Journal of Open Humanities Data

Word Lengths in Classical and Post-Classical Greek

COLLECTION: REPRESENTING THE ANCIENT WORLD THROUGH DATA

DATA PAPER

MATHILDE BRU

ubiquity press

ABSTRACT

The purpose of this dataset is to collate average word lengths in Greek in the Classical period (using a sample from the 5th century BCE) and the average word lengths in the Post-classical period (using a sample from the 1st century CE). This dataset, which is stored as a CSV file in the Harvard Dataverse, has been used to demonstrate that the average length of words in Greek increased significantly between the Classical and Post-classical periods. It comprises two word lists and corresponding word lengths for each sample; a table of the average word lengths (for nouns, adjectives, verbs, adverbs, and total) of each sample and the statistical t-test results showing the significance of the difference in word lengths between Classical and Post-classical Greek. This dataset has a reuse potential for linguists investigating diachronic change in word lengths as well as for historical linguists looking at the evolution of Classical and Post-classical Greek.

CORRESPONDING AUTHOR:
Mathilde Bru
Department of Greek and Latin, University College London, London, UK
mathilde.bru.20@ucl.ac.uk

KEYWORDS:
lexicon; word lengths; Classical Greek; Post-classical Greek; diachrony

TO CITE THIS ARTICLE:
Bru, M. (2023). Word Lengths in Classical and Post-Classical Greek. *Journal of Open Humanities Data*, 9: 19, pp. 1–6. DOI: https://doi.org/10.5334/johd.121

(1) OVERVIEW

REPOSITORY LOCATION

https://doi.org/10.7910/DVN/HKP1VU.

CONTEXT

This dataset was produced as part of a PhD project (ongoing, thesis due to be submitted early 2024), entitled 'A study of variation and change in the Greek lexicon of the Post-classical period.'

(2) METHODS

STEPS

To show that words increased in length on average throughout the history of the Greek language, a core vocabulary was collated for both Classical Greek and Post-classical Greek, and the mean-average counts for the number of syllables of words of both time periods was calculated.[1] Following Fenk-Oczlon & Pilz (2021) and Mikros & Milička (2014), syllable count was chosen as the measure of word length rather than number of contrasting segments, or phonemes, which is the metric used by Nettle (1995). The metric of syllable count was felt to be the best measure of word length, due to the diachronic changes in the pronunciation of graphemes.[2] The average syllable lengths were calculated manually, by going through the word lists and counting the number of syllables in each word. The following boxplots (Figures 1, 2, 3, 4 and 5) show the spread of the distribution of the data.

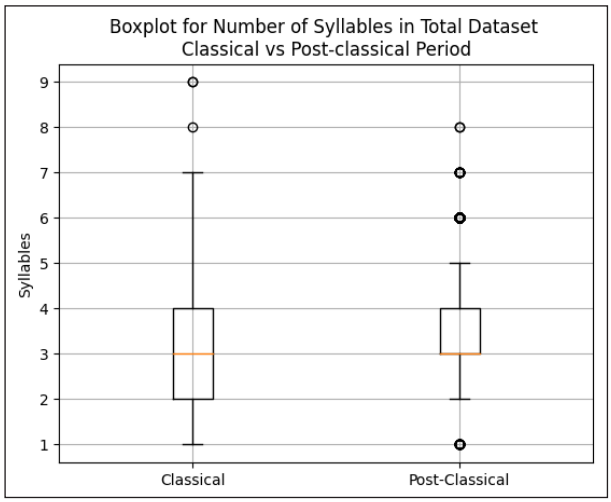

Figure 1 Number of syllables in total dataset (Classical VS Post-classical period).

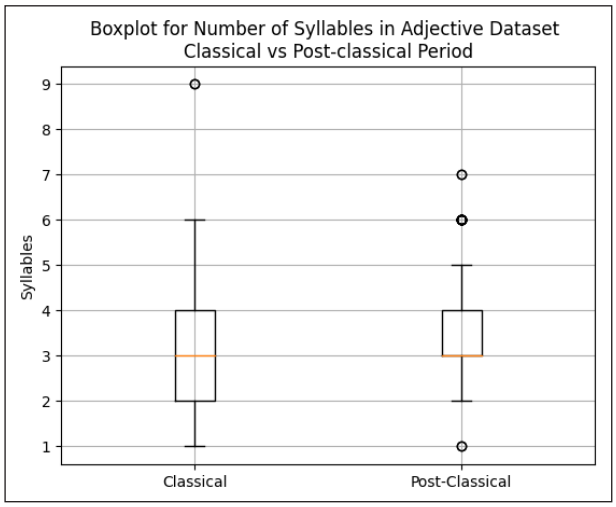

Figure 2 Number of syllables in adjective dataset (Classical VS Post-classical period).

1 See the following section for a description of how the core lexicon for each stage of the language was selected. This investigation follows Nettle (1995: 360–361) in studying words in their dictionary citation form 'as typological differences make cross-language comparisons of morphemes and words in discourse much more problematic.' Moreover, morphological developments between the Classical period and Post-classical period are another factor for lexical change, and this dataset was created to facilitate an investigation of phonological features. Thus, the number of syllables recorded for this study was for the first person singular present indicative; the nominative singular; and the masculine nominative singular form for verbs, nouns and adjectives respectively.

2 Nettle's (1995) study is synchronic; and so it is less affected by this consideration.

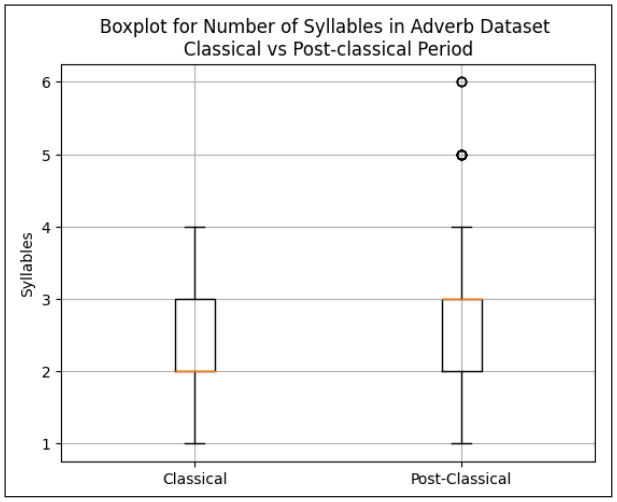

Figure 3 Number of syllables in adverb dataset (Classical VS Post-classical period).

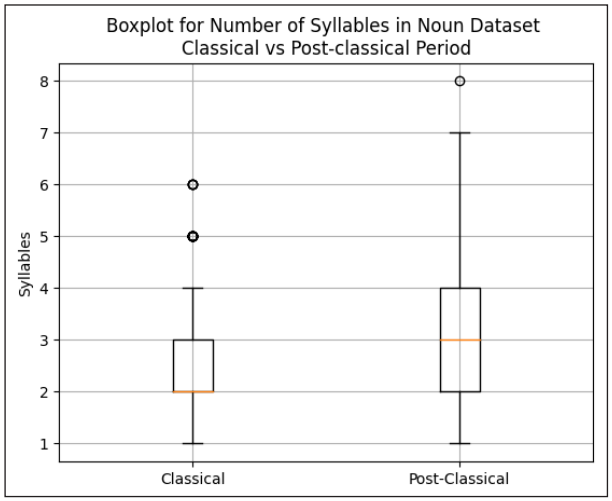

Figure 4 Number of syllables in noun dataset (Classical VS Post-classical period).

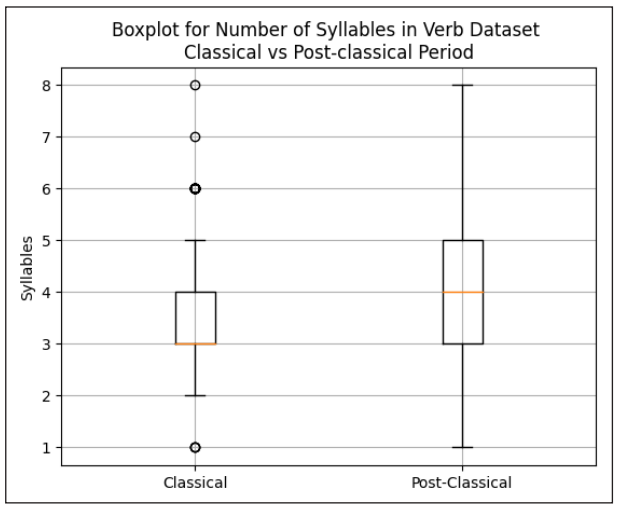

Figure 5 Number of syllables in verb dataset (Classical VS Post-classical period).

Using this dataset, a two-sample t-test was carried out on both the overall average and the average of each word class.

1. In aggregate, in the Classical Greek sample ([Figure 1], M[3] = 3.09, SD[4] = 1.07) words had fewer syllables than in the Post-classical Greek sample (M = 3.48, SD = 1.13), t(3951) = 13.1, p < .001. The Cohen's d[5] is 0.32, showing there is a highly significant small-moderate effect size.

3 Mean number of syllables.

4 Standard deviation.

5 Cohen's d is a standardised effect size that indicates the difference between two means. It is calculated by taking the difference between two means and dividing by the data's standard deviation.

2. Adjectives in the Classical Greek sample ([Figure 2], M = 3.18, SD = 1.07) had fewer syllables than in the Post-classical Greek sample (M = 3.43, SD = 1.02), t(734) = 3.69, p < .001. The Cohen's d is 0.24, showing there is a significant small effect size.
3. Adverbs in the Classical Greek sample ([Figure 3], M = 2.29, SD = 0.76) had fewer syllables than in the Post-classical Greek sample (M = 2.81, SD = 0.98), t(321) = 5.55, p < .001. The Cohen's d is 0.66, showing that there is a significant moderate-large effect size.
4. Nouns in the Classical Greek sample ([Figure 4], M = 2.70, SD = 0.91) had fewer syllables than in the Post-classical Greek sample (M = 3.16, SD = 1.09), t(1382) = 10.35, p < .001. The Cohen's d is 0.45, showing that there is a highly significant moderate effect size.
5. Verbs in the Classical Greek sample ([Figure 5], M = 3.49, SD = 1.04) had fewer syllables than in the Post-classical Greek sample (M = 3.91, SD = 1.07), t(1591) = 9.35, p < .001. The Cohen's d is 0.38, showing that there is a highly significant small-moderate effect size.

SAMPLING STRATEGY

The source used in this thesis to collect a core vocabulary of Classical Greek was the complete word list (2188 lemmas), generated by the *Perseus* software, of Aristophanes' *Clouds*.[6] The source used to collect the core vocabulary of Roman period Greek was the *Vocabulary of the Greek Testament illustrated from the papyri and other non-literary sources* Moulton & Milligan (1914–1929), which collects 4671 lexemes common to both the New Testament and the Roman period inscriptions and documentary papyri.[7] The language of Aristophanes is widely understood by historical linguists to represent something as close as we can get to everyday language in the Classical period,[8] and the language of the New Testament and papyri is used in the same way for scholars working on the Post-classical period.[9] The choice of these two sources remedies two key problems with Nettle's (1995) study: firstly his sample size for each language is small, only 50 head-words, and secondly, these were chosen at random from a dictionary, which means that one sample might include mostly rare or technical words while another might include mostly common, everyday words, and so these might not be truly comparable. Furthermore, the dictionaries in question were of different sizes; and Nettle (1995: 361) himself admits that 'a smaller dictionary would contain generally more common, hence shorter, words.'. While neither of my sources are of course comprehensive, the total number of lexemes collected are significant enough and cover enough core vocabulary to give a representative sample. Although the sample for Post-classical Greek is larger than the sample for Classical Greek, both samples are of a considerable size and contain a similar ratio of different word classes. The following word classes were excluded from the total count in both texts, as they are in all cases significantly rare in both lists, and in some cases irrelevant to a discussion of lexical change: personal and place names; conjunctions; interjections; particles; prepositions; prefixes; pronouns; numerals; articles. Therefore, from both word lists, only nouns, adjectives, verbs and adverbs were taken into account for this investigation. In total, there are 653 nouns, 365 adjectives, 794 verbs, and 129 adverbs, for a total of 1941 surveyed words in Aristophanes' word list. There are 1760 nouns, 612 adjectives, 1686 verbs and 224 adverbs in Moulton and Milligan's Lexicon, for a total of 4282 surveyed words.

(3) DATASET DESCRIPTION

OBJECT NAME

Word lengths in Classical and Post-classical Greek.

FORMAT NAMES AND VERSIONS

Comma Separated Values (CSV).

6 https://vocab.perseus.org/word-list/urn:cts:greekLit:tlg0019.tlg003.perseus-grc2/?page=all.

7 Although only the documentary papyri that had been discovered by the early 20th century, at the time of the book's publication.

8 Cf. e.g. Willi (2003).

9 Cf. e.g. Bentein & Janse (2021).

CREATION DATE

2023-07-20.

DATASET CREATORS

Mathilde Bru, PhD student, UCL.

LANGUAGE

English, Ancient Greek.

LICENSE

CC0 1.10.

REPOSITORY NAME

Dataverse.

PUBLICATION DATE

2023-10-09

(4) REUSE POTENTIAL

This dataset was created to study lexical change in Ancient Greek as part of a Historical Linguistics thesis. However, it is also highly re-usable by modern linguists interested in studying diachronic change in word lengths in a corpus language. Studies which have investigated variation in word lengths include Nettle (1995; 1998), Wichmann et al. (2011) and Fenk-Oczlon & Pilz (2021). These papers have demonstrated that there is a negative correlation between phoneme inventory and word length, something which can now be shown to be true for Classical and Post-classical Greek: the Greek of the Post-classical period had fewer phonemes than in the Classical period,[10] and, as the data show, the lexemes of the Post-classical period were longer than those of the Classical period.[11] Previous studies have all have so-far focussed on *synchronic* comparison between multiple languages. For example, Nettle (1995) compares ten modern languages and repeats his findings in a 1998 paper comparing twelve West African languages; Wichmann et al. (2011) show using data from over 3000 languages collected in the Automated Similarity Judgment Program (ASJP) that average word length and phoneme inventory sizes are negatively correlated;[12] and Fenk-Oczlon & Pilz (2021) analyse parallel text material from 61 languages and also find a negative correlation between phoneme inventory size and mean length of words, measured as number of syllables. This dataset is the first to collect relevant data on a single language *diachronically* (i.e. as opposed to its synchronic application on multiple languages which are being compared), and as such would be useful for linguists looking for evidence to show that the negative correlation between phoneme inventory and word length is found diachronically.

This dataset is also the first to show that negative correlation between phoneme inventory and word length holds true for ancient, as well as modern languages. In addition to the re-use potential for linguists, it would be of use for classicists and historical linguists looking at the diachronic evolution of Greek and needing data showing the average word lengths in the four main inflectional word classes of Greek in two different time periods. This would be useful for specialists in Classical and Post-classical Greek language and literature, as it would facilitate studies on the evolution of the ancient language, from the Classical to the Post-classical period.

10 Classical Greek had an inventory 36 phonemes in total; Post-classical Greek had around 21.

11 This is because the fewer the number of phonemes in a language, the longer the words need to be in order for them to be distinguishable.

12 They also observe that the lower limit to the number of phonemes that a language can tolerate is 10–12 (e.g. Pirahã, a language spoken in Amazonas, Brazil, and Rotakas, spoken in New Guinea).

ACKNOWLEDGEMENTS

My PhD supervisor, Professor Stephen Colvin, helped me develop the idea for this dataset and read a draft of this paper, and three anonymous reviewers provided me with invaluable feedback, which is reflected in the final version of this paper.

FUNDING INFORMATION

This dataset was produced as part of my PhD, which is funded by the AHRC London Arts & Humanities Partnership Research Studentship (2020–2023).

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATIONS

Mathilde Bru orcid.org/0000-0003-4470-0167
Department of Greek and Latin, University College London, London, UK

REFERENCES

Bentein, K., & **Janse, M.** (Eds.). (2021). *Varieties of Post-classical and Byzantine Greek*, Vol. 331. Berlin/Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110614404

Fenk-Oczlon, G., & **Pilz, J.** (2021). Linguistic Complexity: Relationships Between Phoneme Inventory Size, Syllable Complexity, Word and Clause Length, and Population Size. *Frontiers in Communication*, *6*, 1–7. DOI: https://doi.org/10.3389/fcomm.2021.626032

Mikros, G. K., & **Milička, J.** (2014). Distribution of the Menzerath's law on the syllable level in Greek texts. In G. Altmann, R. Čech, J. Mačutek & L. Uhlířová (Eds.), *Empirical Approaches to Text and Language Analysis* (pp. 181–189).

Moulton, J. H., & **Milligan, G.** (1914–1929). *The Vocabulary of the Greek Testament illustrated from the papyri and other non-literary sources*. London: Hodder and Stoughton Limited.

Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics*, *33*(2), 359–367.

Nettle, D. (1998). Coevolution of phonology and the lexicon in twelve languages of West Africa. *Journal of Quantitative Linguistics*, *5*(3), 240–245. DOI: https://doi.org/10.1080/09296179808590132

Wichmann, S., Rama, T., & **Holman, E. W.** (2011). Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology*, *15*, 177–197. DOI: https://doi.org/10.1515/lity.2011.013

Willi, A. (2003). *The Languages of Aristophanes: Aspects of Linguistic Variation in Classical Attic Greek*. Oxford: Oxford University Press.

TO CITE THIS ARTICLE:
Bru, M. (2023). Word Lengths in Classical and Post-Classical Greek. *Journal of Open Humanities Data*, 9: 19, pp. 1–6. DOI: https://doi.org/10.5334/johd.121

Submitted: 27 July 2023
Accepted: 17 October 2023
Published: 06 November 2023

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.