Academia letters: A simple, yet Statistically Robust Methodology for the Analysis of Textual Data.

Emmanuel Pantos[1], Fabricio T.P. Ono[2], Manosh Chowdhury[3],

[1]DLS Visiting Scientist, Didcot, South Oxfordshire, UK

[2]Department of Letters, Universidade Federal de Mato Grosso do Sul, Três Lagoas, Brazil

[3]Department of Anthropology, Jahangirnagar University, Bangladesh

Keywords: Stylometry, Authorial Style, Forensic Linguistics, Homeric Epics, Classical Greek Literature, Os Lusíadas, Tagore, Bengali

Introduction

This brief report is intended as a primer of a methodology for the analysis of digitally available texts in any language and of any period or genre using fairly elementary tools. Any aspiring analyst, young in age or younger in mind though hepta+/-generian, comfortable with spreadsheet data-handling and basic statistical concepts, can use these methods effectively, without dependence on corpus databases or black-box software packages. We demonstrate these methods using "training sets" of well-known-and-loved texts in English, Greek, Portuguese and Bengali (Bangla) and four recently published AL articles. Brief mention is made of the application of the same approach to a variety of ancient and modern texts, including the works of Homer and Aeschylus and the Portuguese epic poem *Os Lusíadas*.

Basic procedures

Any text in digital form is first processed to yield the document's 'Lexicon'[1] - the list of uniquely-spelled word-forms. A word-counting app calculates the frequency of occurrence to produce a .csv file copied into a spreadsheet. This is the very first step in the characterization of any text in any language or genre (Fig. 1). It is just one facet of a multi-dimensional analysis procedure, each aspect of which gives complementary stylistic information.

[1] Early 17th century modern Latin, from Greek lexikon (*βιβλίον*) "(book) of words", from *λέξις* "word", from *λέγειν* "speak". https://www.lexico.com/definition/lexicon [Last accessed 14 April 2021].

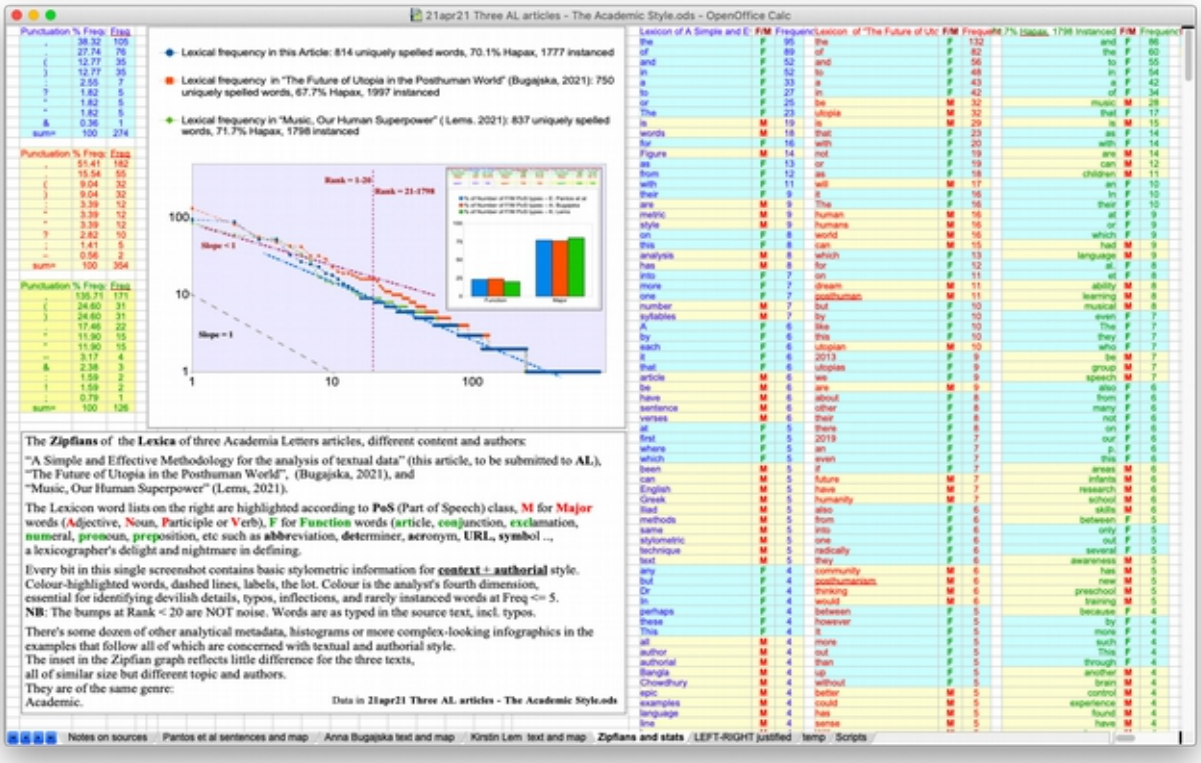

Figure 1. The Zipfians[2] (Piantadosi, 2014) *of the Lexicon of this text (v.0 21apr21) compared to two recent publications (Bugajska, 2021), (Lems, 2021). Every detail in the colour-coded infographics is a stylometric measure of text-content and authorial-style combined: the "Academic Style".*

The text is divided into "sentence segments" delimited by punctuation marks marking points where a poet, speaker or orator would naturally stop for breath, emphasis or dramatic effect. The histogram and the statistics computed provide stylometric information (Fig. 2a-c). In poetry, caesurae[3] occurring mid-verse are often followed by an 'enjambment' where a sentence spans[4] two or more verses. In the *Iliad* some 30% of the verses are enjambed as are 22% of the verses in the *Odyssey*. This often results in an undulating pattern in the "Location Histogram" of the number of words in each sentence.

[2] Named after Prof. George Kingsley Zipf (https://en.wikipedia.org/wiki/George_Kingsley_Zipf) and https://en.wikipedia.org/wiki/Zipf%27s_law [Last accessed 14 April 2021].

[3] In Greek and Latin verse a break between words within a metrical foot. https://www.lexico.com/definition/caesura and https://en.wikipedia.org/wiki/Caesura [Last accessed 14 April 2021].

[4] "In poetry, enjambment (from the French *enjambement*) is incomplete syntax verse-end where a phrase continues from one poetic line to the next, without terminal punctuation. Lines without enjambment are end-stopped.
In reading, the delay of meaning creates a tension that is released when the word or phrase that completes the syntax is encountered (called the rejet); the tension arises from the "mixed message" produced both by the pause of the line-end, and the suggestion to continue provided by the incomplete meaning. In spite of the apparent contradiction between rhyme, which heightens closure, and enjambment, which delays it, the technique is compatible with rhymed verse. Enjambment has a long history in poetry. Homer used the technique, and it is the norm for alliterative verse where rhyme is unknown". https://en.wikipedia.org/wiki/Enjambment, [Last accessed 14 April 2021].

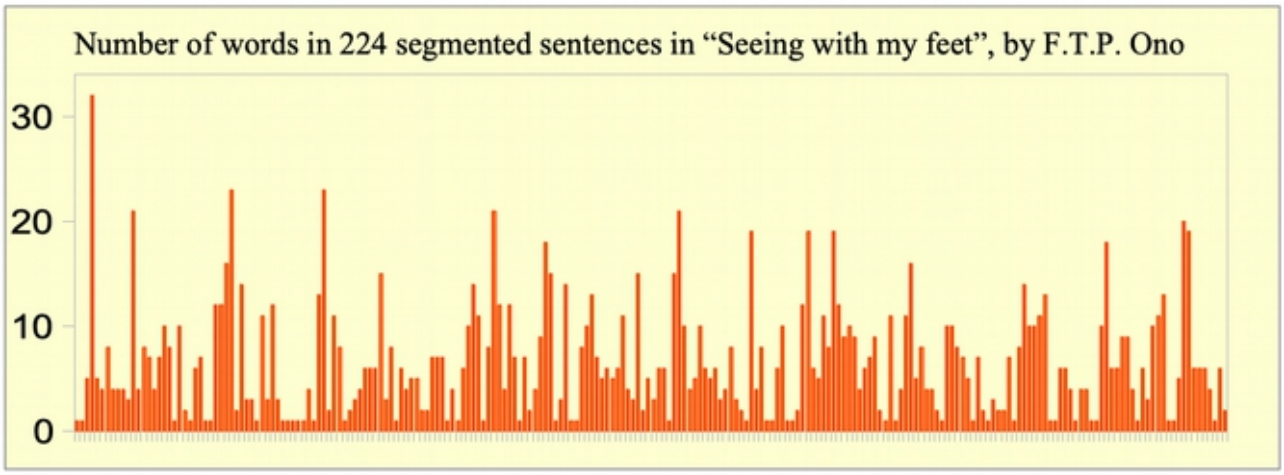

Figure 2a. Location Histogram of the number of words in each segmented sentence from an AL publication (Ono, 2021). The histogram reflects the document's textual AND authorial style. What is it that makes one write long or short sentences, in many cases flanked by a rising/falling intensity, as if the author seeks to stress a point? Is it grammar rules or is it natural tendency or perhaps personal choice in the use and the type of punctuation marks? People don't have a grammar book next to them when they write or speak, do they?

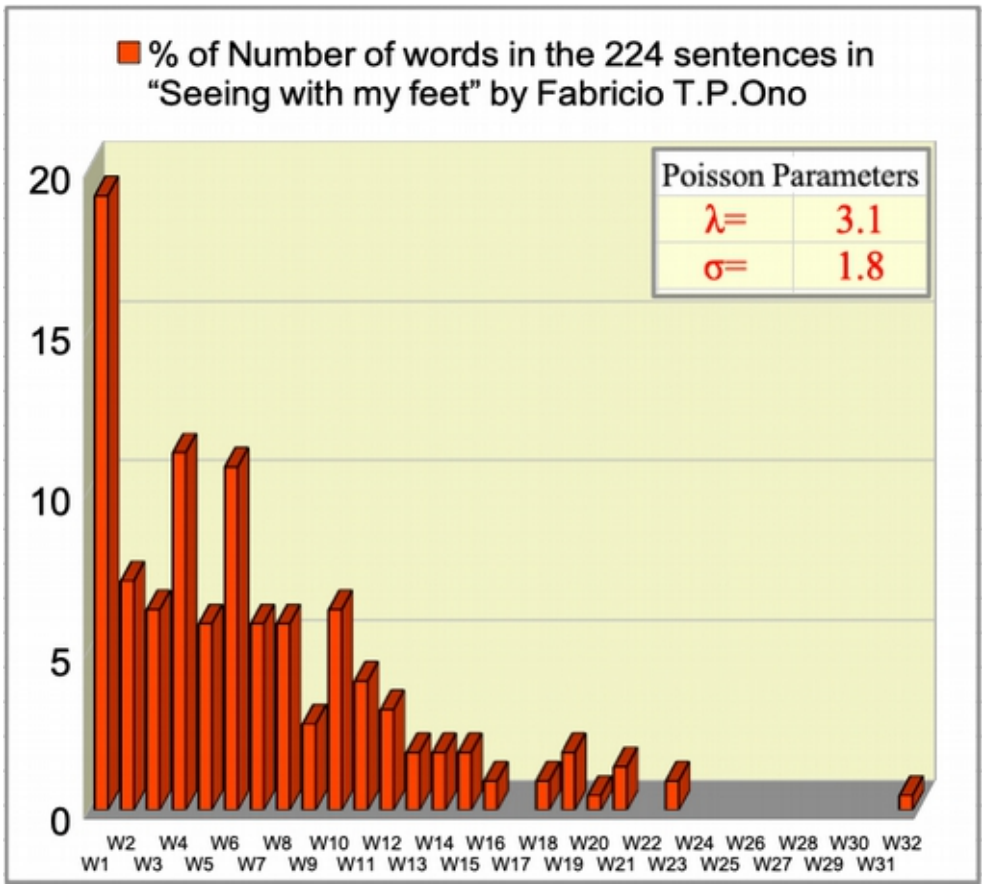

Figure 2b. The Poisson parameters λ=(mean value) and σ=(SQRT(mean)) complement the “textual texture” of the document histogrammed above.

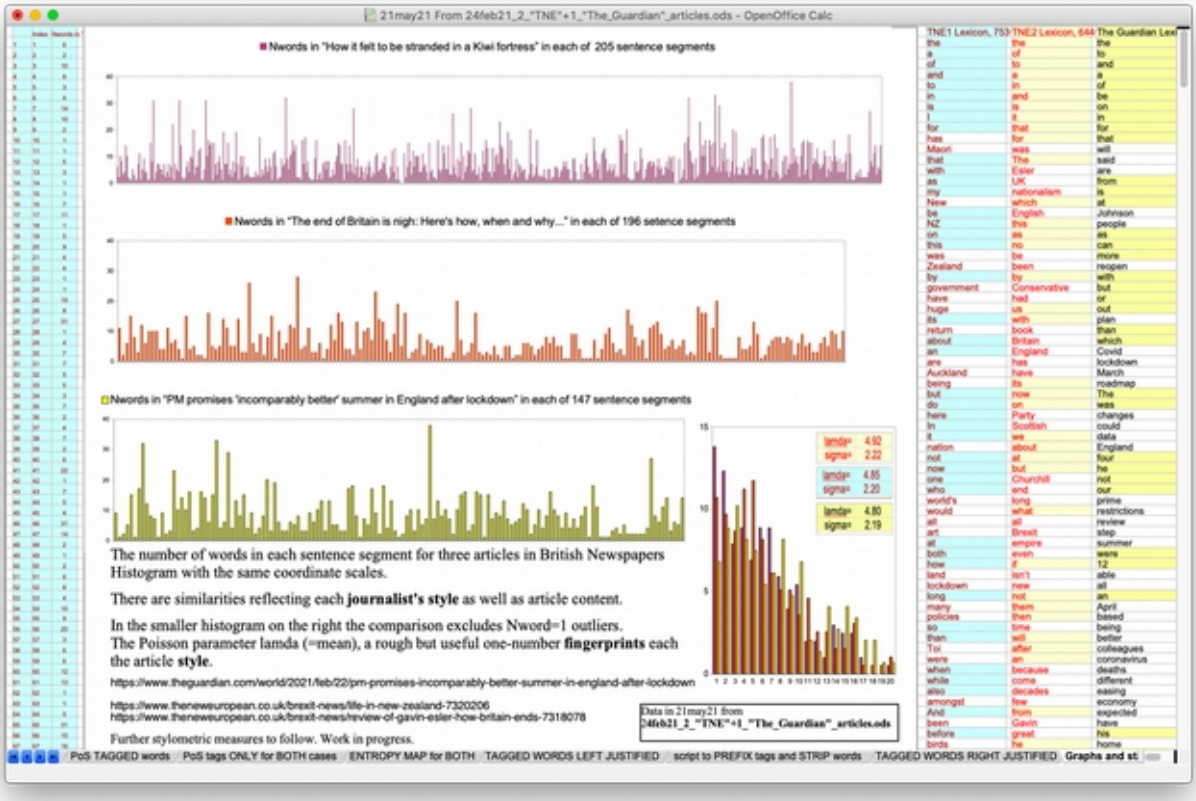

Figure 2c. Journalistic style: three articles, one in The New European[5] *and two in The Guardian*[6]*.*

It is worth noting that more than 50% of the words at the start of sentences are "function"[7] words while the sentence-terminating words are mostly of "major"[8] type[9]. The text you are reading now conforms to this global rule of storytelling, from the *Iliad* and *Genesis* to the speeches of popular(ist) politicians.

In the beginning was the Word,

F F M M F M

and the Word was with God,

F F M M F M

and the Word was God.

F F M M M

---

[5] *How it felt to be stranded in a Kiwi fortress,* TNE, #232, Feb. 18-24, 2021. https://www.theneweuropean.co.uk/brexit-news/life-in-new-zealand-7320206 Garth Cartwright.

[6] *PM promises incomparably better summer in England after lockdown* https://www.theguardian.com/world/2021/feb/22/pm-promises-incomparably-better-summer-in-england-after-lockdown Heather Stewart, Aubrey Allegretti and Jessica Elgot and *The end of Britain is nigh: here's how, when and why...* https://www.theneweuropean.co.uk/brexit-news/review-of-gavin-esler-how-britain-ends-7318078, James Hawes.

[7] **adv**erbs, **art**icles, **conj**unctions, **excl**amations, **num**erals, **pron**ouns and **prep**ositions + a few more (**det**erminers, **abbr**eviations, **acr**onyms or special non-word symbols called collectively tokens. A lexicographer's delight.

[8] Adjectives, Nouns, Participles and Verbs.

[9] Words in colour or **bold type** or in UPPPER CASE are also part of style and technique, annoyingly so, perhaps.

Consider the first verse of the *Iliad*[10] interpreted as a sequence of PoS (**P**art_**o**f_**S**peech) attributes (tags), used for metric analysis.

IL.1.1 **μῆνιν ἄειδε θεὰ [1]Πηληϊάδεω Ἀχιλῆος[1]**
IL.1.1 **SING, goddess, the anger of Peleus' son Achilleus**

+N_ +V_ +N_ +A_ +N_

(NOUN_VERB_NOUN_ADJECTIVE_NOUN)

A connected phrase of five major words tells us what the Iliad is all about: work of a genius.

There are only a few handfulls of major-words-only verses in the *Iliad* and *Odyssey*[11].

Histograms of particular word PoS-tags may differ from author to author (and from text to text by the same author!) and are key stylometric measures of context, technique and authorial-style combined. An important observation to make at this point is that content and technique are the major contributors to the statistics. The role of the symbols + - _ in the PoS-tags (e.g. +M_ for "major" and **-F_** for "function") will be addressed elsewhere.[12] However, basically, **-F_** or **+M_** PoS-tagged words can be joined together, either as distinct PoS-classes or as selected top-ranking words of either class, forming phrases characteristic of the authorial-style, e.g. Homer's distinct formulaic style of repeating verses or phrases at the start, middle or end of verse[13].

The same type of analysis can be carried out on the "metric pattern" of verses where words have been split into syllables each of which is allocated a "metric value": **2** for long, **1** for short[14].

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	NSYL	
ΜΗ	ΝΙΝ	Α	ΕΙ	ΔΕ	ΘΕ	Α	ΠΗ	ΛΗΙ	Α	ΔΕ	ΩΙ	Α	ΧΙ	ΛΗ	ΟΣ			
2	1	1	2	1	1	2	2	2	1	1	2	1	1	2	2		24	(=6*4)
Dactyl			**| Dactyl**			**|Spondee|**		**Dactyl**			**| Dactyl**			**|Spondee**			

The average metric value of all the verses is yet another stylometric measure (see Fig. 4). This information can also be represented in tabular form with a column for each syllable position. The order/disorder of each syllable column is calculated by a "Metric Entropy" app the output of which is displayed as a histogram (Fig. 3) which provides a more accurate comparison of the authorial style of different poetic texts or parts of, e.g. different rhapsodies in the *Iliad* or the *Odyssey*.

According to Shannon's information theory (Shannon, 1948)[15], the second law of thermodynamics implies that high-entropy==>disorder and vice-versa, low-entropy==>increased-order. The "Metric Entropy" histogram gives us a much more accurate stylometric measure than the "Average Metric" one.

---

[10] https://homer.library.northwestern.edu [Last accessed 14 April 2021].

[11] Here's another two: Δαναοῖσιν ἀγῶνα δίδούς ἐτέλεσσεν Ἀχιλλεύς. Πριάμωι νέκυν υἷος λαβὼν γέρα δῶκεν Ἀχιλλεύς. See also https://www.academia.edu/1817596/Recapturing_a_Homeric_Legacy_Images_and_Insights_from_the_Venetus_A_Manuscript_of_the_Iliad?email_work_card=view-paper [Last accessed 02 May 2021], written in the Byzantine font.

[12] E. Pantos et al, in preparation.

[13] E. Pantos, in preparation.

[14] The duration of long syllables is not necessarily twice that of a short syllable. Syllable duration is a matter of performance style, just as musical performances by different orchestras and conductors may differ.

[15] https://en.wikipedia.org/wiki/Information_theory, [and https://evo2.org/information-theory-made-simple/ and Odysseus and Calypso, (https://geometax12.blogspot.com/2018/ and https://geometax12.blogspot.com/2018/07/blog-post_9.html (in Greek, article 53. Η Φύση αγαπάει την αταξία!) [Last accessed 18 April 2021].

This is a particularly distinct aspect of poetic technique, be it in the dactylic epic hexameters of Homer, Hesiod and Virgil, or the works of lyric[16] Greek poets[17] and tragedians[18] such as Pindar, Aeschylus, Sophocles and Euripides.[19]

Every poem has rhythmic patterns just as every musical composition does.[20] Aeschylus' tragedies in particular have been referred to as lyric operas[21], with "song and dance" choral interludes separating character dialogues or monologues. A few more examples follow in Fig. 3-9.

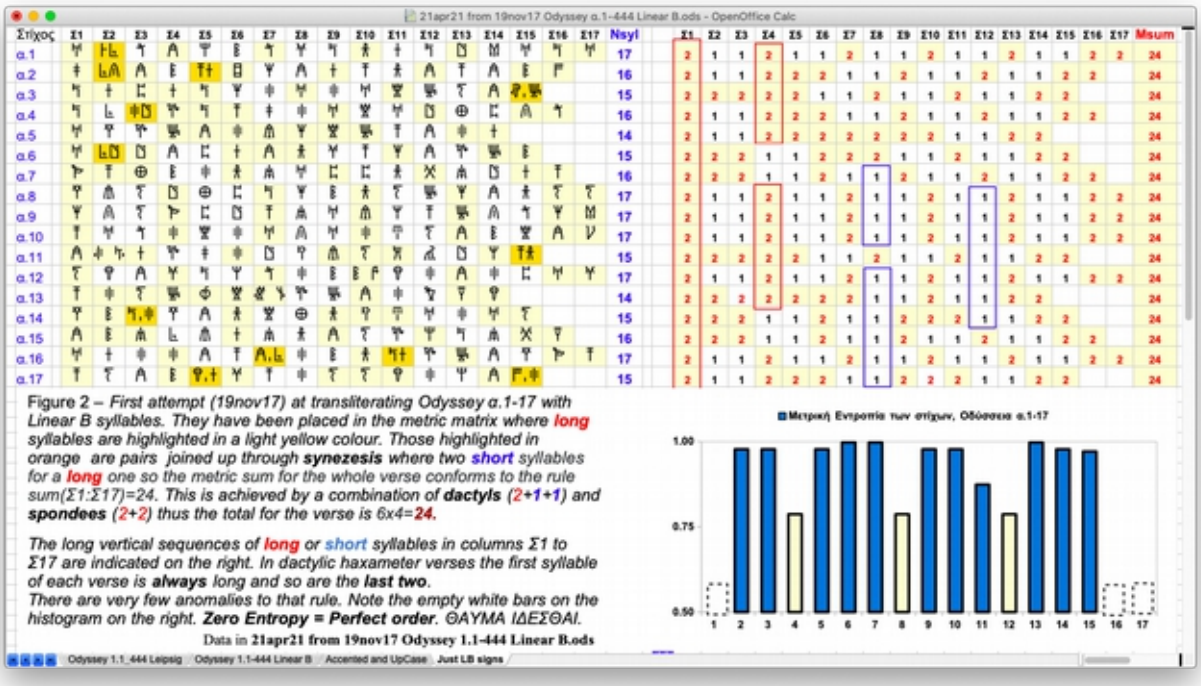

Figure 3. Odyssey α.1-17 where Homer's syllables have been replaced by Linear B syllabograms[22].

More examples of application of the methodology

The following presents some examples of analysis of a collection of short texts used in the development of the methodology[23]:

The Girl from Ipanema, in Brazilian Portuguese[24] and English, music by Antônio Carlos Jobim.

---

[16] https://en.wikipedia.org/wiki/Lyric_poetry [Last accessed 16 April 2021].

[17] https://en.wikipedia.org/wiki/Nine_Lyric_Poets [Last accessed 16 April 2021].

[18] https://en.wikipedia.org/wiki/List_of_ancient_Greek_playwrights [Last accessed 16 April 2021].

[19] Works analysed include Homer's *Iliad*, *Odyssey*, *Hymns*; Hesiod's *Theogony*, *Works and Days*, *Shield of Herakles*; Apollonius Rhodius' *Argonautica;* Virgil's *Aeneid;* Nonnus of Panopolis' *Dionysiaca.*

[20] An excellent book by Michael Spitzer, see https://literaryreview.co.uk/symphony-of-a-thousand-millennia [Last accessed 15 April 2021]. Also, see and listen to https://www.youtube.com/watch?v=UAmuQBnNty8, *Rediscovering Ancient Greek Music - A performance reconstructs the past*, and https://www.youtube.com/watch?v=-vKkK-x89Y4, *Ο Επιτάφιος του Σεικίλου - Το αρχαιότερο ολοκληρωμένο τραγούδι του κόσμου (Seikilos' Epitaph*, the oldest complete song in the world), [Last accessed 28 May 2021].

[21] https://oll.libertyfund.org/title/blackie-the-lyrical-dramas-of-aeschylus and https://andrewesimpson.com/oresteia/index.html [Last accessed 16 April 2021].

[22] The Linear B tablets discovered so far have not provided us with much poetry or prose; they are mostly accountants' stock-taking lists. But, who/when/where/how did the Mycenaean scribes get the syllabogram idea from, adopting it to their language? Ideas travel faster than camels in the desert or horses in the Eurasian steppe, which for quite a while now has been thought of as the region from which the Proto-Indo-European (PIE) languages spread out, eastward and westward.

[23] Further details may be obtained from the corresponding author (manolis.pantos@gmail.com).

[24] https://en.wikipedia.org/wiki/The_Girl_from_Ipanema . Music by Antônio Carlos Jobim and Portuguese lyrics by Vinícius de Moraes. English lyrics were written later by Norman Gimbel [Last accessed 15 April 2021].

I Keep Six Honest Serving-Men (They taught me all I knew), Rudyard Kipling[25]

Playing with Paran (=Heart/Desire), in Bengali (Bangla)[26] and English, Rabindranath Tagore[27]

Ithaca, in English and Greek, Constantine Cavafy[28],

and a wickedly funny and striking AL article in the original Bangla (Chowdhury, 2021a) and in English (Chowdhury, 2021b).

The objective was to study the influence of language as well as poetic technique on the authorial signature. We hope these works will appeal to our readers' literary tastes.

First, let us Bossa-Nova to the tune of the delightful Brazilian masterpiece, before diving into the deep waters of Forensic Linguistics,[29] Stylometry[30] and related topics (Coulthard and Johnson, 2007; Manousakis and Stamatatos, 2018). Just a few screenshots (infographics) from the analyst's development diary with their internal explanatory captions in information-rich colour will suffice.

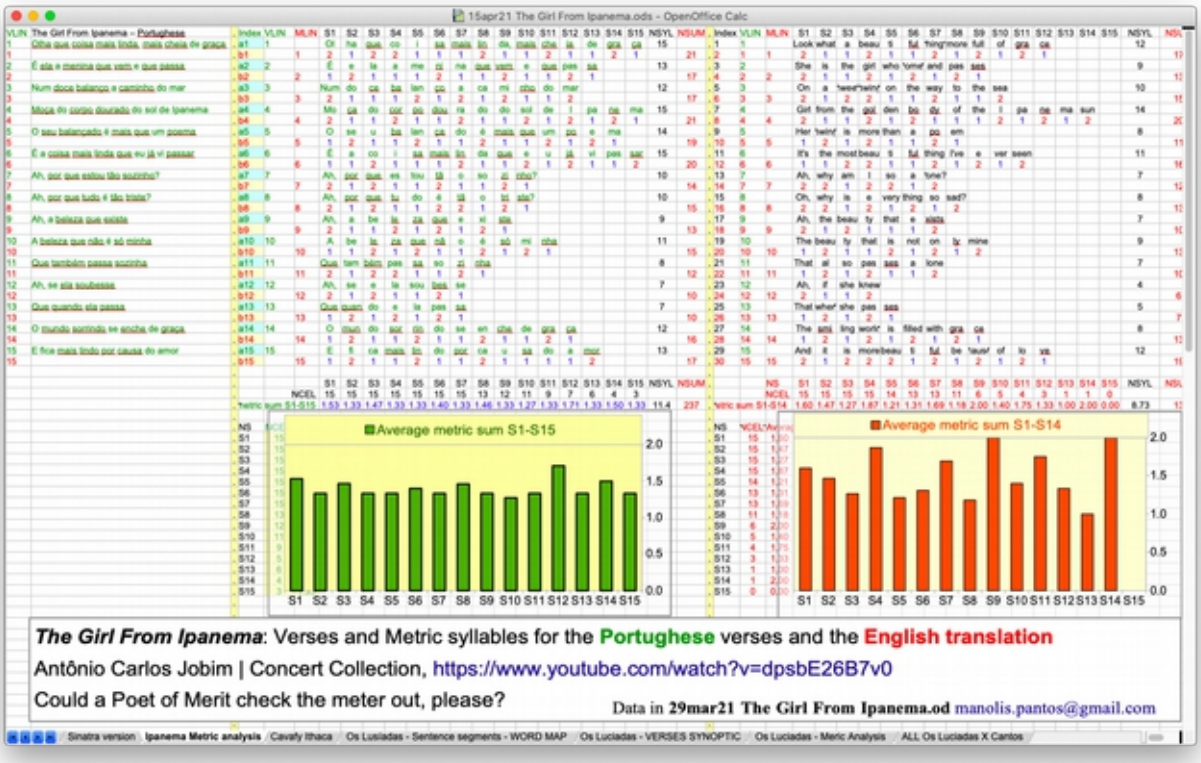

Figure 4. "The Girl from Ipanema" – Verses and metric syllables.

[25] https://sergiocaredda.eu/inspiration/i-keep-six-honest-serving-men-a-poem-by-rudyard-kipling/ [Last accessed 15 April 2021].

[26] https://en.wikipedia.org/wiki/Rabindranath_Tagore [Last accessed 15 April 2021].

[27] https://bn.wikisource.org/wiki/[´ŒÙ/P/W L W/E, XP [Last accessed 21 April 2021].

[28] https://www.poetryfoundation.org/poems/51296/ithaka-56d22eef917ec and https://lyricstranslate.com/en/ithaki-ithaki-ithaka.html-0. [Last accessed 15 April 2021].

[29] https://en.wikipedia.org/wiki/Forensic_linguistics [Last accessed 15 April 2021].

[30] https://en.wikipedia.org/wiki/Stylometry [Last accessed 15 April 2021].

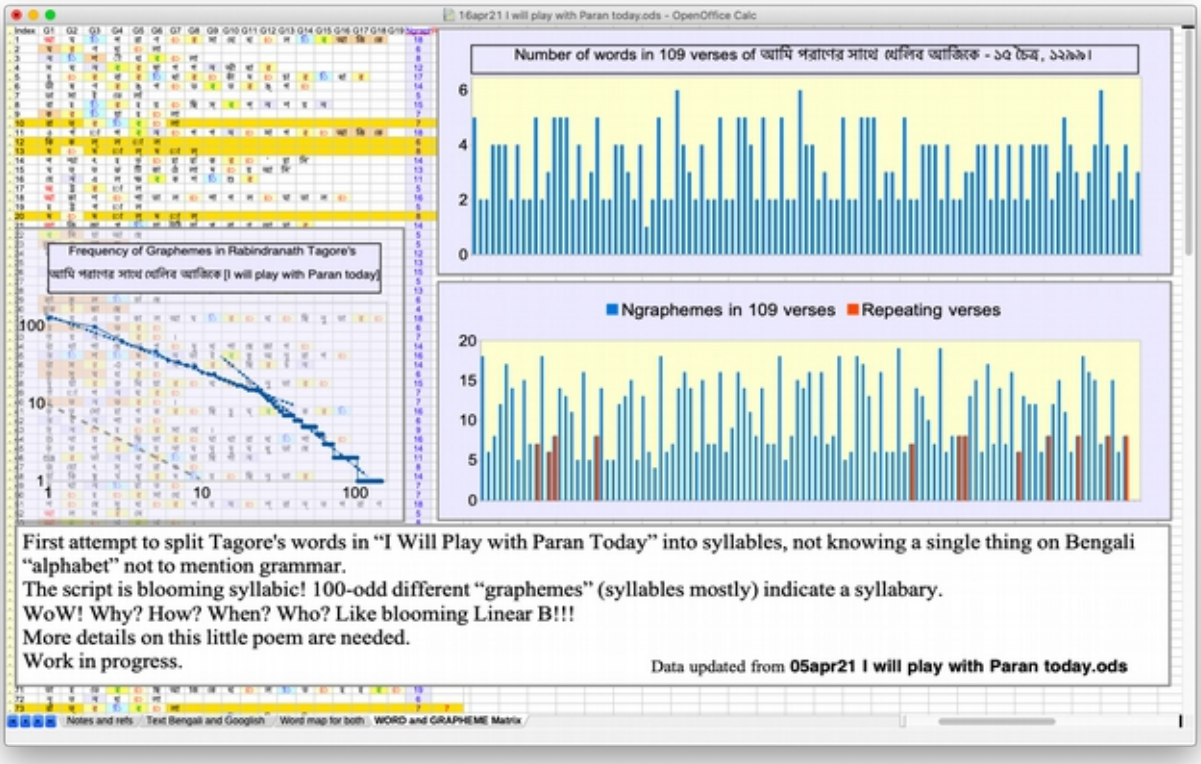

Figure 5. "Playing with Paran", Rabindranath Tagore. The Bangla script is alpha-syllabic (evolved from Sanskrit) just as Linear B is. Bangla is the official state script of Bangladesh.

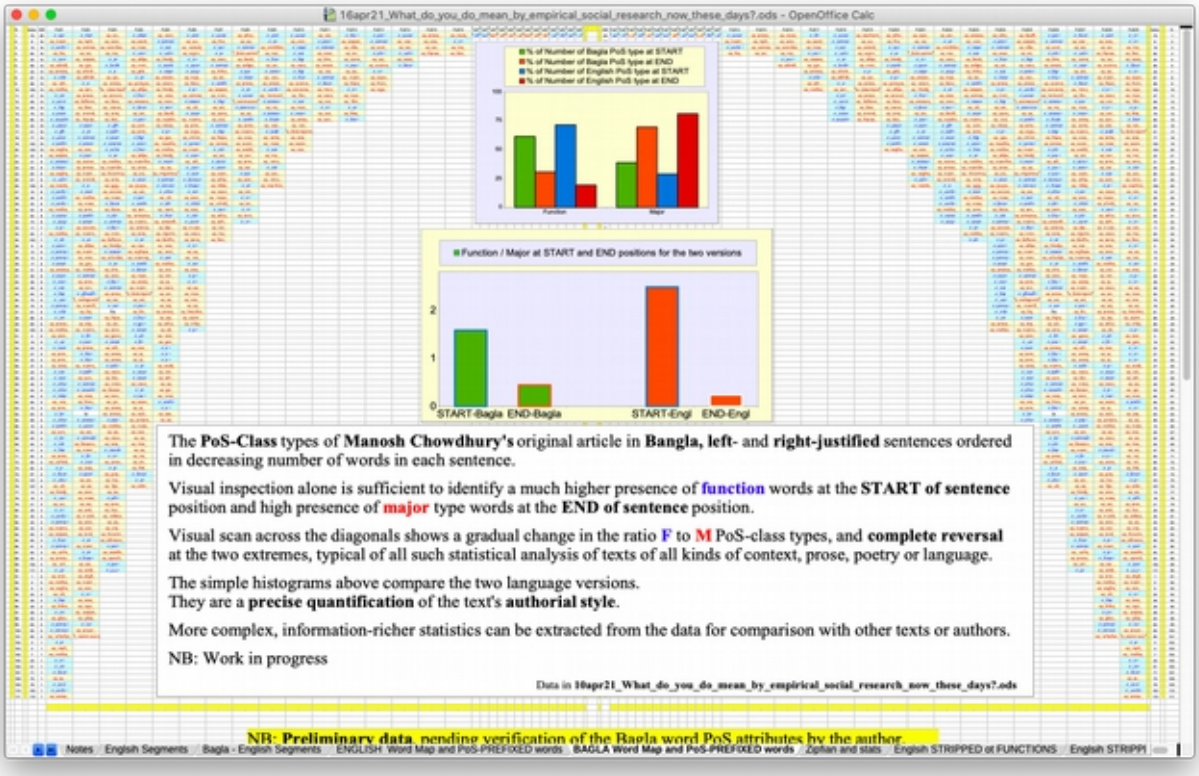

Figure 6a. Left- and right-justified PoS-Class types of original article in Bangla (Chowdhury, 2021a). Here the PoS-tagged words come into play for the statistical comparison of start or end of sentence-terminating word types.

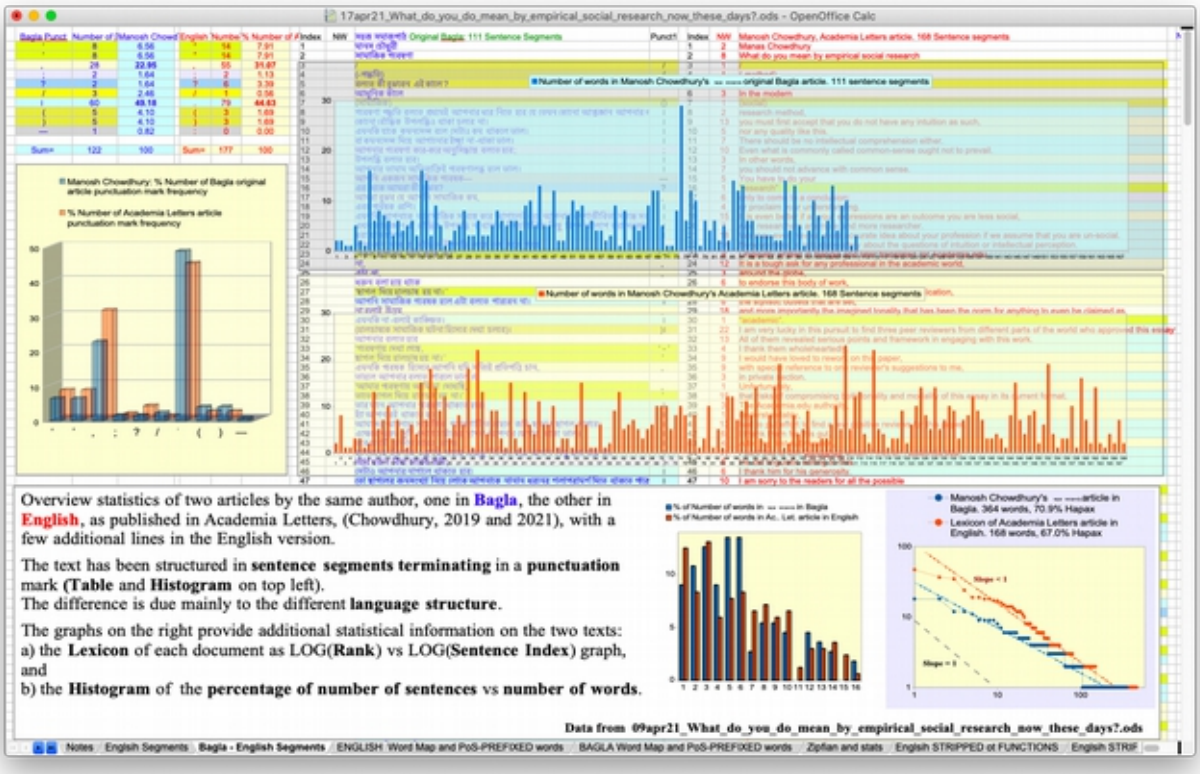

Figure 6b. The same texts as above, different aspects of analysis.

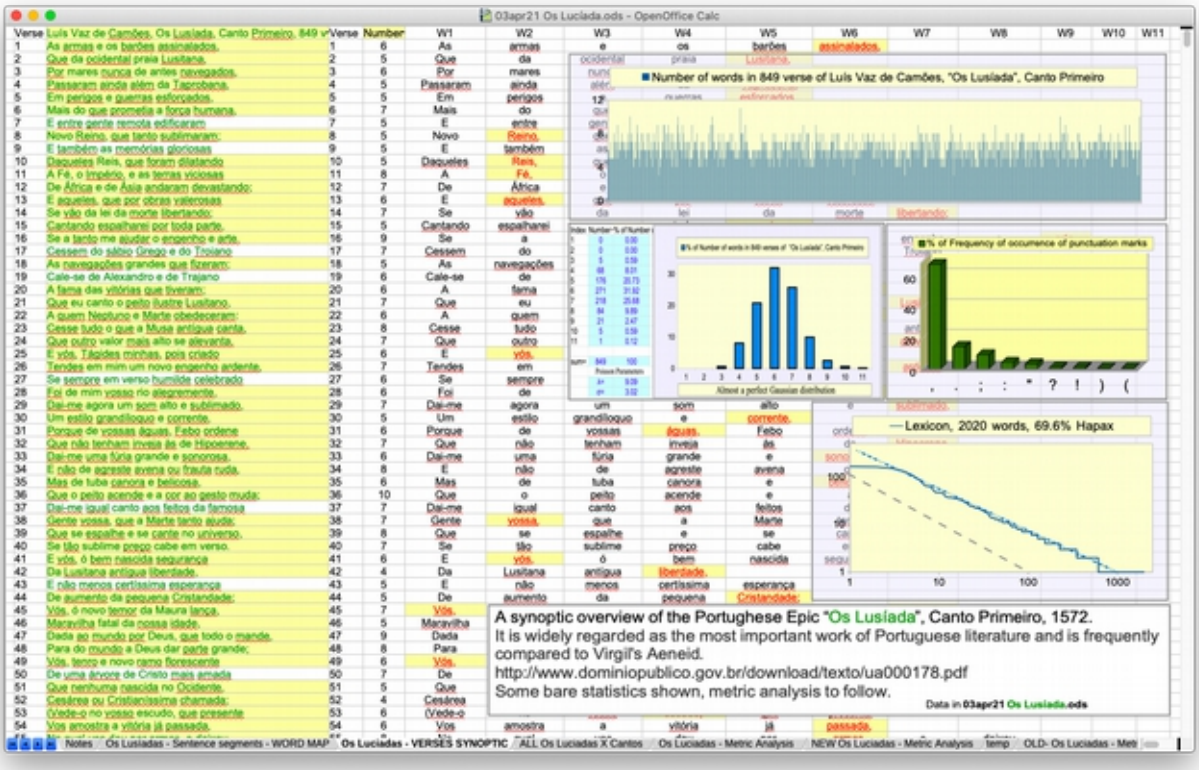

Figure 7. "Os Lusíadas": an epic poem mythologizing the circumnavigation of the globe by Vasco De Gamma, on the model of the Aeneid and the Odyssey.

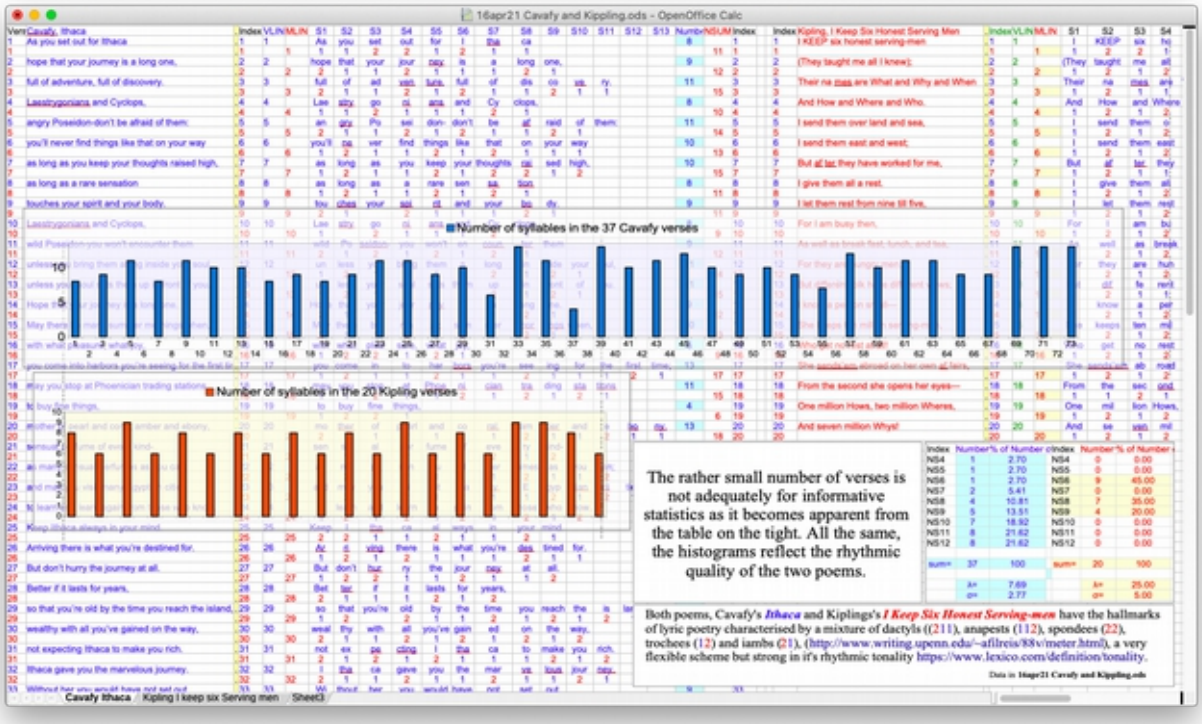

Figure 8. Cavafy's "Ithaca" and Kipling's "I Keep Six Honest Serving-Men" - Metric verses.

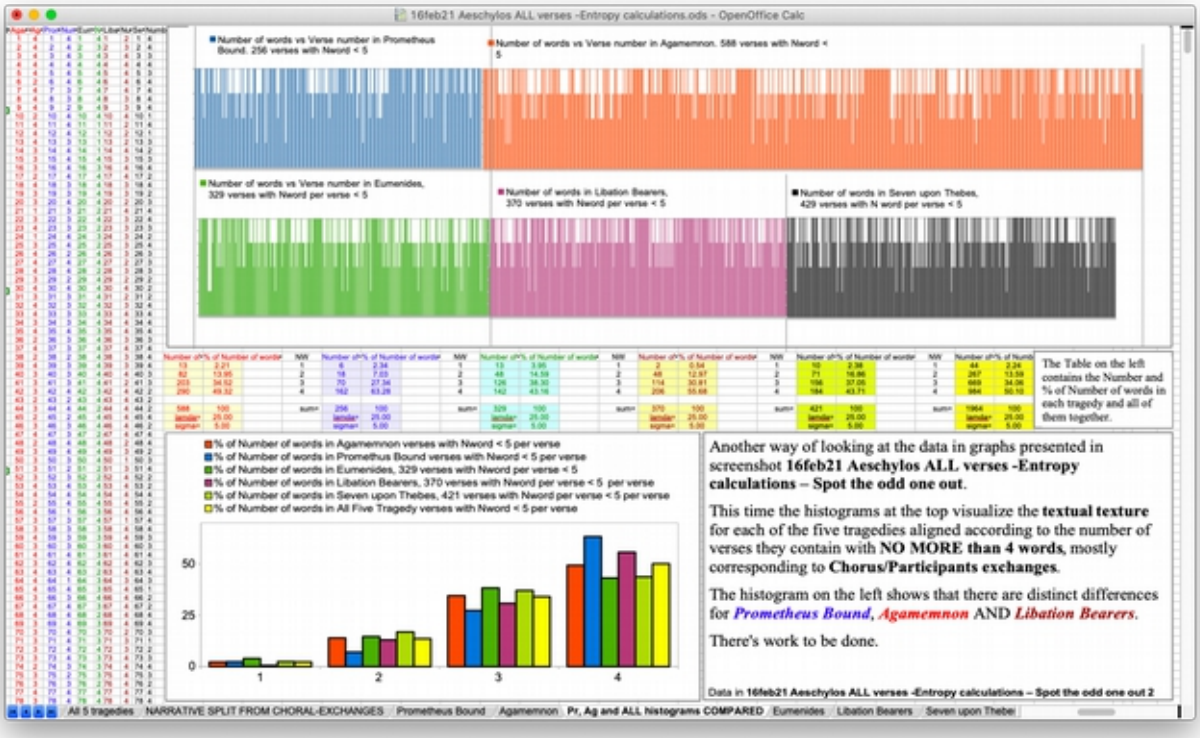

Figure 9. Overview of analysis on five of the seven surviving Aeschylus' tragedies. A rather dense screenshot of longer texts representing a fraction of the stylometric analysis processed thus far.

Closing remarks

The field of Forensic Linguistics (addressing, amongst other things, issues of author identity), has thus far employed traditional methods of literary analysis. In the case of Homer's *Iliad*/*Odyssey* this has a long history, from the days of Strabo the geographer and Alexandrian scholarship to the present (West 1999, 2011, 2014)[31].

A subfield of Forensic Linguistics has recently opened up using advanced computational methods to analyse the authorial style of Beowulf[32], Shakespeare's sonnets[33], the Federalist papers[34] and lately Aeschylus' and Euripides' tragedies (Manousakis, 2020), (Manousakis and Stamatatos, 2018).

Outsiders perhaps can not compete with these AI-type approaches, but *there is more than one way to skin a cat*, as the English saying goes[35].

The examples presented above cover only essential aspects of the methodology and not the whole spectrum of possibilities such as ngrams[36] or other "dark arts of the trade[37]". The tools employed are essentially the Apple OpenOffice toolkit, Spreadsheet plotting and colour annotated graphics (infographics[38]), TextEdit and a handful of simple pieces of code to:

a) produce a document's lexicon, the basis of further analytical processing;

b) compute the metric entropy of verse metric values and compare different author works;

c) count words in each verse or sentence segment, and create "word maps";

d) reverse the word- or letter-order in a sentence, for certain global editing operations.

Added to these is a number of sed scripts[39], specific to each document, for prefixing PoS-tags to words based on the document's lexicon.

---

[31] The work of Martin West - a giant in the field of Greek archaic and classical literature, and music - has been the key motivation and starting point of the leading author of this article, although not all Homeric scholars have agreed with West's suggestions, e.g. Gregory Nagy, chairman and co-founder of the Centre for Hellenic Studies at Harvard, and Richard Janko at the University of Michigan, https://en.wikipedia.org/wiki/Richard_Janko. [Last accessed 23 April 2021]

[32] https://amp.theguardian.com/books/2019/apr/08/beowulf-old-english-poem-work-one-author-research-suggests?CMP=share_btn_tw&__twitter_impression=true [Last accessed 16 April 2021].

[33] https://www.npr.org/templates/story/story.php?storyId=104317503&t=1618552735098 [Last accessed 16 April 2021].

[34] https://www.researchgate.net/publication/243773802_The_Federalist_Revisited_New_Directions_in_Authorship_Attribution [Last accessed 16 April 2021].

[35] https://grammarist.com/phrase/more-than-one-way-to-skin-a-cat/ [Last accessed 21 May 2021].

[36] https://en.wikipedia.org/wiki/N-gram [Last accessed 21 May 2021].

[37] This refers to good knowledge of the art of the possible in spreadsheet data handling that comes with experience.
very difficult to explain to a novice spreadsheet user.
It comes with a few years of experience.

[38] https://www.lexico.com/definition/infographic

[39] https://www.gnu.org/software/sed/manual/html_node/sed-script-overview.html [Last accessed 16 April 2021].

Three cheers for the tool makers - Dr S. D. Pantos, Dr C. Brew and Dr V. Marshall (STFC, Rutherford-Appleton Laboratory), Dr L. Kozlowski (Institute of Informatics, University of Warsaw) and G. Milne (EP's old colleague at Daresbury Laboratory, now in New Zealand) - for their generous "gifts" to the project of simple-looking but very effective codes in *awk* or *perl,* and for instructions on how to use *sed and grep scripts*. All in "computerese" but very poetical all the same (ποιητὴς, Gk. for someone who makes things, a creative word-tool-maker). Special thanks also to G. Metaxas of Athens for a plethora of golden nuggets of information about Greek technical terms (e.g. *Enjambment=Διασκελισμός)* and much more in his blogsite, https://geometax12.blogspot.com/2018/07/blog-post_9.html. We are also indebted to Drs A. Bugajska and K. Lemms for their permission to use their publications as training-sets and Dr A. Pantos for proof-reading the almost final version.

Works cited

Bugajska, A. (2021). The Future of Utopia in the Posthuman World. *Academia Letters*, Article 155, https://doi.org/10.20935/AL155.

Chowdhury, M. (2021a). [/U/(D= ? 8SZK/ /(-Q (L) SX8L =1SE8SP +# =/8X? (What do you mean by empirical social research / (- method) now these days?). *Shamprotik.com*.

Chowdhury, M. (2021b). What do you mean by empirical social research / (- method) now these days. *Academia Letters*, Article 793.https://doi.org/10.20935/AL793

Coulthard, M., Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence,* Routlege, London.

Lems, K. (2021). Music, Our Human Superpower. *Academia Letters*, Article 304, https://doi.org/10.20935/AL304.

Manousakis, N., Stamatatos, E. (2018). *Digital Scholarship in the Humanities*, Vol. 33, Issue 2, pp. 347–361, https://doi.org/10.1093/llc/fqx021

Manousakis, N. (2020). *Prometheus Bound - a Separate Authorial Trace in the Aeschylean Corpus.* Trends in classics - supplementary volumes, 98.

Ono, F.P. (2021). Seeing with my feet. *Academia Letters*, Article 401. https://doi.org/10.20935/AL401.1

Piantadosi, S.T. (2014). *Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions*. Psychon Bull Rev **21,** 1112–1130 (2014). https://doi.org/10.3758/s13423-014-0585-6

Shannon, C.E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journa*l, Volume: 27, Issue: 3. DOI: 10.1002/j.1538-7305.1948.tb01338.x

West, M.L. (1999). The Invention of Homer. *The Classical Quarterly*, Volume 49, Issue 2, pp. 364–382. DOI: https://doi.org/10.1093/cq/49.2.364

West, M.L. (2011), *The Making of the Iliad*, Oxford University Press, Published to Oxford Scholarship Online: DOI:10.1093/acprof:osobl/9780199590070.001.0001

West, M.L. (2014), *The Making of the Odyssey*, Oxford University Press. Published to Oxford Scholarship Online: March 2015, DOI:10.1093/acprof:oso/9780198718369.001.0001

….....

Note for the AL editor and reviewers:

Dear colleagues,

This is the first in a series of articles to be submitted in the near future, two of which (footnotes 12 and 13) are at the top of a long list, product of work over the last 10 years since my retirement from STFC, Daresbury Laboratory.

The style of this paper is somewhat, radically perhaps, different from other publications I have harvested from Academia.edu for which I am grateful as a key source of information related to my work on a variety of topics authored by scholars in a wide range of disciplines.

Indeed, AL was the route of my first contact with the two co-authors and the two contributors to the "training set" of texts where the common denominator was length and format (style) but completely different context.

I may have exceeded the number of infographics normally present in such brief *Letters*. If you find that this is a diabolical liberty I took, there may be alternative ways of reducing them and accomodating them as, for instance, "Additional Material" as for articles in *Nature Letters, Physics Review Letters* or other academic journals of great impact factor (PNAS etc). I could add a reference to my personal Academia.edu website (Work in Progress) or a blog (I much prefer website), which means extra work I would prefer to avoid so I can focus on the much more interesting, in my humble opinion, work in the pipeline.

Another alternative may be to resubmit the substance of the article as a longer article with the same infographics where fair justice to the methodology can be made. Not so long ago I received an announcement that the editors were considering having a different stream for longer articles as in other internet journals. This would be very welcome for me and collaborators otherwise we would have to resort to other Open Access internet journals. I am happy with you and getting happier as my experience of use, content and options available is steadily increasing.

Looking forward to your decision,

Yours sincerely

Dr Emmanuel (Manolis) Pantos,

BSc (National Kapodistrian University, Athens, Greece, Physiscs, 1968),

MSc and PhD (Manchester University, UK), Physics, 1969-73).

manolis.pantos@gmail.com

….....

[Total word count excluding header (Title, Affiliations and Keywords, footnotes and references), 1582 words/tokens on last count, 30may21.]