15683 for *Iliad*).
- ▶ Preprocessing: all non-alphabetic characters (punctuation, numbers, etc.) removed, all characters converted to lowercase.

# Exploratory Analysis and Recurrent Verses

- ▶ A given verse consists of  $42 \pm 1$  characters,  $7 \pm 1$  words.
- ▶ Mean word length is of 5.04 (Il.) or 5.00 (*Od.*) characters.

# Exploratory Analysis and Recurrent Verses

- ▶ A given verse consists of  $42 \pm 1$  characters,  $7 \pm 1$  words.
- ▶ Mean word length is of 5.04 (Il.) or 5.00 (Od.) characters.
- ▶ Recurrent verses: the five most frequent verses in the *Iliad* and *Odyssey* comprise 0.33% (51/15683) and 1% (109/12107) of the respective poems.

	Verse	#
I	kai min phōnēsas epea pteroenta prosēuda	15
I	hoi d' hote dē schedon ēsan ep' allēloisin iontes	10
I	ton d' apameibomenos prosephē podas ōkys achilleus	9
I	hōs eipōn otryne menos kai thymon hekastou	9
I	atreidē kydiste anax andrōn agamemnon	8
O	ton d' au tēlemachos pepnymenos antion ēuda	30
O	ton d' apameibomenos prosephē polymētis odysseus	25
O	ēmos d' ērigeneia phanē rododaktylos ēōs	20
O	tin d' apameibomenos prosephē polymētis odysseus	19
O	hōs ephat' autar ego min ameibomenos proseeipon	15

Most frequently occurring verses in the *Iliad* (I) and *Odyssey* (O) verses, with total number of occurrences (#) in that work.

# Perplexity

- ▶ 5-gram LMs were trained on texts consisting of 23 books from one of the *Iliad* or *Odyssey*, respectively.
- ▶ These models are used to compute one *PPL* score per verse.
  - ▶ The higher the *PPL* score, the more unexpected the verse, given the model.

- ▶ 5-gram LMs were trained on texts consisting of 23 books from one of the *Iliad* or *Odyssey*, respectively.
- ▶ These models are used to compute one *PPL* score per verse.
  - ▶ The higher the *PPL* score, the more unexpected the verse, given the model.
- ▶ The single most linguistically unexpected verse is I.5.887, followed by I.11.385.

---

I.5.887	hē ke zōs amenēnos ea chalkoio typēsi
I.11.385	toxota lōbētēr kera aglae parthenopipa
I.5.723	chalkea oktanēma sidēreō axoni amphis
I.2.363	hōs phrētē phrētrēphin arēgē phyla de phylois
I.13.589	thrōskōsin kyamoi melanochroes ē erebinthoi

---

O.11.320	anthēsai pykasai te genys euanythei lachnē
O.19.177	dōriees te trichaikes dioi te pelasgoi
O.11.415	ē gamō ē eranō ē eilapinē tethalyiē
O.5.368	hōs d' anemos zaēs ēiōn thēmōna tinaxē
O.12.453	autis arizēlōs eirēmēna mythologēuein

---

The five most linguistically unexpected verses per poem (ranked by *PPL* descending).

# Whence Perplexity?

1. **Question 1:** are there more precisely definable features that impact the Perplexity of a verse?

# Whence Perplexity?

1. **Question 1:** are there more precisely definable features that impact the Perplexity of a verse?
2. Correlation Analysis:
  - ▶ Frequency of words (TFIDF)
  - ▶ Frequency of 5-gram sequences
  - ▶ Frequency of named entities

# Whence Perplexity?

1. **Question 1:** are there more precisely definable features that impact the Perplexity of a verse?
2. Correlation Analysis:
  - ▶ Frequency of words (TFIDF)
  - ▶ Frequency of 5-gram sequences
  - ▶ Frequency of named entities
3. **Question 2:** are particular verses of the *Iliad* especially surprising for an LM trained on the *Odyssey* (and vice-versa)?
  - ▶ We refer to this measure as **cross-score**.

# Term Frequency

1. Inverse verse frequency (*IVF*) of words calculated by defining verses as documents, using standard IDF, with scikit-learn's `TfidfVectorizer` Pedregosa et al. (2011), using default values, and the `idf_` attribute.
2. Higher number of verses in which a word occurs  $\Rightarrow$  lower *IVF*
3. Reasonably good correlation between maximum *IVF* for all words in a verse and Perplexity of a verse.

	Spearman's $\rho$	Pearson's $\rho$ ( $\log_e$ -transformed)
<i>Iliad</i>	0.669	0.581
<i>Odyssey</i>	0.702	0.618

- $\Rightarrow$  Low-frequency lexemes (= high *IVF*) act as a good proxy for uncommon character transitions, which in turn directly affect the LM and *PPL* scores.

# Character 5-Grams

1. Correlation between frequency of specific character 5-grams ( $C5F$ ) and  $PPL$ .
2. Number of instances (tokens) of each  $C5F$  in each poem was calculated using NLTK's `ngram` module (Bird et al. 2009).
3. Reasonably good negative correlation between minimum  $C5F$  and  $PPL$  of a verse (in  $[-1, -0.6]$ ), even stronger correlation between median  $C5F$  and  $PPL$  (shown here).

	Spearman's $\rho$	Pearson's $\rho$ ( $\log_e$ -transformed)
<i>Iliad</i>	-0.646	-0.688
<i>Odyssey</i>	-0.667	-0.715

- ⇒ The lower the median frequency of all 5-grams in a verse, the greater the LM's surprise for the verse as measured by  $PPL$ .
- Furthermore, reasonably good negative correlations hold between  $IVF$  and minimum  $C5F$  (in  $[-1, -0.6]$ ).

# Named Entities

1. Correlation between the frequency of named entities (personal names like *Achilleus*) and *PPL*.
  2. Automatically identified using a Transformer-based recognizer trained on Ancient Greek (Yousef et al. 2022).
  3. Most verses (*Iliad*: 65%; *Odyssey*: 73%) contain no personal names.
  4. Correlation (Pearson's  $\rho$ ) between the number of personal names and *PPL* of a verse is negligible: -0.007 (*Iliad*), -0.003 (*Odyssey*).
- ⇒ Personal names do not impact the linguistic unexpectedness of a verse.

# Cross-poem Modeling

1. One LM per poem trained on all 24 books was used to compute *PPL* per verse per poem.
2. Difference in *PPL* for a verse (cross-score) is calculated as:  
 $d^v(\text{Source}, \text{Other}) = PPL(v, \text{Source}) - PPL(v, \text{Other})$ .
3. A large positive cross-score indicates that the verse is more unexpected for the source model than for the other model (despite the bias in favor of the source model).

I.6.490	all' eis oikon iousa ta s' autēs erga komize	5059.86
I.1.485	nēa men hoi ge melainan ep' ēpeiroio eryssan	4717.82
I.13.821	hōs ara hoi eiponti epeptato dexios ornīs	3963.98
O.24.488	bē de kat' oulympoio karēnōn aixasa	4789.3
O.22.124	hippourin deinon de lophos kathyperthen eneuen	3075.1
O.11.270	tēn echen amphitryōnos uios menos aien ateirēs	2706.8

Verses with high cross-scores

4. Significantly higher proportion of verses with positive cross-score in the *Iliad* ( $\chi^2 = 25.17, p < 0.01$ ) may point to a lower level of linguistic homogeneity in the *Iliad*.

# Findings and Discussion

1. Both lexical frequency and phonotactics (as captured by 5-gram frequency) likely impact the behavior of the LMs.
2. Named entities have no impact on the Perplexity of a verse.
3. Verses with high  $PPL$ , but no especially infrequent terms, can be identified by examining the ratio  $\frac{PPL}{IVF}$  of a verse.

# Potential Applications

1. **Pedagogical:** are verses with high *PPL* scores more challenging for students of Ancient Greek to translate?

# Potential Applications

1. **Pedagogical:** are verses with high *PPL* scores more challenging for students of Ancient Greek to translate?
2. **Style:** do more “traditional”/formulaic verses exhibit lower *PPL* scores?

# Potential Applications

1. **Pedagogical:** are verses with high *PPL* scores more challenging for students of Ancient Greek to translate?
2. **Style:** do more “traditional”/formulaic verses exhibit lower *PPL* scores?
3. The general methodology can readily be extended to other texts in other languages for the purpose of assessing linguistic unexpectedness.
  - ▶ Further testing on a variety of texts and languages will help to establish to what extent *PPL* is influenced by term frequency and character *n*-gram frequency.

Thank you!

- Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural Language Processing with Python. O'Reilly Media, Sebastopol, CA.
- Chiara Bozzone. 2014. Constructions: A New Approach to Formularity, Discourse, and Syntax in Homer. Ph.D. thesis, University of California, Los Angeles.
- Chiara Bozzone. 2024. Homer's Living Language. Cambridge University Press, Cambridge.
- Chiara Bozzone and Ryan Sandell. 2022. One or many Homers? Using quantitative authorship analysis to study the Homeric question. In Proceedings of the 32nd Annual UCLA Indo-European Conference: November 5th, 6th, and 7th, 2021, pages 21–48. Helmut Buske Verlag.
- Georg Danek. 1988. Studien zur Dolonie. Österreichische Akademie der Wissenschaften, Vienna.
- Georg Danek. 2012. The Doloneia revisited. In Øvind Andersen and Dag T. T. Haug, editors, Relative Chronology in Early Greek Epic Poetry, pages 106–121. Cambridge University Press, Cambridge.

- Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. 2020. Statistical significance testing for natural language processing. Synthesis Lectures on Human Language Technologies, 13(2):1–116.
- Maria Fasoï, John Pavlopoulos, and Maria Konstantinidou. 2021. Computational authorship analysis of Homeric language. In The 1st Digital Humanities Workshop, Kiev, Ukraine.
- Alex Graves. 2013. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.
- Kyuyeon Hwang and Wonyong Sung. 2017. Character-level language modeling with hierarchical recurrent neural networks. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5720–5724. IEEE.
- Patrick Juola. 2006. Authorship attribution. Foundations and Trends in Information Retrieval, 1:233–334.
- Jakub Kabala. 2020. Computational authorship attribution in medieval Latin corpora: the case of the Monk of Lido (ca. 1101–08) and Gallus Anonymous (ca. 1113–17). Lang. Resources & Evaluation, 54(1):25–56.

Albert Bates Lord. 1960. The Singer of Tales. Harvard University Press, Cambridge, MA.

David B. Monro and Thomas W. Allen, editors. 1920. Homeri Opera, 3rd edition. Clarendon Press, Oxford.

A. T. Murray, editor. 1919. Homer, Odyssey. Harvard University Press, Cambridge, MA.

Milman Parry. 1971. The Making of Homeric Verse: The Collected Papers of Milman Parry. Oxford University Press, Oxford.

John Pavlopoulos and Maria Konstantinidou. 2023. Computational authorship analysis of the Homeric poems. International Journal of Digital Humanities, 5:45–64.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3):538–556.

Frank Turner. 2011. The Homeric question. In Ian Morris and Barry Powell, editors, A New Companion to Homer, pages 123–145. Brill, Leiden.

Tariq Yousef, Chiara Palladino, and Stefan Jänicke. 2022.  
Transformer-based named entity recognition for Ancient Greek.