ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Προσδιορισμός ταυτότητας άγνωστων κειμένων με τη χρήση μηχανικής μάθησης»

«Εικόνα»

Του φοιτητή
Αργυριάδη Κωνσταντίνου
Αρ. Μητρώου: 2020015

Επιβλέπων
Ιλιούδης Χρήστος
Καθηγητής

Ημερομηνία 23/01/2026

Τίτλος Δ.Ε. Προσδιορισμός ταυτότητας άγνωστων κειμένων με τη χρήση μηχανικής μάθησης
Κωδικός Δ.Ε. 25117
Ονοματεπώνυμο φοιτητή Κωνσταντίνος Αργυριάδης
Ονοματεπώνυμο εισηγητή Ιλιούδης Χρήστος
Ημερομηνία ανάληψης Δ.Ε. 14/02/2025
Ημερομηνία περάτωσης Δ.Ε. 23/01/2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Πρόλογος

Το θέμα που επέλεξα συμπίπτει με τα ενδιαφέροντά μου πάνω στην γλωσσολογία και γενικότερα στη γλώσσα. Μέσα από τη μελέτη του κατανόησα τον τρόπο με τον οποίο χαρακτηριστικά της γλώσσας μπορούν να ποσοτικοποιηθούν και να αναλυθούν υπολογιστικά. Καταπιάστηκα με βασικές έννοιες και εργαλεία τα οποία χρησιμοποιούνται σε LLMs και την υπολογιστική γλωσσολογία.

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει την υφολογική ποικιλία και τις εσωτερικές συγγένειες του Ιπποκρατικού corpus μέσω υπολογιστικών μεθόδων υφομετρίας και μη εποπτευόμενης μηχανικής μάθησης. Το corpus αποτελείται από ανώνυμες ιατρικές πραγματείες, οι οποίες παρουσιάζουν σημαντική θεματική, γλωσσική και υφολογική ετερογένεια, γεγονός που έχει οδηγήσει σε μακροχρόνια φιλολογική συζήτηση σχετικά με την πατρότητα, τη χρονολόγηση και την ομαδοποίησή τους. Η μελέτη βασίζεται σε χαρακτηριστικά χαμηλού επιπέδου, συγκεκριμένα character n-grams και συχνότερες λέξεις (Most Frequent Words), τα οποία έχουν αποδειχθεί ιδιαίτερα κατάλληλα για υφομετρική ανάλυση σε μικρά και μορφολογικά πλούσια corpora. Τα κείμενα αναπαρίστανται ως διανύσματα υψηλής διαστασιμότητας και αναλύονται με τη χρήση της Evidence Accumulation Clustering, η οποία συνδυάζει πολλαπλές εκτελέσεις k-means σε έναν πίνακα συνσυσχέτισης και εφαρμόζει ιεραρχική συσταδοποίηση για την ανάδειξη σταθερών υφολογικών σχέσεων. Τα αποτελέσματα οπτικοποιούνται με δενδρογράμματα και πολυδιάστατη κλιμάκωση, επιτρέποντας τη διερεύνηση τόσο τοπικών όσο και γενικότερων υφολογικών δομών. Η ανάλυση αναδεικνύει επαναλαμβανόμενες και ερμηνεύσιμες συστάδες κειμένων, οι οποίες σε μεγάλο βαθμό συμφωνούν με τις εκτιμήσεις της φιλολογικής έρευνας, ιδίως ως προς τη διάκριση τεχνικών, χειρουργικών και θεωρητικών πραγματειών. Η εργασία καταδεικνύει ότι οι υπολογιστικές μέθοδοι μπορούν να λειτουργήσουν συμπληρωματικά προς τη φιλολογική ανάλυση, προσφέροντας ένα ποσοτικό και αναπαραγώγιμο εργαλείο για τη μελέτη της αρχαίας ιατρικής γραμματείας.

«Identifying texts of unknown author with machine learning»

Argiriadis Konstantinos

Abstract

The present thesis examines stylistic variation and internal affinities within the Hippocratic corpus through computational stylometry and unsupervised machine learning methods. The corpus consists of anonymous medical treatises that exhibit significant thematic, linguistic, and stylistic heterogeneity, a fact that has given rise to long-standing philological debate concerning their authorship, dating, and classification. The study relies on low-level features, specifically character n-grams and Most Frequent Words, which have proven particularly suitable for stylometric analysis in small and morphologically rich corpora. The texts are represented as high-dimensional vectors and analyzed using Evidence Accumulation Clustering, which combines multiple k-means runs into a co-association matrix and applies hierarchical clustering to reveal stable stylistic relationships. The results are visualized through dendrograms and multidimensional scaling, allowing for the exploration of both local and global stylistic structures. The analysis highlights recurring and interpretable text clusters that largely align with findings from philological research, particularly with regard to the distinction between technical, surgical, and theoretical treatises. The thesis demonstrates that computational methods can function complementarily to philological analysis, offering a quantitative and reproducible tool for the study of ancient medical literature.

Ευχαριστίες

Ευχαριστώ τον επιβλέποντα καθηγητή μου, Χρήστο Ιλιούδη, για την πολύτιμη καθοδήγηση.

Περιεχόμενα

Κατάλογος Σχημάτων

Συντομογραφίες

Δ.Ε. – Διπλωματική Εργασία

ΔΙ.ΠΑ.Ε. – Διεθνές Πανεπιστήμιο της Ελλάδος

MFW – Most Frequent Words

EAC – Evidence Accumulation Clustering

MDS – Multidimensional Scaling

TF-IDF – Term Frequency – Inverse Document Frequency

CTS – Canonical Text Services

URN – Uniform Resource Name

PoS – Part of Speech

ANN – Artificial Neural Networks

MLP – Multilayer Perceptron

SVM – Support Vector Machine

DBN – Deep Belief Network

AST – Abstract Syntax Tree

PAN – Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (benchmark)

LLM – Large Language Model

GMM – Gaussian Mixture Models

DBSCAN – Density-Based Spatial Clustering of Applications with Noise

Κεφάλαιο 1ο: Εισαγωγή

1.1 Το Ιπποκρατικό Ζήτημα

Το ιπποκρατικό corpus έχει αποτελέσει αντικείμενο εκτενούς φιλολογικής μελέτης. Πρόκειται για ένα ετερογενές σύνολο 60 περίπου ιατρικών κειμένων, τα οποία πραγματεύονται μεγάλη ποικιλία θεμάτων, όπως η ανατομία, η παθολογία, η θεραπεία, η γυναικολογία, η ιδεολογία. Ενώ είναι όλα γραμμένα στην ιωνική διάλεκτο, παρουσιάζουν σημαντικές διαφορές μεταξύ τους ως προς τη δομή, τη σύνταξη και το λεξιλόγιο, γεγονός που τα τοποθετεί και σε διαφορετικές εποχές. Η έκτασή τους παρουσιάζει μεγάλη διακύμανση. Τα κείμενα είναι όλα ανώνυμα και ως επί το πλείστον δε συναντώνται μέσα τους ονομαστικές αναφορές σε πρόσωπα και γιατρούς της εποχής. Είναι δύσκολο να γίνει λόγος για γνησιότητα ή μη των κειμένων, υπό το πρίσμα της ερώτησης αν έχουν γραφτεί ή όχι από τον ίδιο τον Ιπποκράτη[1]. Γενικά, η προέλευση του ογκώδους και πολύπτυχου υλικού του ιπποκρατικού corpus παραμένει σ' έναν μεγάλο βαθμό απροσδιόριστη.

Σύμφωνα με την Craik, ήδη από την αρχαιότητα είχε επισημανθεί ότι η συλλογή αυτή δεν μπορεί να αποτελεί έργο ενός και μόνο συγγραφέα· η ετερογένεια των κειμένων καθιστά κάθε μονοσήμαντη απόδοση στον ιστορικό Ιπποκράτη μεθοδολογικώς επισφαλή. Οι πραγματείες διαφέρουν ως προς το είδος (πλήρη θεωρητικά δοκίμια, συνόψεις, σημειώσεις, αφηγηματικές καταγραφές περιπτώσεων, κανονιστικά κείμενα), ενώ στην εσωτερική τους διάρθρωση περιέχουν συχνά ίχνη σύνθεσης από πολλαπλές πηγές, επιβεβαιώνοντας ότι μεγάλο μέρος του corpus διαμορφώθηκε μέσα σε ένα περιβάλλον συλλογικής ιατρικής γνώσης και ανταλλαγής, χωρίς σαφείς κανόνες για την έννοια της «δημοσίευσης» ή της «πνευματικής ιδιοκτησίας».

Η πρώτη συστηματική απόπειρα ταξινόμησης των ιπποκρατικών πραγματειών εντοπίζεται στο έργο του Ερωτιανού, ο οποίος διαμόρφωσε μία θεματική οργάνωση της συλλογής, βασισμένη σε έναν προγενέστερο κατάλογο του Βακχείου της Τανάγρας. Ο Ερωτιανός, με κριτήριο το θέμα της κάθε πραγματείας, διακρίνει το ιπποκρατικό corpus σε πέντε κατηγορίες [1, σ. xxvi]:

- «Σημειωτικά» (σχολιασμός και ερμηνεία συμπτωμάτων).
- «Αιτιολογικά» και «φυσικά».
- Θεραπευτικά/πρακτικά (εδώ υπάρχουν δύο υποκατηγορίες, τα χειρουργικά και τα διαιτητικά).
- Μικτά
- Έργα πάνω στην «τέχνη» της ιατρικής (δεοντολογικά/ιδεολογικά, εδώ περιλαμβάνεται και ο Όρκος)

Περίπου έναν αιώνα μετά, ο Γαληνός επιχειρεί δικούς του σχολιασμούς πάνω στα ιπποκρατικά έργα. Τον απασχόλησε, μάλιστα, έντονα το ζήτημα της αυθεντικότητας των κειμένων ως δημιουργημάτων του Ιπποκράτη. Ξέρουμε πως έγραψε ένα βιβλίο πάνω στα ιπποκρατικά κείμενα, το οποίο όμως δεν έχει διασωθεί[1, σ. xxi]. Η συμβολή του αφορά περισσότερο την κριτική-ερμηνευτική ταξινόμηση παρά μια συστηματική θεματική κατανομή.

Κατά τη βυζαντινή περίοδο, τα ιπποκρατικά κείμενα αναπαράγονταν σε ποικίλους κώδικες, χωρίς ενιαίο κανόνα ως προς τη σειρά ή ακόμη και το περιεχόμενο. Οι δύο βασικές οικογένειες χειρογράφων, των M (Marcianus gr. 269) και V (Vaticanus gr. 276), διαφέρουν σημαντικά μεταξύ τους[1, σ. xxiv].

Πάνω στο πρότυπο του Ερωτιανού βασίστηκαν όχι μόνο οι εκδότες της Αναγέννησης, αλλά και η εμβληματική έκδοση του Littré του 19ου αιώνα.

Σε αντιδιαστολή με τις αρχαίες προσεγγίσεις ταξινόμησης, η σύγχρονη μελέτη της Craik αντιμετωπίζει το ιπποκρατικό corpus με μεγαλύτερη επιφύλαξη ως προς οποιαδήποτε συστηματική κατηγοριοποίηση. Η Craik επιλέγει να παρουσιάσει τις πραγματείες αλφαβητικά, αποφεύγοντας να εισαγάγει εκ των προτέρων υποθέσεις για κοινές συγγραφικές καταβολές ή θεματικά σύνολα. Προτείνει βέβαια, για λόγους πρακτικότητας, επτά ευρείες θεματικές ομάδες (επιστημονικές αρχές, ανατομία και φυσιολογία, νοσολογία–παθολογία–θεραπεία, χειρουργική, περιστατικά και σημεία, γυναικολογία και εμβρυολογία, καθοδήγηση και ιατρικά ιδεώδη), επισημαίνοντας πως η ταξινόμηση αυτή δεν συνεπάγεται κοινό συγγραφέα ούτε υποδηλώνει εσωτερική ομοιογένεια· αντιθέτως, θεωρεί ότι τα περισσότερα έργα υπερβαίνουν τα όρια μίας μόνο κατηγορίας και ότι κάθε τέτοιος διαχωρισμός παραμένει, σε κάποιο βαθμό, τεχνητός και συμβατικός. Η προσέγγιση αυτή διαφοροποιείται από εκείνη του Jouanna, ο οποίος, αν και αναγνωρίζει την ετερογένεια του corpus, επιχειρεί να αναδείξει ιστορικά και φιλολογικά κριτήρια που επιτρέπουν την ανασύσταση επιμέρους ρευμάτων μέσα στην ιπποκρατική παράδοση. Έτσι, ο Jouanna προτείνει διακρίσεις που σχετίζονται με σχολές (Κώα και Κνιδία), με τη φυσιολογική ή πρακτική κατεύθυνση των κειμένων και με πιθανές υφολογικές συγγένειες[2, σσ. 35–50]. Ενώ η Craik προκρίνει μία μη δεσμευτική, λειτουργική ομαδοποίηση που υπογραμμίζει τη ρευστότητα του corpus, ο Jouanna επιχειρεί μία περισσότερο ιστορικο-κριτική και αυστηρή αρχιτεκτονική του, επιδιώκοντας την ανάδειξη διακριτών παραδόσεων και θεωρητικών κατευθύνσεων στο εσωτερικό του. Για τις ανάγκες της εργασίας θα βασιστούμε κυρίως στην ανάλυση της Craik.

1.2 Υφομετρία

Η υφομετρία (stylometry) αποτελεί την ποσοτική μελέτη του ύφους, βασισμένη στη στατιστική ανάλυση μετρήσιμων γλωσσικών χαρακτηριστικών ενός κειμένου. Οι απαρχές της ως τέτοιας εντοπίζονται στα μέσα του 19ου αιώνα. Ήδη το 1851 ο Augustus de Morgan είχε προτείνει ότι διαφορές στη συχνότητα των μηκών των λέξεων θα μπορούσαν να αξιοποιηθούν για την επίλυση συγγραφικών αμφισβητήσεων. Η ιδέα αυτή οδήγησε, λίγες δεκαετίες αργότερα, στην πρώτη συστηματική προσπάθεια χειροκίνητης ποσοτικής ανάλυσης από τον Thomas C. Mendenhall, ο οποίος στα τέλη του 19ου αιώνα συνέκρινε τις κατανομές μήκους λέξεων στα έργα των Bacon, Marlowe και Shakespeare, με στόχο να διερευνήσει την πραγματική πατρότητα των σαιξπηρικών δραμάτων[3]. Ορόσημο για τη σύγχρονη υφομετρία αποτελούν οι μελέτες των Mosteller και Wallace πάνω στα Federalist Papers[4], όπου για πρώτη φορά εφαρμόστηκαν συστηματικά υπολογιστικές μέθοδοι για τη διερεύνηση συγγραφικής ταυτότητας.

Πρόκειται για μία προσέγγιση που έχει αξιοποιηθεί ευρέως στη διερεύνηση ζητημάτων συγγραφικής ταυτότητας και ομοιογένειας κειμένων. Στο πλαίσιο του Ιπποκρατικού Ζητήματος, όπου η πολυμορφία των πραγματειών και η αβεβαιότητα της γνησιότητας αποτελούν καίρια προβλήματα, η υφομετρία προσφέρει ένα χρήσιμο συμπληρωματικό εργαλείο που μπορεί να υποστηρίξει (ή να αμφισβητήσει) παραδοσιακές φιλολογικές εκτιμήσεις. Η ανάλυση συνήθως εστιάζει σε χαρακτηριστικά όπως η συχνότητα λειτουργικών λέξεων, η λεξιλογική ποικιλία, το μήκος προτάσεων και η κατανομή σταθερά χρησιμοποιούμενων μορφοσυντακτικών μοτίβων. Όπως θα δούμε παρακάτω, με την ανάπτυξη της υπολογιστικής γλωσσολογίας και της τεχνητής νοημοσύνης, οι υφομετρικές μέθοδοι έχουν επεκταθεί σε πιο σύνθετα μοντέλα ταξινόμησης και ομαδοποίησης.

1.3 Στόχοι και διάρθρωση

Στην παρούσα εργασία, η υφομετρική ανάλυση αξιοποιείται με στόχο να εξεταστεί κατά πόσον ορισμένα κείμενα του ιπποκρατικού corpus παρουσιάζουν συγκλίσεις ή αποκλίσεις που μπορεί να υποδηλώνουν κοινή προέλευση ή διαφορετικές γραφικές παραδόσεις. Τα αποτελέσματα των πειραμάτων θα αντιπαραβληθούν με τα πορίσματα της φιλολογικής έρευνας, τόσο της σύγχρονης όσο και της αρχαίας που έχει αναφερθεί.

Η διάρθρωση της εργασίας έχει ως εξής:

- Στο δεύτερο κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο της υφομετρικής ανάλυσης. Αρχικά εξετάζονται οι βασικές κατηγορίες υφομετρικών χαρακτηριστικών, όπως οι λειτουργικές λέξεις, τα λεξικολογικά χαρακτηριστικά, τα character n-grams και τα μορφοσυντακτικά χαρακτηριστικά, με αναφορά στα πλεονεκτήματα και τους περιορισμούς κάθε προσέγγισης. Στη συνέχεια γίνεται επισκόπηση των τεχνικών μηχανικής μάθησης που χρησιμοποιούνται στη σύγχρονη υφομετρία, με διάκριση μεταξύ εποπτευόμενων και μη εποπτευόμενων μεθόδων, καθώς και των embedding-based προσεγγίσεων. Το κεφάλαιο λειτουργεί ως γέφυρα μεταξύ της φιλολογικής προβληματικής και της υπολογιστικής μεθοδολογίας που ακολουθεί.
- Το τρίτο κεφάλαιο περιγράφει αναλυτικά το μεθοδολογικό πλαίσιο της έρευνας. Αρχικά τονίζεται ο περιορισμός στον όγκο του υλικού και αιτιολογείται η επιλογή μη εποπτευόμενων μεθόδων λόγω της απουσίας αξιόπιστων ετικετών συγγραφέα στο Ιπποκρατικό corpus. Ακολουθεί εκτενής παρουσίαση του αλγοριθμικού υποβάθρου, συμπεριλαμβανομένων των μέτρων ομοιότητας, των διαμεριστικών και ιεραρχικών αλγορίθμων συσταδοποίησης, των ensemble μεθόδων και ειδικότερα της Evidence Accumulation Clustering. Στη συνέχεια περιγράφεται βήμα προς βήμα το υπολογιστικό pipeline, από την ανάκτηση και προεπεξεργασία των κειμένων, την εξαγωγή χαρακτηριστικών (character n-grams και Most Frequent Words), έως τη συσταδοποίηση και την οπτικοποίηση των αποτελεσμάτων.
- Το τέταρτο κεφάλαιο παρουσιάζει και σχολιάζει τα αποτελέσματα της ανάλυσης. Αρχικά εξετάζονται τα αποτελέσματα που προκύπτουν από τα character n-grams και στη συνέχεια εκείνα που βασίζονται στις συχνότερες λέξεις, με και χωρίς φιλτράρισμα θεματικά φορτισμένων όρων. Τα δενδρογράμματα και οι MDS οπτικοποιήσεις αναλύονται ως προς τη σταθερότητα και τη δομή των συστάδων και αντιπαραβάλλονται με τις καθιερωμένες ταξινομήσεις της σύγχρονης φιλολογικής έρευνας. Το κεφάλαιο ολοκληρώνεται με συγκριτική αποτίμηση των δύο προσεγγίσεων, αναδεικνύοντας κοινά και αποκλίνοντα μοτίβα.
- Στο πέμπτο κεφάλαιο σκιαγραφούνται κατευθύνσεις για μελλοντική έρευνα (επέκταση του υλικού της έρευνας, αξιοποίηση διαφορετικών αναπαραστάσεων των κειμένων, βασισμένων σε μορφοσυντακτικά και ρυθμικά χαρακτηριστικά).
- Το έκτο κεφάλαιο συνοψίζει τα βασικά συμπεράσματα της έρευνας, επαναξιολογώντας τους αρχικούς στόχους της μελέτης.

Κεφάλαιο 2ο: Τεχνικές υφομετρίας

2.1 Υφομετρικά χαρακτηριστικά (features)

Ένα εκ των δύο κύριων μερών της υφομετρικής διαδικασίας αποτελεί η επιλογή των χαρακτηριστικών του κειμένου που θα μετρηθούν και πάνω στα οποία θα βασιστεί η ανάλυση. Υπάρχει πληθώρα τέτοιων χαρακτηριστικών και για την επιλογή των βέλτιστων πρέπει σε κάθε περίπτωση να λαμβάνονται υπόψη οι ιδιαιτερότητες των κειμένων προς διερεύνηση. Παρακάτω παρατίθενται και εξετάζονται με παραδείγματα οι βασικές κατηγορίες που συναντώνται στη σύγχρονη βιβλιογραφία.

2.1.1 Λειτουργικές λέξεις (Function words)

Οι λειτουργικές λέξεις (function words) θεωρούνται ισχυρό και αποτελεσματικό χαρακτηριστικό για την υφομετρία. Εξαιτίας της υψηλής τους συχνότητας και της φύσης τους, που καθιστά δύσκολη τη συνειδητή ρύθμισή τους και άρα τις εκούσιες αλλοιώσεις ύφους, αποτελούν σημαντική μέτρηση για τον προσδιορισμό του ύφους ενός συγγραφέα. Επιπλέον, τέτοιες λέξεις είναι σε μεγάλο βαθμό ανεξάρτητες από το θέμα και το είδος του κειμένου και τα ποσοστά εμφάνισής τους αναμένονται σταθερά στα διάφορα κείμενα που μπορεί να γράψει ο ίδιος συγγραφέας[5].

Οι δείκτες που βασίζονται σε λειτουργικές λέξεις χρησιμοποιούν μία προσέγγιση της αντίληψης bag-of-words, σύμφωνα με την οποία το κείμενο είναι ένα σύνολο ανεξάρτητων λέξεων. Η παραδοχή αυτή αγνοεί εντελώς τη συντακτική δομή και τις ακολουθιακές σχέσεις που υπάρχουν στο κείμενο. Στο πλαίσιο αυτό, ερευνητές όπως ο Hoover[6] προτείνουν ότι η αξιοποίηση της ακολουθιακής πληροφορίας, δηλαδή των πραγματικών συνδυασμών και διαδοχών λέξεων μέσα στο κείμενο, θα μπορούσε να οδηγήσει σε πιο προηγμένους υφομετρικούς δείκτες.

Τα παραπάνω οδήγησαν τους Boukhaled και Ganascia στη σύγκριση της bag-of-words προσέγγισης με την ακολουθιακή, εφαρμόζοντας τις δύο εναλλακτικές σε κείμενα κλασικών Γάλλων συγγραφέων. Διαπιστώθηκε όμως πως η πρώτη, παρά τους περιορισμούς της, απέφερε καλύτερα αποτελέσματα, ενισχύοντας τις αποφάνσεις για την αποτελεσματικότητα της bag-of-words προσέγγισης.

2.1.2 Λεξικολογικά χαρακτηριστικά

Στον πυρήνα αυτής της κατηγορίας βρίσκονται οι έννοιες των tokens (το σύνολο όλων των λέξεων ενός κειμένου, συμπεριλαμβανομένων των επαναλήψεων) και των types (ο αριθμός των διαφορετικών λέξεων). Με βάση αυτά προκύπτει ο λόγος types/tokens, καθώς και πιο εξελιγμένοι δείκτες λεξιλογικής ποικιλίας, όπως οι δείκτες Yule's K και Herdan's C, οι οποίοι επιχειρούν να μετρήσουν πόσο πλούσιο ή επαναληπτικό είναι το λεξιλόγιο ενός κειμένου, περιορίζοντας τα προβλήματα που δημιουργεί η διακύμανση του μήκους των κειμένων. Άλλα λεξικολογικά χαρακτηριστικά είναι τα hapax legomena (λέξεις που εμφανίζονται μόνο μία φορά), τα οποία χρησιμοποιούνται ως δείκτης λεξιλογικής ποικιλίας, καθώς και η συχνότητα τεχνικών ή εξειδικευμένων όρων, όταν το κείμενο ανήκει σε ειδικό αντικείμενο. Τέτοιου τύπου μετρήσεις προσεγγίζουν το κείμενο από την άποψη του λεξιλογίου, κατά πόσο δηλαδή αυτό ποικίλει ή κατά πόσο ένας συγγραφέας επανέρχεται στις ίδιες λέξεις και όρους.

2.1.3 Χαρακτηριστικά βασισμένα σε χαρακτήρες (character-based)

Τα character-based χαρακτηριστικά αποτελούν μία ιδιαίτερα αποδοτική κατηγορία υφομετρικών δεικτών, καθώς βασίζονται όχι σε λέξεις ή συντακτικές δομές, αλλά στην κατανομή μεμονωμένων χαρακτήρων και ακολουθιών χαρακτήρων (character n-grams). Η βασική τους αρχή είναι ότι οι μικρές μονάδες, όπως γράμματα, δίψηφα, σημεία στίξης, συγκροτούν χαρακτηριστικά μοτίβα στο γραπτό ύφος ενός συγγραφέα, τα οποία είναι δύσκολο να ελεγχθούν συνειδητά και παραμένουν σχετικά σταθερά σε διαφορετικά κείμενα.

Αξίζει να σημειωθεί ότι η σημασία τους στη σύγχρονη υφομετρία επιβεβαιώνεται και από το έργο του Koppel, ο οποίος ανέδειξε ότι η συγγραφική ταυτότητα μπορεί να ανιχνευθεί ακόμη και μέσω «επιφανειακών» δεικτών[8]. Ο Koppel δείχνει πως η στυλιστική ιδιομορφία εκφράζεται κυρίως μέσα από σταθερά μοτίβα μικρών γλωσσικών μονάδων τα οποία παραμένουν ανθεκτικά σε θεματικές ή ειδολογικές μεταβολές και άρα λειτουργούν ως αξιόπιστοι δείκτες ύφους. Υποστηρίζει ότι τέτοιοι (μη σημασιολογικοί) δείκτες διακρίνουν αποτελεσματικά τα συγγραφικά ύφη όχι μέσω των επιφανειακών διαφοροποιήσεων του περιεχομένου, κάτι που θεμελιώνει και τη μέθοδο unmasking, όπου εξετάζεται η ανθεκτικότητα των ταξινομητών καθώς αφαιρούνται σταδιακά τα πιο διακριτικά χαρακτηριστικά όπως το θέμα ή το είδος του κειμένου.

Τα n-grams έχουν το πλεονέκτημα ότι παρακάμπτουν την ανάγκη λεξικής ή μορφολογικής ανάλυσης, γεγονός που τα καθιστά ιδιαίτερα χρήσιμα σε γλώσσες με πλούσια μορφολογία ή περιορισμένη υπολογιστική υποστήριξη, καθώς και σε περιβάλλοντα όπου τα εργαλεία επισημείωσης παρουσιάζουν χαμηλή ακρίβεια.

2.1.4 Μορφοσυντακτικά χαρακτηριστικά

Τα μορφοσυντακτικά χαρακτηριστικά αποτυπώνουν τον τρόπο με τον οποίο οργανώνεται γραμματικά και συντακτικά ο λόγος ενός συγγραφέα. Οι μετρήσεις που αντλούνται από τα εν λόγω χαρακτηριστικά αφορούν τη συχνότητα γραμματικών κατηγοριών (π.χ ρήματα, αντωνυμίες), τις αναλογίες ρηματικών χρόνων και εγκλίσεων, την προτίμηση παρατακτικής ή υποτακτικής σύνδεσης, όπως επίσης δείκτες συντακτικής πολυπλοκότητας (μέσο μήκος πρότασης, κατανομή δευτερευουσών προτάσεων).

Ο Gorman[7] αναλύει τη χρησιμότητα της μορφοσυντακτικής επισημείωσης ως υφομετρικού εργαλείου, διακρίνοντας τρεις βασικές κατηγορίες χαρακτηριστικών: τα μορφολογικά, τα συντακτικά και τα σύνθετα (μορφοσυντακτικά n-grams).

- Ως μορφολογικά χαρακτηριστικά θεωρεί τις γραμματικές τιμές που αποδίδονται σε κάθε λέξη**,** όπως μέρος του λόγου (PoS – Part of Speech) , γένος, αριθμός, πτώση, χρόνος, έγκλιση κλπ., οι οποίες επιτρέπουν την απομάκρυνση από το λεξιλόγιο και άρα από την επίδραση του θέματος του κειμένου.
- Συντακτικά χαρακτηριστικά είναι οι σχέσεις εξάρτησης κάθε λέξης μέσα στη δομή της πρότασης (π.χ. υποκείμενο, αντικείμενο, επιθετικός προσδιορισμός), καθώς και μετρικές όπως η απόσταση εξάρτησης και η κατεύθυνση εξάρτησης. Η υφομετρική ανάλυση, δηλαδή, βασίζεται πάνω σε δομικά στοιχεία του κειμένου.
- Τα σύνθετα χαρακτηριστικά προκύπτουν από συνδυασμούς μορφολογικών και συντακτικών πληροφοριών**.** Ουσιαστικά πρόκειται για «μορφοσυντακτικά n-grams» που αναπαριστούν ακολουθίες λέξεων όχι γραμμικά, αλλά σύμφωνα με την ιεραρχία της εξάρτησης στη δομή της πρότασης.

Συνολικά, ο Gorman υποστηρίζει ότι αυτά τα χαρακτηριστικά είναι topic-agnostic, δηλαδή σε μεγάλο βαθμό ανεξάρτητα από το θεματικό περιεχόμενο, κάτι που τα καθιστά κατάλληλα για μελέτες υφομετρίας. Επιπλέον, αποδεικνύει ότι τα υπάρχοντα συστήματα UDPipe, παρότι αναπόφευκτα κάνουν λάθη, παράγουν επισημειώσεις επαρκώς ακριβείς ώστε τα σφάλματά τους να μην υπονομεύουν ουσιαστικά τα αποτελέσματα των υφομετρικών ταξινομήσεων**.** Τα μορφοσυντακτικά χαρακτηριστικά παραμένουν δηλαδή στιβαρά και αποτελεσματικά παρά τα αναμενόμενα λάθη των αναλυτών.

2.2 Τεχνικές Μηχανικής Μάθησης

Στο δεύτερο σκέλος της υφομετρικής διαδικασίας, τα χαρακτηριστικά τα οποία έχουν επιλεγεί και εξαχθεί από τα κείμενα πρέπει να εξεταστούν ώστε να προκύψει το επιθυμητό αποτέλεσμα, δηλαδή η αποτίμηση ομοιοτήτων και διαφορών ανάμεσα σε κείμενα, η ομαδοποίησή τους, η απόδοση ή επαλήθευση της συγγραφικής ταυτότητας, η διαπίστωση αλλοίωσης ή λογοκλοπής, καθώς και άλλα συμπεράσματα. Σε όλες τις σύγχρονες εφαρμογές τις υφομετρίας, οι προσεγγίσεις της επιστημονικής κοινότητας χρησιμοποιούν συχνότατα εργαλεία μηχανικής μάθησης που επεξεργάζονται τα δεδομένα (δηλαδή τα επιλεγμένα χαρακτηριστικά) και εξάγουν τις ζητούμενες πληροφορίες.

2.2.1 Εποπτευόμενα (supervise) μοντέλα

Οι εποπτευόμενες μέθοδοι (supervised learning) βασίζονται στην ύπαρξη ετικετών συγγραφέα. Εκπαιδεύουν μοντέλα που μαθαίνουν να αναγνωρίζουν υφολογικά μοτίβα με βάση παραδείγματα.

Ο Yavanoglu[9] εφαρμόζει ένα σύστημα συγγραφικής αναγνώρισης βασισμένο σε τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks - ANN), αξιοποιώντας 41 υφομετρικά χαρακτηριστικά τουρκικών δημοσιογραφικών άρθρων, τα οποία ανήκουν κυρίως στις λεξικές και συντακτικές κατηγορίες. Για την εκπαίδευση χρησιμοποιεί πολυεπίπεδα perceptrons (MLP) με διαφορετικές αρχιτεκτονικές, ενώ ο αλγόριθμος Levenberg–Marquardt επιλέγεται ως ο πιο αποτελεσματικός για τη βελτιστοποίηση του μοντέλου. Το σύστημα εκπαιδεύεται πάνω σε ένα μεγάλο πλήθος κειμένων (22.000 άρθρα) που καλύπτει μία γκάμα από διαφορετικά είδη (οικονομία, πολιτική, κοινωνικά), επιτυγχάνοντας πολύ υψηλά ποσοστά αναγνώρισης που φτάνουν έως και το 98%, με κατώτερο όριο 80%, ανάλογα με το είδος. Φαίνεται ότι ακόμη και με σχετικά μικρό μήκος κειμένων (περίπου 900 λέξεις), τα νευρωνικά δίκτυα αποδίδουν ικανοποιητικά σε εφαρμογές συγγραφικής ταυτοποίησης.

Ο Verhoeven και οι συνεργάτες του[10] εφαρμόζουν μία μέθοδο ταξινόμησης για την αναγνώριση φύλου Σλοβένων χρηστών του Twitter, αξιοποιώντας Support Vector Machines (SVM) σε πειράματα 10-fold cross-validation. Οι συγγραφείς συγκρίνουν δύο τρόπους προεπεξεργασίας, την token-based προσέγγιση, όπου το κείμενο διατηρείται στη μορφή των αρχικών λέξεων, και τη lemma-based, όπου εφαρμόζεται lemmatization, δηλαδή αναγωγή των λέξεων στη βασική τους μορφή. Τα χαρακτηριστικά περιλαμβάνουν word 1-grams/2-grams και character 3-grams/4-grams. Η token-based μέθοδος επιτυγχάνει σαφώς υψηλότερη ακρίβεια (92,6% έναντι 87,9%), καθώς η λημματοποίηση αφαιρεί κρίσιμα μορφολογικά γνωρίσματα του φύλου στη σλοβενική. Η διαφορά αυτή είναι ουσιαστική και καταδεικνύει ότι η μορφολογία αποτελεί ισχυρό δείκτη για τον προσδιορισμό του φύλου, στη συγκεκριμένη γλώσσα τουλάχιστον. Ωστόσο η lemma-based προσέγγιση δεν αποδίδει πολύ χαμηλότερα και αναδεικνύει καθαρότερα θεματικές και στιλιστικές τάσεις.

Ο Brocardo και οι συνεργάτες του[11] εφαρμόζουν μια εποπτευόμενη μέθοδο συγγραφικής επαλήθευσης με Deep Belief Networks (Gaussian–Bernoulli DBN), χρησιμοποιώντας ένα μεγάλο σύνολο λεξικών, συντακτικών και character-based χαρακτηριστικών. Η τεχνική βασίζεται στην

ανάλυση μικρών κειμένων (blocks 140–500 χαρακτήρων), ώστε να προσομοιωθεί συνθήκη continuous authentication σε e-mails και tweets (αντί να ελέγχεται ο χρήστης μία φορά, το σύστημα ελέγχει διαρκώς αν αυτός που γράφει είναι πράγματι ο νόμιμος χρήστης). Πριν από την ταξινόμηση, όλα τα κείμενα υφίστανται tokenization και εξαγωγή n-grams, ενώ εφαρμόζονται μέθοδοι επιλογής και συγχώνευσης χαρακτηριστικών. Το μοντέλο επιτυγχάνει πολύ χαμηλά EER (5,48%–16,7%), δείχνοντας ότι μέθοδοι deep learning μπορούν να υπερέχουν έναντι των κλασικών ταξινομητών (όπως SVM) σε εργασίες συγγραφικής επαλήθευσης με μικρά σε μέγεθος κείμενα.

Ο Ríos-Toledo και οι συνεργάτες του[12] ασχολούνται με την ανίχνευση διαχρονικών αλλαγών στο ύφος λογοτεχνών, εκπαιδεύοντας Logistic Regression και SVM σε χαρακτηριστικά που προκύπτουν από tokenization, POS tagging και dependency parsing. Κεντρικό ρόλο έχουν τα POS n-grams, τα οποία αποτυπώνουν επαναλαμβανόμενα μορφοσυντακτικά μοτίβα, ανεξάρτητα από το λεξιλόγιο. Μαζί με word και character n-grams, αξιολογούνται ως προς την ικανότητά τους να διακρίνουν πρώιμες από όψιμες φάσεις γραφής. Σύμφωνα με τα αποτελέσματα, ιδίως τα POS και syntactic n-grams παρέχουν υψηλή ακρίβεια.

Στο [13], οι Caliskan-Islam et al προτείνουν μια ισχυρή μέθοδο απόδοσης συγγραφής σε προγραμματιστές, βασισμένη κυρίως σε χαρακτηριστικά που εξάγονται από Abstract Syntax Trees (ASTs). Οι συγγραφείς δείχνουν ότι τα AST-based συντακτικά features αποτελούν τα πιο διακριτικά στοιχεία. Η μέθοδός τους, βασισμένη σε Random Forest classifiers, επιτυγχάνει πολύ υψηλές ακρίβειες (έως 98% σε 250 συγγραφείς και 93% σε 1600). Επίσης, παρουσιάζουν ότι η μέθοδος παραμένει ανθεκτική σε απλές τεχνικές παραμόρφωσης, ενώ κλιμακώνεται αποτελεσματικά σε μεγάλα σύνολα συγγραφέων.

2.2.2 Μη εποπτευόμενα (unsupervised) μοντέλα

Οι μη εποπτευόμενες μέθοδοι αποκτούν ολοένα και μεγαλύτερη σημασία, ιδίως σε περιπτώσεις όπου δεν υπάρχουν διαθέσιμες ετικέτες συγγραφέων ή όταν η εκ των προτέρων κατάτμηση του υλικού είναι αβέβαιη. Οι προσεγγίσεις αυτές επιδιώκουν να αναδείξουν ενδογενή στιλιστικά πρότυπα αποκλειστικά μέσω της δομής των δεδομένων, χωρίς εξωτερική επιτήρηση, και έχουν αποδειχθεί ιδιαίτερα χρήσιμες σε ιστορικά ή πολυσυγγραφικά κείμενα.

Οι Fifield, Follan και Lunde[14] προτείνουν μία μη εποπτευόμενη μέθοδο συγγραφικής ανάλυσης, σχεδιασμένη για περιπτώσεις όπου δεν υπάρχουν ετικέτες συγγραφέων. Η τεχνική τους βασίζεται στη διάσπαση του κειμένου σε επικαλυπτόμενα παράθυρα σταθερού μεγέθους και στην επανάληψη της συσταδοποίησης πολλές φορές με διαφορετικές μετατοπίσεις, έτσι ώστε κάθε clustering να «βλέπει» ελαφρώς διαφορετικά τμήματα του ίδιου κειμένου. Κάθε τέτοια κατάτμηση ομαδοποιείται με ανεξάρτητο clustering αλγόριθμο, παράγοντας πολλαπλές ταξινομήσεις που συλλαμβάνουν τις τοπικές υφολογικές ομοιότητες και τις μεταβάσεις. Η μέθοδος αυτή, των επαναλαμβανόμενων επικαλυπτόμενων συσταδοποιήσεων, εφαρμόζεται με επιτυχία σε αγγλικά και αρχαιοελληνικά κείμενα.

Στο άρθρο των Layton, Watters και Layton[15] παρουσιάζεται ένα πλαίσιο υφομετρικής ανάλυσης με στόχο τη διάκριση συγγραφέων σε περιβάλλοντα όπου δεν υπάρχουν διαθέσιμες ετικέτες. Η μελέτη βασίζεται αποκλειστικά σε στατιστικές ιδιότητες του κειμένου (character/word n-grams) και αξιοποιεί τεχνικές ομαδοποίησης για την ανάδυση στιλιστικών προτύπων. Χρησιμοποιούν μία EAC (Evidence Acumulation Clustering) μέθοδο, που ξεκινά με πολλαπλούς ανεξάρτητους υπολογισμούς clustering

(κυρίως μέσω k-means) πάνω στα ίδια δεδομένα, κάθε φορά με διαφορετικές αρχικοποιήσεις, ώστε να παραχθεί μια ποικιλία ταξινομήσεων. Οι ομαδοποιήσεις συνδυάζονται σε έναν co-association matrix, ο οποίος αποτυπώνει τη συχνότητα με την οποία δύο τμήματα κειμένου καταλήγουν στο ίδιο cluster και αξιοποιείται ως νέα απόσταση για τελική ιεραρχική ομαδοποίηση, η οποία παράγει πιο σταθερές και ανθεκτικές ομάδες. Η επιλογή του αριθμού των clusters γίνεται με το κριτήριο IPS, το οποίο εντοπίζει το σημείο στο οποίο περαιτέρω διάσπαση δεν βελτιώνει τη συνοχή των ομάδων.

Στο [16], οι Martín-del-Campo-Rodríguez et al. κατάφεραν σε αρκετές περιπτώσεις να πετύχουν καλύτερες αποδόσεις στο πρόβλημα της μη εποπτευόμενης απόδοσης συγγραφέα (author clustering) από το καθιερωμένο benchmark του διαγωνισμού PAN2017. Η μέθοδός τους συνδυάζει επιλογή χαρακτηριστικών με βελτιωμένη αντιμετώπιση των αποστάσεων ανάμεσα σε κείμενα, χρησιμοποιώντας διάφορες τεχνικές feature selection, με κορυφαία την MAD-based (Median Absolute Deviation· υψηλό MAD υποδεικνύει ότι το χαρακτηριστικό παρουσιάζει ουσιαστική διακύμανση ανάμεσα στα κείμενα και, ως εκ τούτου, αξίζει να διατηρηθεί). Η σταθμισμένη απόσταση, πάνω στην οποία εφαρμόζεται η ιεραρχική συσταδοποίηση, υπολογίζεται μέσω παραλλαγής της cosine similarity (weighted cosine), ενός μέτρου ομοιότητας που μειώνει την επίδραση σπάνιων και μη αντιπροσωπευτικών χαρακτηριστικών.

2.2.3 Embeddings

Στο πεδίο της σύγχρονης υφομετρίας, τα embeddings, που παράγονται από μοντέλα μετασχηματιστών (transformers) προσφέρουν μια πιο βαθιά αναπαράσταση της γλώσσας, ευαίσθητη στα συμφραζόμενα, υπερβαίνοντας τα παραδοσιακά χαρακτηριστικά βασισμένα σε συχνότητες. Η χρήση τέτοιων μοντέλων BERT(Bidirectional Encoder Representations from Transformers) επιτρέπει τη σύλληψη λεπτών υφολογικών διαφορών και έχει καταστήσει τα embeddings μια ανερχόμενη κατηγορία εργαλείων στην υφομετρική ανάλυση.

Ένα BERT μοντέλο δέχεται ως είσοδο ένα κείμενο και το μετατρέπει σε μια σειρά από embeddings, δηλαδή σε αριθμητικά διανύσματα υψηλής διάστασης που αναπαριστούν τις λέξεις, τις φράσεις ή και ολόκληρες προτάσεις με τρόπο κατανοητό από υπολογιστικά μοντέλα. Αυτά τα διανύσματα αποτυπώνουν όχι μόνο τη λεξική μορφή, αλλά και τις συντακτικές και σημασιολογικές σχέσεις που αναδύονται στο εσωτερικό του κειμένου. Μέσω μηχανισμών self-attention, το BERT αξιολογεί τη συμβολή κάθε λέξης σε σχέση με όλες τις άλλες, επιτρέποντας σε κάθε embedding να «ενσωματώνει» πληροφορία σχετικά με τα συμφραζόμενα. Η διαφοροποίηση από τα παραδοσιακά στατικά embeddings (όπως Word2Vec ή GloVe) είναι καίρια, καθώς στα μοντέλα τύπου BERT, η ίδια λέξη αποκτά διαφορετική αριθμητική αναπαράσταση ανάλογα με το περιβάλλον στο οποίο εμφανίζεται.

Οι Zamir et al.[17] αναφέρονται συνοπτικά στις ιδιότητες διάφορων μοντέλων BERT (BERT, RoBERTa, DistilBERT, ALBERT και XLM-RoBERTa), τα οποία αξιοποιούν ως εξαγωγείς χαρακτηριστικών σε πολυσυγγραφικά κείμενα εφαρμόζοντας μία εποπτευόμενη προσέγγιση σε τρία επί μέρους προβλήματα, συγκεκριμένα τη διάκριση μονοσυγγραφικών και πολυσυγγραφικών κειμένων, τον εντοπισμό σημείων αλλαγής συγγραφέα, και την πλήρη χαρτογράφηση όλων των υφολογικών μεταβολών μέσα στο κείμενο. Κάθε transformer encoder, όπως υλοποιείται στα μοντέλα που προαναφέρθηκαν, παράγει τα δικά του embeddings για κάθε τμήμα κειμένου. Τα embeddings αυτά τροφοδοτούνται σε έναν ανεξάρτητο εποπτευόμενο ταξινομητή για κάθε μοντέλο (fully connected / feed-forward layer), ο οποίος εκπαιδεύεται να προβλέπει αν δύο παράγραφοι ανήκουν στον ίδιο συγγραφέα ή αν υπάρχει αλλαγή ύφους. Μετά την ολοκλήρωση της εκπαίδευσης των επί μέρους μοντέλων εφαρμόζεται η διαδικασία late fusion, κατά την οποία οι τελικές πιθανότητες που παράγει ο

κάθε ταξινομητής συγχωνεύονται. Οι προβλέψεις των πέντε μοντέλων συνδυάζονται σε ένα ενιαίο αποτέλεσμα μέσω σταθμισμένου συνδυασμού, όπου τα βάρη των μοντέλων καθορίζονται αυτόματα με αλγορίθμους βελτιστοποίησης (PSO, Nelder–Mead, Powell).

Οι De Langhe et al. [18] εισάγουν μία πρωτοποριακή προσέγγιση για τη μη εποπτευόμενη απόδοση συγγραφέα σε κείμενα μεσαιωνικών λατινικών, βασισμένη σε contextual embeddings από μοντέλα μετασχηματιστών. Για κάθε κείμενο εξάγονται αναπαραστάσεις από τα τέσσερα τελευταία encoder layers πολλών transformer-based μοντέλων (Latin RoBERTa, mBERT, mDeBERTaV3, Longformer), οι οποίες στη συνέχεια ομαδοποιούνται με agglomerative clustering και χαρτογραφούνται μέσω self-organizing maps (SOM). Από τις μετρικές αξιολόγησης των αποτελεσμάτων φαίνεται ότι ο multilingual Longformer επιτυγχάνει τις πιο συνεκτικές συστάδες, ιδίως λόγω της δυνατότητας επεξεργασίας μεγαλύτερων κειμενικών ακολουθιών (4096 tokens).

Κεφάλαιο 3ο: Μεθοδολογία

3.1 Στόχος και μεθοδολογικό πλαίσιο

Έχουμε δείξει ότι το ιπποκρατικό corpus αποτελεί ένα χαρακτηριστικό παράδειγμα κειμενικής ετερογένειας στην αρχαία ελληνική γραμματεία. Όπως έχει επισημάνει και ο Lesky[19], οι ιπποκρατικές πραγματείες δεν συγκροτούν έργο ενός ενιαίου συγγραφέα, αλλά προϊόν διαφορετικών χεριών, εποχών και επιστημονικών συμφραζομένων. Παράλληλα όμως, αναγνωρίζει την ύπαρξη μίας κοινής ιατρικής γραφικής παράδοσης, η οποία χαρακτηρίζεται από λιτότητα, λειτουργικότητα και έμφαση στην εμπειρική παρατήρηση. Αυτό συγκλίνει με την άποψη του Pigeaud[20], σύμφωνα με την οποία η ιπποκρατική γραφή δεν είναι απλώς μέσο καταγραφής γνώσεων, αλλά συνειδητή πρακτική συγκρότησης της ιατρικής τέχνης. Εντοπίζεται δηλαδή ένα ύφος προσανατολισμένο στη σαφήνεια, τη διδακτικότητα και την πρακτική εφαρμογή. Η γλώσσα των ιπποκρατικών κειμένων, δηλαδή, παρά την εσωτερική της ποικιλία, ενσωματώνει σταθερές υφολογικές επιλογές.

Στο πλαίσιο αυτό, ο στόχος της παρούσας έρευνας δεν είναι η απόδοση συγκεκριμένων πραγματειών σε μεμονωμένους συγγραφείς, αλλά η διερεύνηση υφολογικών συγγενειών και διαφοροποιήσεων εντός του corpus. Ειδικότερα, εξετάζεται κατά πόσον τα κείμενα παρουσιάζουν μετρήσιμα υφολογικά μοτίβα, ικανά να οδηγήσουν στη συγκρότηση ομάδων με αυξημένη εσωτερική ομοιογένεια, ομάδων που ενδεχομένως αντανακλούν κοινές συγγραφικές πρακτικές, παραδόσεις ή λειτουργικά είδη λόγου.

Η μεθοδολογία καθορίζεται σε μεγάλο βαθμό από τη φύση του διαθέσιμου υλικού. Με το corpus στην ολότητά του να αποτελεί ήδη ένα πολύ μικρό σώμα κειμένων, το υλικό που συγκεντρώθηκε μέσω του Perseus API προέρχεται από 18 πραγματείες, με τη μεγαλύτερο κείμενο να ανέρχεται περίπου στις 1200 λέξεις, ενώ το μικρότερο στις 5. Επομένως, το υλικό είναι εξαιρετικά περιορισμένο. Όπως δείχνουν οι Eder, et al.[21], σε μορφολογικά πλούσιες γλώσσες όπως τα αρχαία ελληνικά, και ιδίως σε περιορισμένα σύνολα δεδομένων, οι συχνότερες λέξεις(Most Frequent Words – MFW) παραμένουν από τους πιο σταθερούς δείκτες συγγραφικού ύφους, υπερέχοντας συχνά έναντι πιο σύνθετων αναπαραστάσεων, ενώ deep learning και BERT απαιτούν μεγάλες και ισορροπημένες ποσότητες δεδομένων, διαφορετικά τείνουν να συγχέουν τις σημασιολογικές και θεματικές διαφορές με τις υφολογικές. Για τον λόγο αυτό, η παρούσα μελέτη υιοθετεί χαρακτηριστικά χαμηλού επιπέδου, όπως character n-grams και λειτουργικές λέξεις που εξάγονται με data-driven τρόπο από το ίδιο το corpus, σύμφωνα με τη λογική των MFW, ώστε να διασφαλιστεί όσο το δυνατόν η στατιστική τους σταθερότητα και η καταλληλότητά τους για μικρά δείγματα.

Όσον αφορά τη μεθοδολογική προσέγγιση που θα ακολουθηθεί για την αλγοριθμική ανάλυση των δεδομένων, η απουσία αξιόπιστων και γενικώς αποδεκτών ετικετών συγγραφέα καθιστά μη εφαρμόσιμη τη χρήση εποπτευόμενων μοντέλων μηχανικής μάθησης. Ως εκ τούτου, η παρούσα μελέτη υιοθετεί μία μη εποπτευόμενη προσέγγιση, στο πλαίσιο της οποίας τα κείμενα δεν ταξινομούνται βάσει προκαθορισμένων κατηγοριών, αλλά ομαδοποιούνται σε συστάδες που προκύπτουν αποκλειστικά από τα υφολογικά χαρακτηριστικά τους. Οι συστάδες αυτές ερμηνεύονται ως ενδείξεις υφολογικής συγγένειας μεταξύ των κειμένων, σύμφωνα με τις επιλεγμένες αναπαραστάσεις (char n-grams και MFW). Στην ακόλουθη υποενότητα παρουσιάζεται μία αναλυτική επισκόπηση των βασικών κατηγοριών αλγορίθμων συσταδοποίησης, καθώς και η αιτιολόγηση της τελικής επιλογής των μεθόδων που εφαρμόζονται στην παρούσα ανάλυση.

3.1.1 Αλγοριθμικό υπόβαθρο

3.1.1.1 Μη εποπτευόμενη συσταδοποίηση: βασικές αρχές

Η συσταδοποίηση (clustering) αποτελεί κεντρική τεχνική της μη εποπτευόμενης μηχανικής μάθησης και αποσκοπεί στην οργάνωση ενός συνόλου δεδομένων σε ομάδες (συστάδες), έτσι ώστε τα στοιχεία εντός της ίδιας συστάδας να εμφανίζουν μεγαλύτερη μεταξύ τους ομοιότητα σε σχέση με στοιχεία που ανήκουν σε διαφορετικές συστάδες. Σε αντίθεση με τις εποπτευόμενες μεθόδους, δεν υπάρχει προκαθορισμένη έννοια «σωστής» ή «λανθασμένης» ταξινόμησης, και η αξιολόγηση των αποτελεσμάτων βασίζεται σε έμμεσα κριτήρια, όπως η συνοχή, η σταθερότητα και η ερμηνευσιμότητα των συστάδων [22].

Στη γλωσσομετρική και υφομετρική ανάλυση, η συσταδοποίηση χρησιμοποιείται κυρίως για τη διερεύνηση της εσωτερικής δομής ενός corpus, την ανίχνευση ομάδων κειμένων με κοινά υφολογικά χαρακτηριστικά και την ανάδειξη πιθανών συγγραφικών ή σχολικών συγγενειών χωρίς την επιβολή προκαταρκτικών υποθέσεων [23].

Καθοριστικό ρόλο στη συσταδοποίηση διαδραματίζει η επιλογή του μέτρου ομοιότητας ή απόστασης, καθώς αυτό καθορίζει τον τρόπο με τον οποίο συγκρίνονται οι διανυσματικές αναπαραστάσεις των κειμένων. Στα κειμενικά δεδομένα, όπου οι αναπαραστάσεις είναι συχνά υψηλής διαστασιμότητας και έντονα αραιές, η επιλογή κατάλληλου μέτρου είναι κρίσιμη τόσο για τη σταθερότητα όσο και για την ερμηνευσιμότητα των αποτελεσμάτων [22], [24].

3.1.1.2 Μέτρα ομοιότητας και απόστασης

Τα μέτρα ομοιότητας και απόστασης αποτελούν το θεμέλιο κάθε αλγορίθμου συσταδοποίησης. Στη διανυσματική αναπαράσταση κειμένων, έχουν προταθεί και χρησιμοποιηθεί ποικίλα μέτρα, καθένα με διαφορετικές μαθηματικές ιδιότητες και συμπεριφορά σε υψηλές διαστάσεις.

- Ευκλείδεια απόσταση (L2):
  Βασίζεται στο μήκος της ευθείας που συνδέει δύο διανύσματα στον χώρο. Αν και διαισθητικά απλή, η ευκλείδεια απόσταση παρουσιάζει σοβαρά προβλήματα σε υψηλές διαστάσεις, όπου οι αποστάσεις τείνουν να εξισώνονται (φαινόμενο της «συγκέντρωσης αποστάσεων») [25].
- Manhattan απόσταση (L1):
  Ορίζεται ως το άθροισμα των απόλυτων διαφορών των συνιστωσών. Είναι πιο ανθεκτική σε ακραίες τιμές, αλλά λιγότερο διαδεδομένη σε υφομετρικές εφαρμογές.
- Cosine ομοιότητα:
  Μετρά τη γωνία μεταξύ δύο διανυσμάτων και αγνοεί το μέτρο τους. Είναι ιδιαίτερα κατάλληλη για κειμενικά δεδομένα, καθώς εστιάζει στη σχετική κατανομή των χαρακτηριστικών και όχι στο συνολικό μήκος του κειμένου. Διακρίνονται παραλλαγές όπως:
  - απλή cosine similarity,
  - cosine distance (1 − cosine similarity),

- cosine similarity μετά από L2 κανονικοποίηση.

Η cosine ομοιότητα έχει επικρατήσει στη σύγχρονη υφομετρία και στην ανάλυση κειμένων γενικότερα, καθώς προσαρμόζεται καλύτερα στη φύση των αραιών, υψηλής διαστασιμότητας δεδομένων [22], [26].

3.1.1.3 Διαμεριστικοί αλγόριθμοι συσταδοποίησης

Οι διαμεριστικοί (partitioning) αλγόριθμοι χωρίζουν το σύνολο των δεδομένων σε προκαθορισμένο αριθμό συστάδων k. Ο πιο διαδεδομένος εκπρόσωπος αυτής της κατηγορίας είναι ο αλγόριθμος k-means [27].

Ο k-means λειτουργεί επαναληπτικά: αρχικά επιλέγονται k κέντρα (centroids), κάθε σημείο ανατίθεται στο πλησιέστερο κέντρο βάσει κάποιου μέτρου απόστασης και στη συνέχεια τα κέντρα επανυπολογίζονται ως ο μέσος όρος των σημείων της συστάδας. Η διαδικασία επαναλαμβάνεται έως τη σύγκλιση.

Στην πράξη, ο k-means είναι υπολογιστικά αποδοτικός και κλιμακώνεται καλά σε υψηλές διαστάσεις, γεγονός που τον καθιστά κατάλληλο για γλωσσομετρικές αναπαραστάσεις όπως τα character n-grams. Ωστόσο, παρουσιάζει γνωστούς περιορισμούς: απαιτεί προκαθορισμένο αριθμό συστάδων, είναι ευαίσθητος στις αρχικές συνθήκες και μπορεί να παράγει διαφορετικά αποτελέσματα σε διαφορετικές εκτελέσεις [22], [27].

Πέραν του k-means, μια συγγενής αλλά πιο ανθεκτική μέθοδος είναι ο k-medoids. Σε αντίθεση με τον k-means, όπου κάθε συστάδα εκπροσωπείται από το μέσο όρο των σημείων της, ο k-medoids επιλέγει ως κέντρο (medoid) ένα πραγματικό σημείο των δεδομένων. Αυτό καθιστά τον αλγόριθμο λιγότερο ευαίσθητο σε ακραίες τιμές και θόρυβο.

Ωστόσο, ο k-medoids παρουσιάζει:

- αυξημένη υπολογιστική πολυπλοκότητα, καθώς απαιτεί υπολογισμό αποστάσεων μεταξύ όλων των ζευγών σημείων,
- μικρότερη κλιμακωσιμότητα σε υψηλές διαστάσεις,
- και συχνά χαμηλότερη απόδοση σε μεγάλα, αραιά διανύσματα όπως αυτά που προκύπτουν από n-grams[28].

Παράλληλα, μέθοδοι όπως τα Gaussian Mixture Models (GMMs) εισάγουν ισχυρότερες στατιστικές παραδοχές, όπως η κανονικότητα των κατανομών εντός κάθε συστάδας. Τέτοιες παραδοχές είναι δύσκολο να τεκμηριωθούν για υφομετρικά δεδομένα, ιδιαίτερα σε μικρά corpora [29].

Για τους λόγους αυτούς, ο k-means παραμένει πρακτική επιλογή όταν συνδυάζεται με ensemble τεχνικές που μετριάζουν την αστάθειά του.

3.1.1.4 Ιεραρχικοί αλγόριθμοι συσταδοποίησης

Οι ιεραρχικοί αλγόριθμοι δεν απαιτούν προκαθορισμένο αριθμό συστάδων και παράγουν δενδροειδή αναπαράσταση (dendrogram). Στη συγχωνευτική (agglomerative) εκδοχή, κάθε στοιχείο ξεκινά ως αυτόνομη συστάδα και σε κάθε βήμα συγχωνεύονται οι δύο πλησιέστερες συστάδες [22].

Υπάρχει επίσης η διαιρετική (divisive) προσέγγιση, όπου όλα τα δεδομένα ξεκινούν σε μία ενιαία συστάδα και διασπώνται επαναληπτικά. Παρότι θεωρητικά ελκυστική, η διαιρετική συσταδοποίηση είναι υπολογιστικά πιο απαιτητική και εφαρμόζεται σπανιότερα στην πράξη [22].

Η έννοια της «εγγύτητας» μεταξύ συστάδων καθορίζεται από το κριτήριο σύνδεσης (linkage). Πέραν του average linkage, που χρησιμοποιείται στη μελέτη αυτή, διακρίνονται:

- Single linkage: ελάχιστη απόσταση στοιχείων, τείνει να δημιουργεί «αλυσίδες».
- Complete linkage: μέγιστη απόσταση στοιχείων, παράγει πιο συμπαγείς συστάδες αλλά είναι ευαίσθητο σε ακραίες τιμές.
- Ward's method: ελαχιστοποιεί την ενδοσυσταδιακή διακύμανση, αλλά προϋποθέτει ευκλείδεια απόσταση [22], [30].

Η επιλογή του average linkage αποτελεί συμβιβασμό μεταξύ αυτών των άκρων.

Παρά τα πλεονεκτήματά τους, οι ιεραρχικοί αλγόριθμοι είναι ιδιαίτερα ευαίσθητοι στον θόρυβο και δεν επιτρέπουν αναθεώρηση προηγούμενων συγχωνεύσεων, γεγονός που μπορεί να οδηγήσει σε ασταθή αποτελέσματα όταν εφαρμόζονται απευθείας σε υψηλής διαστασιμότητας υφομετρικά δεδομένα.

3.1.1.5 Μέθοδοι βασισμένες στην πυκνότητα και στο γράφημα

Άλλες κατηγορίες μη εποπτευόμενων αλγορίθμων περιλαμβάνουν μεθόδους βασισμένες στην πυκνότητα, όπως ο DBSCAN, καθώς και μεθόδους βασισμένες σε γράφημα ή φασματική ανάλυση [31].

Ωστόσο, η εφαρμογή τους σε υφομετρικά δεδομένα παρουσιάζει σοβαρές δυσκολίες. Η έννοια της «πυκνότητας» σε υψηλές διαστάσεις είναι συχνά ασαφής, καθώς σε πολυδιάστατους χώρους όλα τα σημεία τείνουν να απέχουν περίπου το ίδιο μεταξύ τους, καθιστώντας δυσχερή τον ορισμό πυκνών περιοχών [25].

3.1.1.6 Ensemble και consensus clustering

Για την αντιμετώπιση της αστάθειας των μεμονωμένων εκτελέσεων συσταδοποίησης, έχει προταθεί η χρήση ensemble ή consensus μεθόδων. Στο πλαίσιο αυτό, πολλαπλές εκτελέσεις ενός αλγορίθμου συσταδοποίησης, με διαφορετικές παραμετροποιήσεις ή αρχικοποιήσεις, συνδυάζονται ώστε να εξαχθεί μια πιο σταθερή και αξιόπιστη ομαδοποίηση.

Η Evidence Accumulation Clustering (EAC) αποτελεί χαρακτηριστικό παράδειγμα τέτοιας προσέγγισης. Αντί να βασίζεται σε μία μόνο κατάτμηση των δεδομένων, η μέθοδος καταγράφει πόσες φορές δύο κείμενα ανατίθενται στην ίδια συστάδα σε διαφορετικές εκτελέσεις. Η πληροφορία αυτή αποτυπώνεται σε έναν πίνακα συνσυσχέτισης, ο οποίος εκφράζει τη συχνότητα κοινής ανάθεσης και λειτουργεί ως νέο, πιο σταθερό μέτρο εγγύτητας.

3.1.1.7 Επίδραση της αναπαράστασης των δεδομένων στην επιλογή αλγορίθμου

Η επιλογή αλγορίθμου συσταδοποίησης επηρεάζεται και από τη φύση των χαρακτηριστικών που χρησιμοποιούνται. Στην παρούσα μελέτη, τα character n-grams και οι MFW παράγουν διανυσματικές αναπαραστάσεις υψηλής διαστασιμότητας και έντονης αραιότητας.

Ως υψηλή διαστασιμότητα νοείται η αναπαράσταση των δεδομένων σε χώρο με πολύ μεγάλο αριθμό διαστάσεων, όπως συμβαίνει όταν κάθε n-gram ή λέξη αντιστοιχεί σε ξεχωριστή διάσταση. Στην παρούσα μελέτη, τα διανύσματα χαρακτηριστικών περιλαμβάνουν χιλιάδες διαστάσεις, ενώ η έντονη αραιότητα αναφέρεται στο γεγονός ότι κάθε κείμενο ενεργοποιεί μόνο ένα μικρό ποσοστό αυτών των διαστάσεων[22],[25], καθώς τα περισσότερα χαρακτηριστικά δεν εμφανίζονται σε κάθε κείμενο. Ο συνδυασμός αυτών των δύο ιδιοτήτων καθιστά ακατάλληλα πολλά κλασικά μέτρα απόστασης και ενισχύει την ανάγκη για κανονικοποίηση και γωνιακά μέτρα ομοιότητας.

Η κανονικοποίηση των διανυσμάτων επιτρέπει τη χρήση του k-means με τρόπο λειτουργικά ισοδύναμο προς τη χρήση cosine distance[22], ενώ η επαναληπτική εκτέλεση του αλγορίθμου με διαφορετικές παραμέτρους μειώνει την εξάρτηση από τυχαίες αρχικές συνθήκες. Παράλληλα, η χρήση ιεραρχικής συσταδοποίησης πάνω στον πίνακα συνσυσχέτισης της EAC επιτρέπει την ανάδειξη πολυεπίπεδων σχέσεων χωρίς να επιβάλλεται αυστηρός αριθμός συστάδων.

3.1.1.8 Σύνοψη και αιτιολόγηση τελικής επιλογής

Συνοψίζοντας, η επιλογή της μεθοδολογίας που υιοθετείται στη μελέτη αυτή βασίζεται στη συνδυαστική χρήση πολλαπλών εκτελέσεων του k-means και ιεραρχικής συσταδοποίησης μέσω average linkage πάνω στον πίνακα συνσυσχέτισης της EAC. Η προσέγγιση αυτή εξισορροπεί την υπολογιστική αποδοτικότητα των διαμεριστικών αλγορίθμων με τη σταθερότητα και την ερμηνευσιμότητα των ιεραρχικών μεθόδων, και είναι ιδιαίτερα κατάλληλη για μικρά, φιλολογικά ευαίσθητα σώματα κειμένων όπως το ιπποκρατικό.

3.2 Pipeline ανάλυσης

Η παρούσα ενότητα περιγράφει αναλυτικά το υπολογιστικό pipeline που ακολουθήθηκε, από την ανάκτηση και προεπεξεργασία των δεδομένων έως την εξαγωγή χαρακτηριστικών, τη συσταδοποίηση και την οπτικοποίηση των αποτελεσμάτων. Το pipeline εφαρμόστηκε με συνεπή τρόπο τόσο στα character n-grams όσο και στα function words, επιτρέποντας τη συγκριτική αξιολόγηση διαφορετικών αναπαραστάσεων ύφους.

3.2.1 Ανάκτηση και προεπεξεργασία δεδομένων

Για την ανάκτηση των κειμένων χρησιμοποιήθηκε το Perseus Digital Library API, και συγκεκριμένα η CTS (Canonical Text Services) διεπαφή, η οποία αποτελεί το επίσημο πρότυπο του Perseus για την αναγνώριση και ανάκτηση αρχαίων κειμένων βάσει σταθερών αναγνωριστικών (URNs).

Μετά την ανάκτηση, τα κείμενα υπέστησαν ελαφριά αλλά συστηματική προεπεξεργασία, με στόχο τη μείωση της ορθογραφικής και επιφανειακής ποικιλίας χωρίς αλλοίωση του υφολογικού σήματος.

Συγκεκριμένα εφαρμόστηκαν τα εξής βήματα:

- Αφαίρεση σημείων στίξης: αφαιρέθηκαν σημεία στίξης και μη αλφαβητικοί χαρακτήρες, ώστε τα χαρακτηριστικά να βασίζονται αποκλειστικά στη γλωσσική ύλη και όχι σε εκδοτικές συμβάσεις.

- Πεζοποίηση (lowercasing): όλοι οι χαρακτήρες μετατράπηκαν σε πεζούς, ώστε να ενοποιηθούν μορφές της ίδιας λέξης που διαφέρουν μόνο ως προς την κεφαλαιοποίηση. Το βήμα αυτό μειώνει τεχνητές διαφοροποιήσεις που δεν σχετίζονται με ύφος.
- Κανονικοποίηση χαρακτήρων (normalization): εφαρμόστηκε κανονικοποίηση Unicode (π.χ. NFD/NFC), με σκοπό την ενοποίηση εναλλακτικών γραφών του ίδιου χαρακτήρα (ιδίως σε πολυτονικά ελληνικά). Έτσι διασφαλίζεται ότι ο ίδιος χαρακτήρας δεν καταγράφεται ως διαφορετικό feature.
- Διατήρηση διακριτικών (τόνων/πνευμάτων): τα διακριτικά δεν αφαιρέθηκαν, προκειμένου να διατηρηθεί το πρωτογενές υφολογικό σήμα των κειμένων, δηλαδή τα χαρακτηριστικά που απορρέουν άμεσα από τη γλωσσική πρακτική και τον τρόπο γραφής, χωρίς την επιβολή εξωτερικών γλωσσολογικών μοντέλων.
- Αποθήκευση καθαρισμένων κειμένων: τα καθαρισμένα κείμενα αποθηκεύτηκαν σε ενιαία μορφή UTF-8 και χρησιμοποιήθηκαν ως κοινή βάση για όλα τα επόμενα στάδια εξαγωγής χαρακτηριστικών.

3.2.2 Εξαγωγή χαρακτηριστικών και αναπαράσταση κειμένων

Για την αναπαράσταση των κειμένων σε μορφή αριθμητικών διανυσμάτων εφαρμόστηκαν δύο κατηγορίες χαρακτηριστικών: character n-grams και λειτουργικές λέξεις (Most Frequent Words – MFW). Οι δύο αυτές προσεγγίσεις διαφοροποιούνται τόσο ως προς το επίπεδο γλωσσικής ανάλυσης όσο και ως προς το σχήμα στάθμισης που χρησιμοποιήθηκε.

- Στην περίπτωση των character n-grams, τα κείμενα αναλύθηκαν σε ακολουθίες χαρακτήρων μήκους από 3 έως 5 (n = 3…5). Η επιλογή χαρακτήρων αντί λέξεων επιτρέπει την καταγραφή ορθογραφικών, μορφολογικών και υπολεξικών μοτίβων, τα οποία έχουν αποδειχθεί ιδιαίτερα ανθεκτικά σε θεματικές διαφοροποιήσεις και κατάλληλα για υφολογική ανάλυση, ιδίως σε μικρά σύνολα δεδομένων. Οι ακολουθίες χαρακτήρων σταθμίστηκαν με το σχήμα TF-IDF (Term Frequency – Inverse Document Frequency), το οποίο ορίζεται ως:

$$\text{TF-IDF}(t, d) = \text{tf}(t, d) \cdot \log \frac{N}{\text{df}(t)}$$

όπου:

$\text{tf}(t, d)$είναι η συχνότητα του n-gram tστο έγγραφο d,

$\text{df}(t)$είναι ο αριθμός των εγγράφων στα οποία εμφανίζεται το t,

Nείναι το συνολικό πλήθος των εγγράφων του corpus.

Το TF-IDF ενισχύει n-grams που είναι χαρακτηριστικά για συγκεκριμένα κείμενα, ενώ αποδυναμώνει εξαιρετικά συχνές ακολουθίες χαρακτήρων που εμφανίζονται σχεδόν καθολικά

στο corpus. Η κανονικοποίηση των διανυσμάτων έγινε με L2 normalization, ώστε να καταστούν συγκρίσιμα τα κείμενα ανεξαρτήτως μήκους.

- Για τα χαρακτηριστικά λειτουργικών λέξεων, κατασκευάστηκαν πίνακες σχετικών συχνοτήτων ανά κείμενο. Εξετάστηκαν δύο εκδοχές. Η αρχική λίστα των Most Frequent Words (MFW), όπως προέκυψε από το σύνολο του corpus, περιελάμβανε συνολικά 116 λέξεις, οι οποίες εμφανίζονται με υψηλή συχνότητα σε όλα τα κείμενα.Σε δεύτερο στάδιο εφαρμόστηκε φιλτράρισμα με σκοπό την αφαίρεση λέξεων που λειτουργούν πιθανότατα ως κύρια ονόματα ή θεματικοί δείκτες (π.χ. ονόματα προσώπων, τόπων ή τεχνικών όρων με άνιση κατανομή στα κείμενα). Μετά την εφαρμογή αυτού του φιλτραρίσματος, η τελική λίστα περιελάμβανε 102 λέξεις, δηλαδή 14 λιγότερες από την αρχική. Η σύγκριση των δύο αυτών συνόλων (with / without) επιτρέπει την αξιολόγηση της επίδρασης των θεματικά φορτισμένων λέξεων στη συσταδοποίηση και στον εντοπισμό υφολογικών ομοιοτήτων.

Οι σχετικές συχνότητες υπολογίστηκαν ως:

$$f(w,d) = \frac{\text{count}(w,d)}{\sum_{w'} \text{count}(w',d)}$$

όπου $\text{count}(w,d)$είναι η απόλυτη συχνότητα της λέξης w στο έγγραφο d.

Η μικρή ποσοτικά διαφορά μεταξύ των δύο λιστών υποδηλώνει ότι το μεγαλύτερο μέρος του λεξιλογίου υψηλής συχνότητας παραμένει σταθερό, γεγονός που ενισχύει την υπόθεση ότι τα MFW λειτουργούν ως υφολογικοί και όχι θεματικοί δείκτες, ακόμη και σε ένα ετερογενές corpus όπως το ιπποκρατικό.

Και στις δύο περιπτώσεις, η εξαγωγή χαρακτηριστικών πραγματοποιήθηκε με τρόπο data-driven, χωρίς την επιβολή εξωτερικών γλωσσολογικών μοντέλων, ώστε να διατηρηθεί το πρωτογενές υφολογικό σήμα των κειμένων.

3.2.3 Evidence Accumulation Clustering (EAC)

Η Evidence Accumulation Clustering (EAC) αποτελεί μια μεθοδολογία συσταδοποίησης τύπου ensemble, η οποία δεν βασίζεται σε ένα και μοναδικό αποτέλεσμα συσταδοποίησης, αλλά συνθέτει την πληροφορία που προκύπτει από πολλαπλές ανεξάρτητες εκτελέσεις ενός μη ιεραρχικού αλγορίθμου συσταδοποίησης, στην προκειμένη περίπτωση του k-means. Η βασική ιδέα είναι ότι, αν δύο κείμενα τείνουν να καταλήγουν μαζί στην ίδια συστάδα υπό διαφορετικές παραμέτρους και τυχαίες αρχικοποιήσεις, τότε αυτή η συν-εμφάνιση αποτελεί ισχυρή ένδειξη ουσιαστικής ομοιότητας.

Αφετηρία της διαδικασίας είναι ένας πίνακας χαρακτηριστικών, στον οποίο κάθε γραμμή αντιστοιχεί σε ένα κείμενο και κάθε στήλη σε ένα χαρακτηριστικό. Οι αναπαραστάσεις αυτές διαφέρουν ανάλογα με το είδος των χαρακτηριστικών (π.χ. character n-grams ή λειτουργικές λέξεις), αλλά σε κάθε περίπτωση οδηγούν σε έναν διανυσματικό χώρο στον οποίο μπορεί να εφαρμοστεί συσταδοποίηση.

Ο k-means είναι ένας αλγόριθμος μη ιεραρχικής (flat) συσταδοποίησης. Δεδομένου ενός προκαθορισμένου αριθμού συστάδων k, ο αλγόριθμος επιδιώκει να διαχωρίσει τα δεδομένα σε kομάδες, έτσι ώστε κάθε παρατήρηση να ανατεθεί στο κοντινότερο κέντρο συστάδας. Η διαδικασία βασίζεται σε επαναληπτική βελτιστοποίηση: αρχικά επιλέγονται κέντρα, στη συνέχεια τα δεδομένα ανατίθενται στα πλησιέστερα κέντρα, και τα κέντρα επαναϋπολογίζονται ως ο μέσος όρος των σημείων που τους έχουν ανατεθεί. Η διαδικασία επαναλαμβάνεται μέχρι σύγκλισης.

Στο πλαίσιο του EAC, ο k-means δεν χρησιμοποιείται για να παραχθεί ένα τελικό αποτέλεσμα συσταδοποίησης, αλλά ως γεννήτρια πολλών εναλλακτικών διασπάσεων των δεδομένων. Για τον λόγο αυτό, ο αριθμός συστάδων kδεν είναι σταθερός, αλλά επιλέγεται κάθε φορά από ένα προκαθορισμένο εύρος τιμών, ενώ παράλληλα μεταβάλλεται και η τυχαία αρχικοποίηση των κέντρων. Κάθε εκτέλεση παράγει μια ανεξάρτητη ομαδοποίηση των ίδιων κειμένων.

Το κρίσιμο βήμα του EAC είναι η μετάβαση από τις επιμέρους ομαδοποιήσεις σε μια συνολική εκτίμηση ομοιότητας μεταξύ των κειμένων. Για τον σκοπό αυτό κατασκευάζεται ένας πίνακας συν-συσχέτισης, στον οποίο κάθε στοιχείο εκφράζει το ποσοστό των εκτελέσεων του k-means στις οποίες δύο κείμενα κατέληξαν στην ίδια συστάδα. Τα διανύσματα χαρακτηριστικών κανονικοποιούνται με βάση την L2 νόρμα, δηλαδή διαιρούνται με το τετραγωνικό ρίζωμα του αθροίσματος των τετραγώνων των συνιστωσών τους. Με τον τρόπο αυτό, όλα τα διανύσματα αποκτούν μοναδιαίο μήκος και η ευκλείδεια απόσταση που χρησιμοποιεί ο αλγόριθμος k-means καθίσταται ισοδύναμη με τη γωνιακή (cosine) απόσταση, καθώς η σύγκριση βασίζεται πλέον στη διεύθυνση και όχι στο μέτρο των διανυσμάτων.

Τυπικά, για ένα σύνολο R εκτελέσεων, ο πίνακας ορίζεται ως:

$$C_{ij} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}(\ell_i^{(r)} = \ell_j^{(r)})$$

όπου $\ell_i^{(r)}$είναι η ετικέτα συστάδας του κειμένου i στην εκτέλεση r, και $\mathbb{I}$ η ενδεικτική συνάρτηση(παίρνει την τιμή 1 όταν μια λογική συνθήκη ισχύει και την τιμή 0 όταν δεν ισχύει). Η τιμή C_{ij} λαμβάνει τιμές στο διάστημα [0,1] και εκφράζει τη σταθερότητα της συν-ομαδοποίησης των δύο κειμένων.

Σημαντικό στοιχείο της προσέγγισης που πρέπει να ξεκαθαριστεί είναι ότι το ίδιο το αναγνωριστικό της συστάδας δεν έχει καμία σημασία. Δεν μας ενδιαφέρει αν δύο κείμενα ανήκουν στη «συστάδα 1» ή «συστάδα 3», αλλά μόνο αν ανήκουν στην ίδια συστάδα μέσα στην ίδια εκτέλεση. Αυτό καθιστά την προσέγγιση ανεξάρτητη από τον αριθμό συστάδων kκαι επιτρέπει τη συνδυαστική αξιοποίηση εκτελέσεων με διαφορετικές τιμές k.

Για να μπορέσει ο πίνακας συν-συσχέτισης να χρησιμοποιηθεί σε ιεραρχική συσταδοποίηση, μετατρέπεται σε πίνακα αποστάσεων ορίζοντας:

$$D_{ij} = 1 - C_{ij}$$

Με αυτόν τον ορισμό, μικρές τιμές του D_{ij}αντιστοιχούν σε ζεύγη κειμένων που συν-ομαδοποιούνται συστηματικά, ενώ μεγάλες τιμές υποδηλώνουν ζεύγη που σπάνια ή ποτέ δεν εμφανίζονται στην ίδια συστάδα.

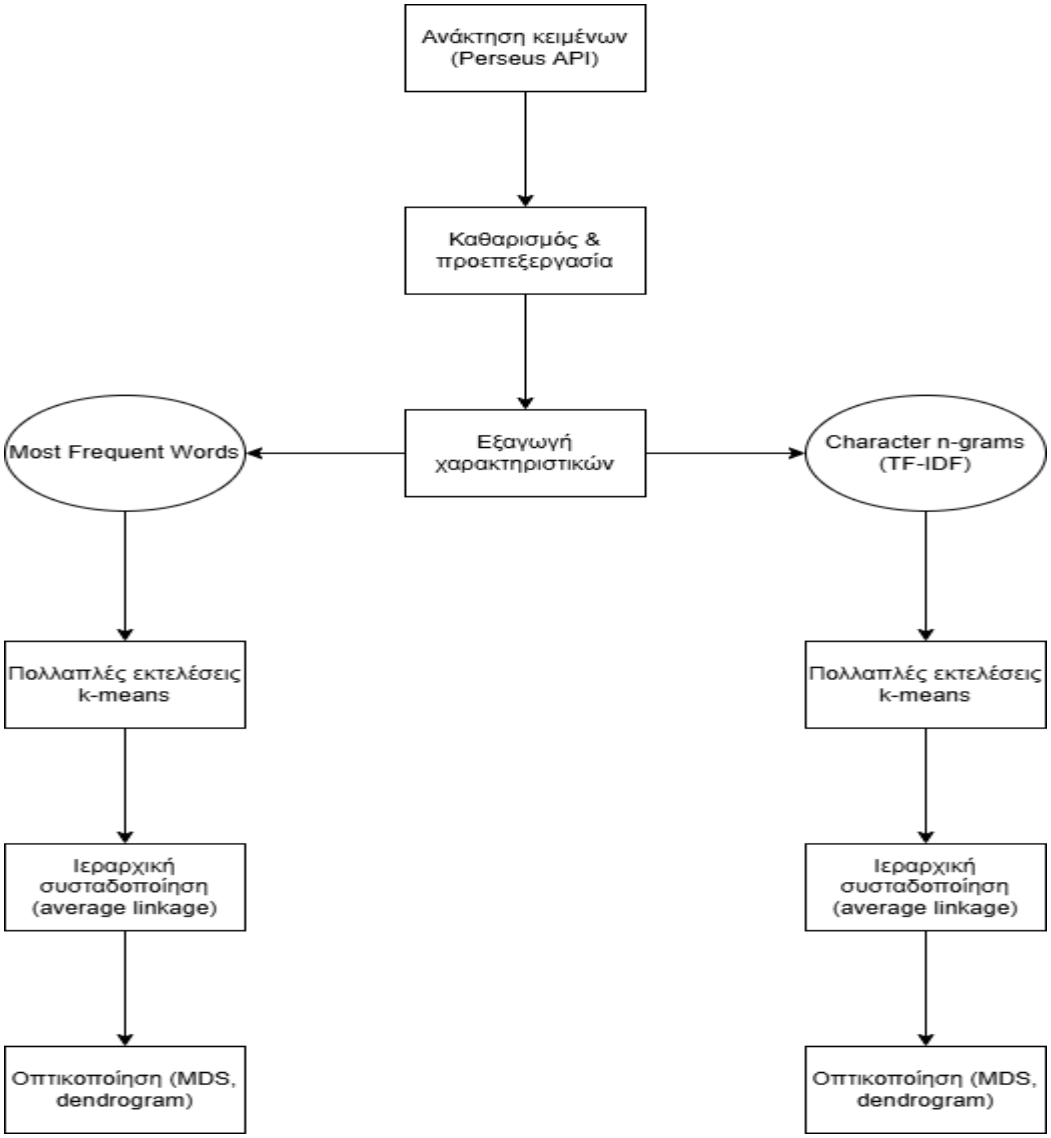
Στο τελικό στάδιο εφαρμόζεται ιεραρχική συσταδοποίηση πάνω στον πίνακα αποστάσεων D. Σε αντίθεση με τον k-means, η ιεραρχική συσταδοποίηση δεν απαιτεί προκαθορισμένο αριθμό συστάδων και δεν παράγει μία μοναδική διάσπαση των δεδομένων. Αντίθετα, κατασκευάζει μια ιεραρχία συγχωνεύσεων, ξεκινώντας από μεμονωμένα αντικείμενα και συγχωνεύοντας σταδιακά συστάδες με βάση ένα κριτήριο απόστασης.

Η μέθοδος average linkage ορίζει την απόσταση μεταξύ δύο συστάδων ως τον μέσο όρο των αποστάσεων όλων των ζευγών στοιχείων που ανήκουν στις δύο συστάδες. Με αυτόν τον τρόπο, κάθε συγχώνευση λαμβάνει υπόψη τη συνολική σχέση μεταξύ των ομάδων και όχι μόνο τα πιο κοντινά ή πιο απομακρυσμένα στοιχεία τους. Το αποτέλεσμα είναι μια πιο εξισορροπημένη ιεραρχική δομή, λιγότερο ευαίσθητη σε ακραίες τιμές.

Η ιεραρχική φύση της μεθόδου έγκειται στο γεγονός ότι κάθε συγχώνευση διατηρείται και ενσωματώνεται σε ανώτερα επίπεδα, δημιουργώντας ένα δενδρικό σχήμα συγγένειας μεταξύ των κειμένων. Αυτό διαφοροποιεί ριζικά την προσέγγιση από τον k-means, ο οποίος παράγει μόνο ένα επίπεδο ομαδοποίησης και δεν παρέχει πληροφορία για τις σχετικές αποστάσεις μεταξύ συστάδων σε διαφορετικά επίπεδα ανάλυσης.

3.2.4 Οπτικοποίηση

Η οπτικοποίηση των αποτελεσμάτων αποσκοπεί στη διερευνητική κατανόηση της δομής ομοιότητας που προκύπτει από τον πίνακα συν-συσχέτισης του EAC. Καθώς ο EAC παράγει έναν πλήρη, συμμετρικό πίνακα ομοιότητας μεταξύ των κειμένων, το πρόβλημα μετασχηματίζεται σε πρόβλημα ανάλυσης αποστάσεων και σχέσεων σε υψηλό διαστατικό χώρο.



Εικόνα 3.1: Συνοπτική απεικόνιση του υπολογιστικού pipeline της μελέτης

Αρχικά, ο πίνακας συνσυσχέτισης C μετατρέπεται σε πίνακα απόστασης D, ο οποίος εκφράζει την αποσυσχέτιση μεταξύ ζευγών κειμένων. Η μετατροπή αυτή βασίζεται στη συμπληρωματική σχέση ομοιότητας–απόστασης, όπου υψηλές τιμές συνσυσχέτισης αντιστοιχούν σε μικρές αποστάσεις. Ο πίνακας D χρησιμοποιείται στη συνέχεια ως είσοδος σε μεθόδους οπτικοποίησης και ιεραρχικής ανάλυσης.

Για τη μείωση της διάστασης και την οπτική αναπαράσταση των σχέσεων εφαρμόζεται η μέθοδος Multidimensional Scaling (MDS). Η MDS επιδιώκει να τοποθετήσει τα κείμενα σε έναν

χαμηλοδιάστατο χώρο (συνήθως δύο διαστάσεων), έτσι ώστε οι μεταξύ τους αποστάσεις να προσεγγίζουν όσο το δυνατόν πιστότερα τις αποστάσεις του αρχικού πίνακα. Σε αντίθεση με γραμμικές τεχνικές όπως η PCA, η MDS δεν βασίζεται στη διακύμανση των αρχικών χαρακτηριστικών, αλλά αποκλειστικά στη δομή των αποστάσεων, γεγονός που την καθιστά καταλληλότερη για πίνακες που προκύπτουν από διαδικασίες συσταδοποίησης και όχι από άμεσες διανυσματικές αναπαραστάσεις.

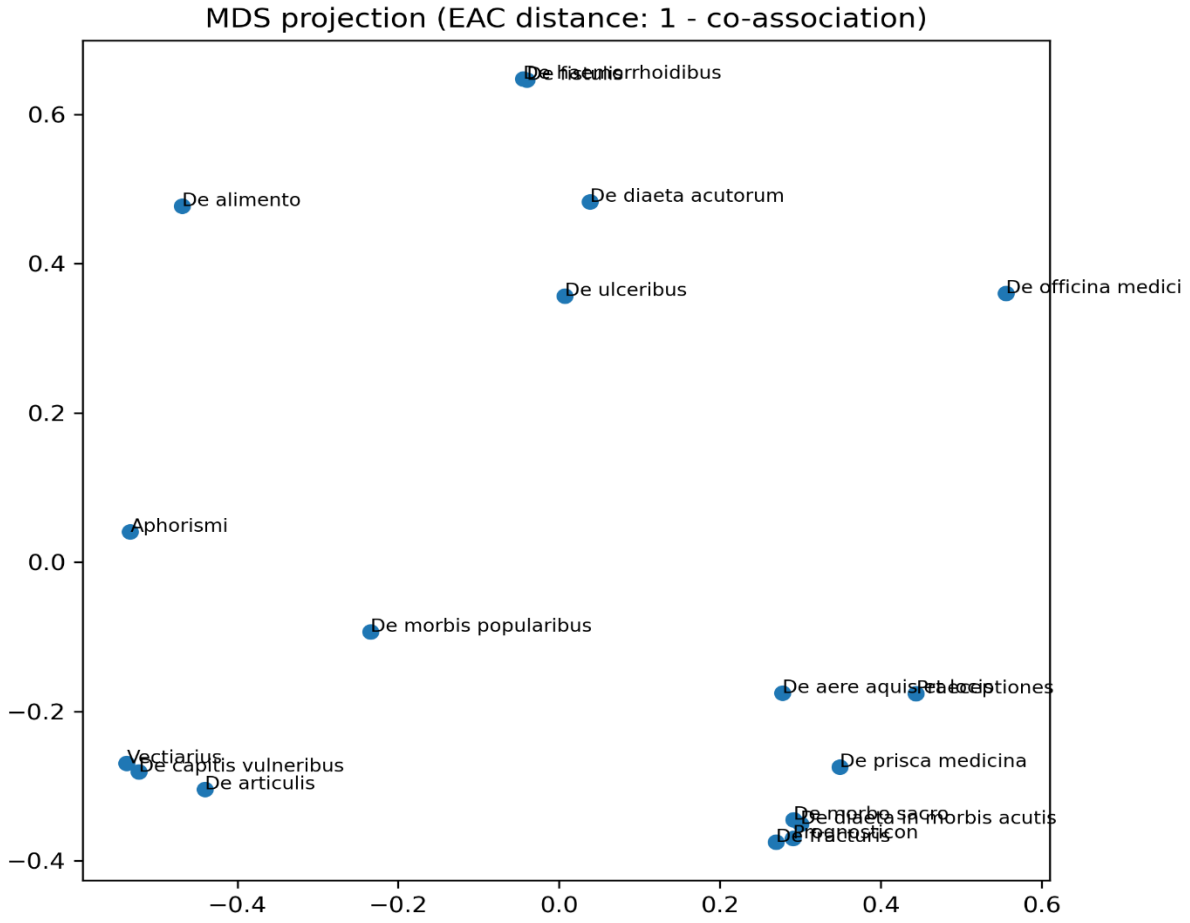
Η μεθοδολογία της εικόνας 3.1 που αναλύθηκε στα προηγούμενα υποκεφάλαια ορίζει ένα ενιαίο και αναπαραγώγιμο πλαίσιο ανάλυσης, το οποίο επιτρέπει τη συγκριτική διερεύνηση διαφορετικών υφολογικών αναπαραστάσεων στο ίδιο σώμα κειμένων.

Κεφάλαιο 4ο: Αποτελέσματα

4.1 Αποτελέσματα με βάση character n-grams

Η προβολή MDS (εικόνα 4.1), βασισμένη στην απόσταση EAC, παρέχει μία συνοπτική γεωμετρική αναπαράσταση των υφολογικών σχέσεων μεταξύ των πραγματειών. Η σχετική εγγύτητα των σημείων αντανακλά τη συχνότητα με την οποία τα αντίστοιχα κείμενα ομαδοποιήθηκαν μαζί στις πολλαπλές εκτελέσεις του k-means.

Παρατηρείται σαφής ομαδοποίηση των πραγματειών De morbis acutis (Περὶ διαίτης ἐν ὀξέσι νοσήμασι), Prognosticon (Προγνωστικόν) και De morbo sacro (Περὶ ἱερῆς νούσου), οι οποίες εμφανίζονται χωρικά συγκεντρωμένες. Η εγγύτητα αυτή συνάδει με τη φιλολογική παράδοση, η οποία



Εικόνα 4.1 MDS οπτικοποίηση για EAC με char n-grams.

συχνά τις εντάσσει στον λεγόμενο «Κνίδιο» ή πρώιμο ιπποκρατικό πυρήνα, με έμφαση στη διάγνωση και την πρόγνωση(αν και, όπως έχει επισημανθεί, αυτή η διάκριση θεωρείται παρωχημένη). Αντιθέτως, πραγματείες όπως το Aphorismi (Ἀφορισμοί) και το De alimento (Περὶ τροφῆς) καταλαμβάνουν απομονωμένες θέσεις στον χώρο, υποδηλώνοντας υψηλότερο βαθμό υφολογικής ιδιαιτερότητας. Η απομόνωση των Αφορισμών είναι δικαιολογημένη, δεδομένης της αποσπασματικής και γνωμικής μορφής του έργου.

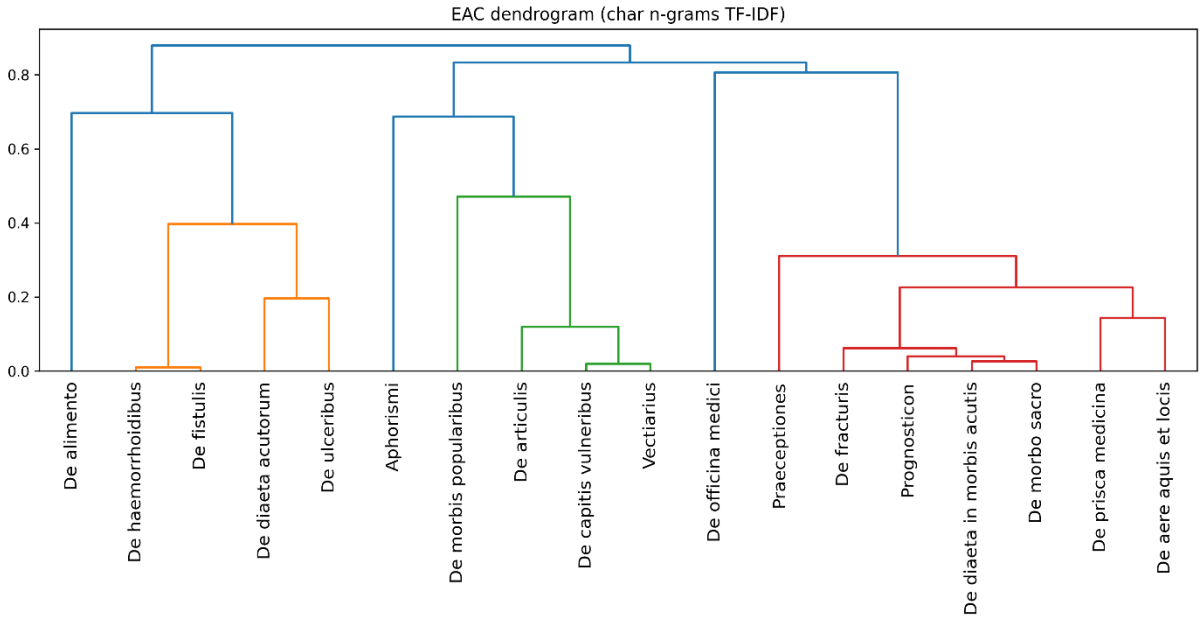
Παρατηρείται σχετική εγγύτητα των De articulis (Περὶ ἄρθρων), De capitis vulneribus (Περὶ τῶν ἐν κεφαλῇ τρωμάτων) και Vectiarius (Μοχλικός), που σχηματίζουν μια χαλαρή ομάδα, συμβατή με τον τεχνικό και χειρουργικό χαρακτήρα των κειμένων.

Το δενδρόγραμμα EAC παρουσιάζει συνεκτική και ερμηνεύσιμη δομή. Η χρήση της συνσυσχέτισης ως μέτρου συχνότητας κοινής ανάθεσης σε συστάδες, επιτρέπει την ανάδειξη σταθερών σχέσεων που επαναλαμβάνονται σε διαφορετικές παραμετροποιήσεις του k-means. Στην ιεραρχική αναπαράσταση της εικόνας 4.2, οι πρώιμες συγχωνεύσεις, αυτές που πραγματοποιούνται σε χαμηλά επίπεδα του δενδρογράμματος, αντανακλούν στενή υφολογική συγγένεια, ενώ οι καθυστερημένες συγχωνεύσεις αποτυπώνουν βαθύτερους και πιο γενικούς διαχωρισμούς στο corpus.

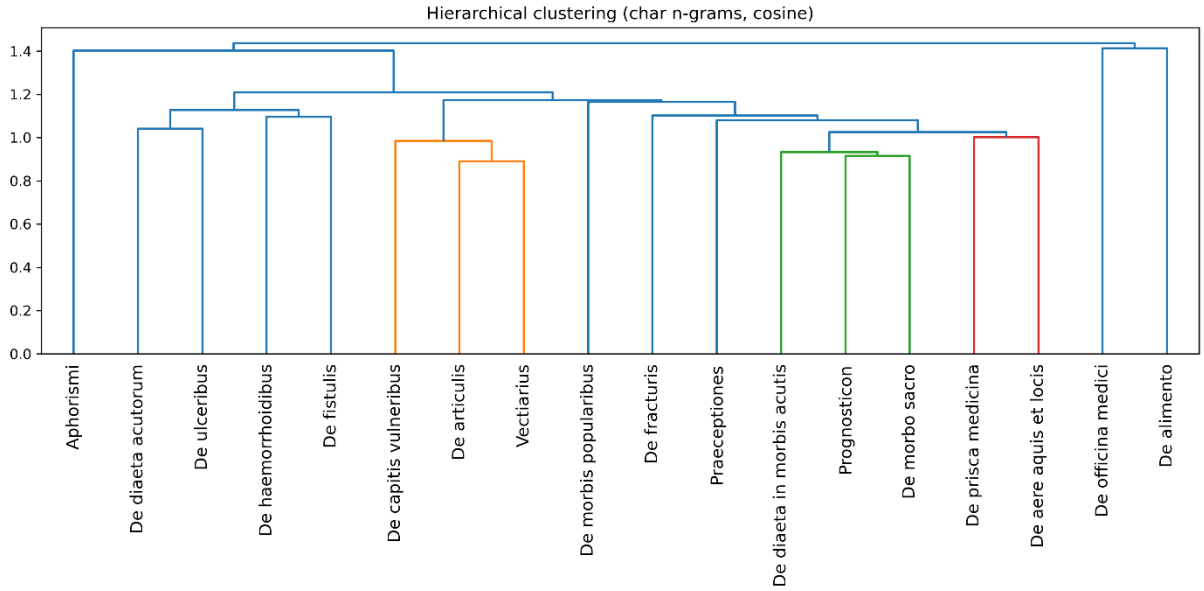
Εντοπίζεται καθαρή συστάδα που περιλαμβάνει τα De morbis acutis (Περὶ διαίτης ἐν ὀξέσι νοσήμασι), Prognosticon (Προγνωστικόν), De morbo sacro (Περὶ ἱερῆς νοῦσου) και De fracturis (Περὶ ἀγμῶν), γεγονός που ενισχύει την υπόθεση κοινών υφολογικών επιλογών, πιθανώς συνδεδεμένων με τη διδακτική και κλινική στόχευση των κειμένων.

Παράλληλα, τα De articulis (Περὶ ἄρθρων), De capitis vulneribus (Περὶ τῶν ἐν κεφαλῇ τρωμάτων) και Vectiarius (Μοχλικός) σχηματίζουν μια διακριτή υποομάδα, επιβεβαιώνοντας τη συνοχή του «χειρουργικού» υποσυνόλου. Η σταθερότητα αυτής της ομάδας στο πλαίσιο του EAC υποδηλώνει ότι η υφολογική συγγένεια δεν αποτελεί τεχνούργημα μεμονωμένης εκτέλεσης.

Αντίθετα, τα Aphorismi (Ἀφορισμοί) και De alimento (Περὶ τροφῆς) παραμένουν περιφερειακά, επιβεβαιώνοντας την ιδιότυπη θέση τους εντός του corpus. Το δενδρόγραμμα της εικόνας 4.3, που προκύπτει από ιεραρχική συσταδοποίηση (χωρίς να προηγούνται τα πολλαπλά runs k-means) με cosine distance επί των character n-grams αποτυπώνει παρόμοιες αλλά λιγότερο σταθερές δομές. Παρατηρούνται τοπικές συγγένειες, όπως η σύζευξη των De articulis (Περὶ ἄρθρων) και Vectiarius (Μοχλικός), καθώς και η σχετική εγγύτητα των Prognosticon (Προγνωστικόν) και De morbis acutis (Περὶ διαίτης ἐν ὀξέσι νοσήμασι). Ωστόσο, ορισμένες συγχωνεύσεις πραγματοποιούνται σε σχετικά υψηλά επίπεδα απόστασης, γεγονός



Εικόνα 4.2: Δενδρόγραμμα για char n-grams EAC



Εικόνα 4.3: Δενδρόγραμμα hierarchical clustering για char n-grams

που υποδηλώνει ότι το αποτέλεσμα είναι ευαίσθητο τόσο στην επιλογή της αρχικής ομαδοποίησης όσο και στη μονοσήμαντη χρήση μίας μόνο εκτέλεσης. Αυτό καθιστά τη διάκριση μεταξύ σταθερών υφολογικών σχέσεων και τυχαίων συσχετίσεων δυσχερή.

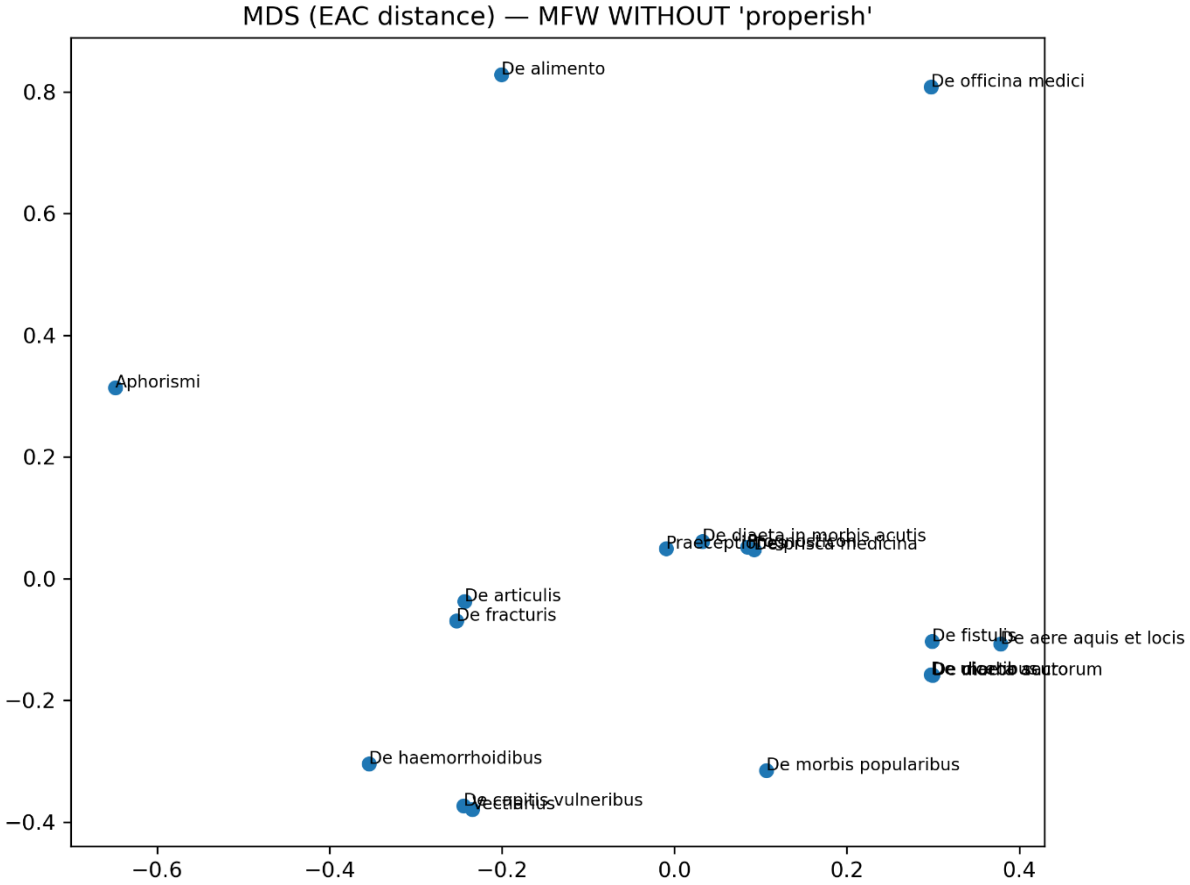
Επικεντρώνοντας το ενδιαφέρον στην εικόνα 4.2, παρατηρούμε ότι:

- η ομαδοποίηση των πραγματειών De ulceribus (Περὶ ἑλκῶν), De fistulis (Περὶ συρίγγων), De haemorrhoidibus (Περὶ αἱμορροΐδων), De diaeta in morbis acutis (Περὶ διαίτης ἐν ὀξέσι νοσήμασι) και De alimento (Περὶ τροφῆς) μπορεί να ερμηνευθεί στο πλαίσιο της έντονα πρακτικής και θεραπευτικής τους στόχευσης. Όπως επισημαίνει η Craik, όλα αυτά τα κείμενα χαρακτηρίζονται από τεχνικό ύφος, τυποποιημένες διατυπώσεις και έμφαση σε διαδικαστικές οδηγίες.
- Η απομόνωση των Αφορισμών σε ξεχωριστό cluster επιβεβαιώνει την ιδιοτυπία του κειμένου, το οποίο, σύμφωνα με τη Craik, διαφέρει ριζικά ως προς το ύφος και τη δομή από τις λοιπές ιπποκρατικές πραγματείες, λόγω της γνωμικής και εξαιρετικά συμπυκνωμένης μορφής του, διαφοροποιώντας το ριζικά σε ύφος από τις άλλες πραγματείες, ακόμη και από κείμενα με συναφή θεματική, όπως το Προγνωστικόν.
- το τρίτο cluster περιλαμβάνει τις πραγματείες De morbis popularibus (Περὶ νούσων δημοσίων), De articulis (Περὶ ἄρθρων), De capitis vulneribus (Περὶ τραυμάτων κεφαλῆς) και Mochlicon (Μοχλικόν). Η ομαδοποίηση αυτών των κειμένων μπορεί να ερμηνευθεί στο πλαίσιο της κοινής τους τεχνικής και περιγραφικής στόχευσης. Όπως επισημαίνει η Craik, οι τελευταίες τρεις πραγματείες αποτελούν ένα σώμα κειμένων που εστιάζει στη μηχανική της θεραπείας, στην ανατομική παρατήρηση και στην πρακτική εφαρμογή χειρισμών, με έντονα επαναλαμβανόμενο λεξιλόγιο και τυποποιημένες συντακτικές δομές. Η πρώτη, αν και θεματικά διαφοροποιημένη, παρουσιάζει κατά την Craik έναν καταγραφικό και περιγραφικό χαρακτήρα, με έμφαση στην παρατήρηση και όχι στη θεωρητική επεξεργασία, γεγονός που μπορεί να εξηγεί τη μερική υφολογική της σύγκλιση με τις καθαρά τεχνικές πραγματείες του cluster.

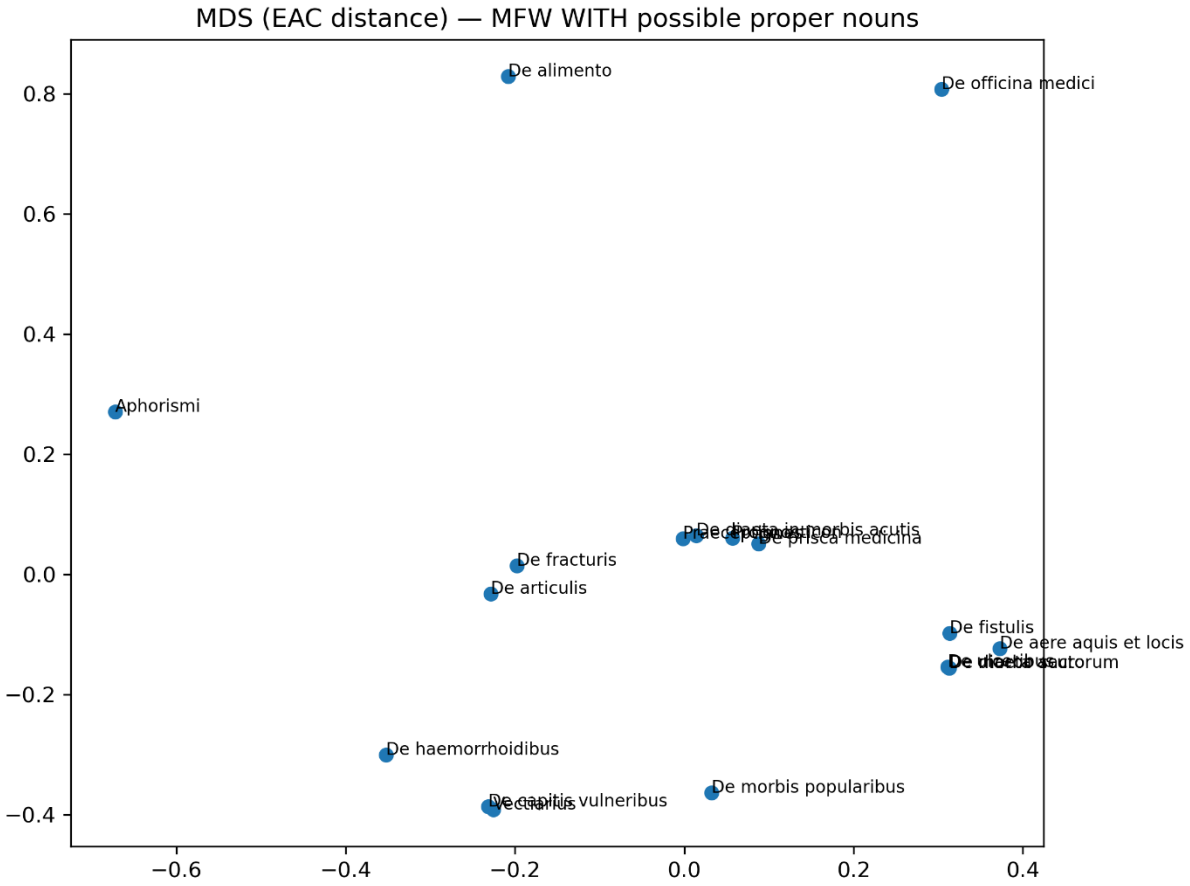
- το τέταρτο cluster συγκροτείται από τις πραγματείες Praeceptiones (Παραινέσεις), Prognosticon (Προγνωστικόν), De diaeta in morbis acutis (Περὶ διαίτης ὀξέων), De morbo sacro (Περὶ ἱερῆς νούσου), De prisca medicina (Περὶ ἀρχαίης ἰητρικῆς), De aere aquis et locis (Περὶ ἀέρων ὑδάτων τόπων) και De officina medici (Περὶ ἰητροῦ ἐργαστηρίου). Η ομαδοποίηση αυτών των κειμένων είναι σε μεγάλο βαθμό συμβατή με τη φιλολογική αποτίμηση της Craik, καθώς πρόκειται για πραγματείες που χαρακτηρίζονται από υψηλή αφαιρετικότητα, επιχειρηματολογικό ύφος και συστηματική ανάπτυξη γενικών αρχών σχετικά με τη φύση της ιατρικής γνώσης, τη διάγνωση και την πρόγνωση, καθώς και τη σχέση περιβάλλοντος και νόσου. Η ένταξη της De officina medici στο ίδιο cluster μπορεί να ερμηνευθεί ως ένδειξη συγγένειας σε επίπεδο μετα-ιατρικού λόγου, δεδομένου ότι το κείμενο αυτό λειτουργεί ως στοχασμός πάνω στην ιατρική πρακτική και τον ρόλο του ιατρού.

4.2 Αποτελέσματα με βάση λειτουργικές λέξεις/MFW

Η ανάλυση με βάση τις συχνότερες λέξεις εφαρμόστηκε σε δύο παραλλαγές του ίδιου χαρακτηριστικού χώρου, οι οποίες διαφέρουν ως προς τον βαθμό φιλτραρίσματος λεξιλογικών στοιχείων με πιθανή θεματική ή ονομαστική φόρτιση. Η σύγκριση των αποτελεσμάτων των εικόνων 4.4, 4.5, 4.6 και 4.7 δείχνει ότι η βασική δομή των συστάδων παραμένει σταθερή και στις δύο περιπτώσεις, τόσο ως προς τη σύνθεση όσο και ως προς τις κύριες συγγένειες μεταξύ των κειμένων. Οι παρατηρούμενες διαφοροποιήσεις εντοπίζονται κυρίως σε δευτερεύοντα χαρακτηριστικά, όπως το ύψος των συγχωνεύσεων και η εσωτερική διάταξη των συστάδων, χωρίς να επηρεάζουν ουσιωδώς τη συνολική ομαδοποίηση.



Εικόνα 4.4: MDS για MFW EAC (χωρίς κύρια ονόματα)



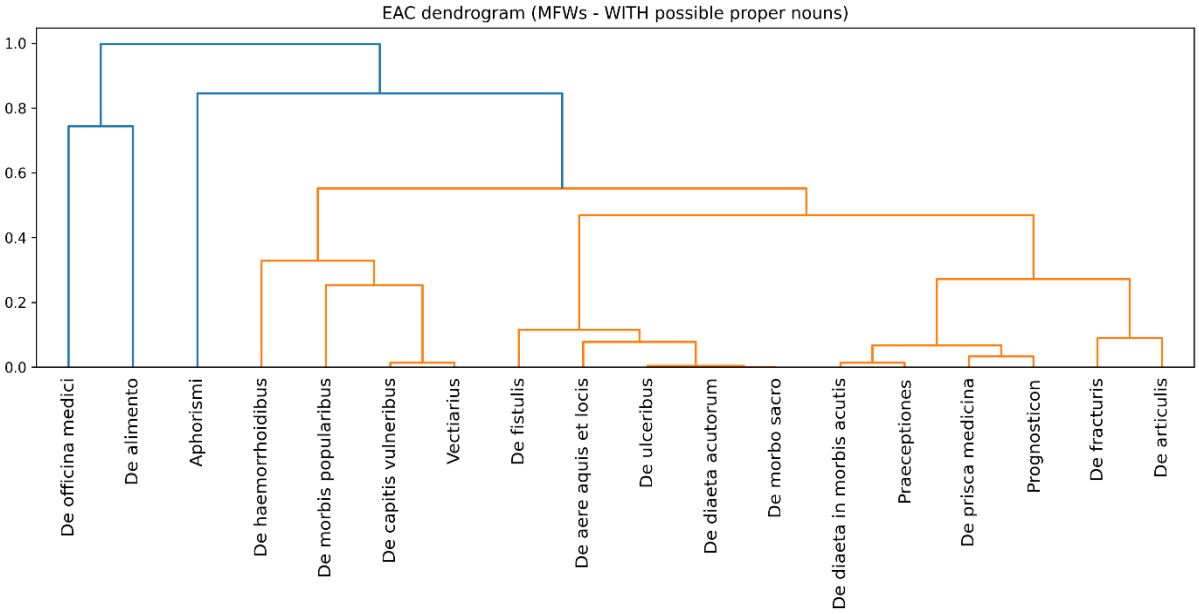
Εικόνα 4.5: MDS για MFW EAC (με κύρια ονόματα)

Οι συστάδες που διακρίνονται είναι οι εξής:

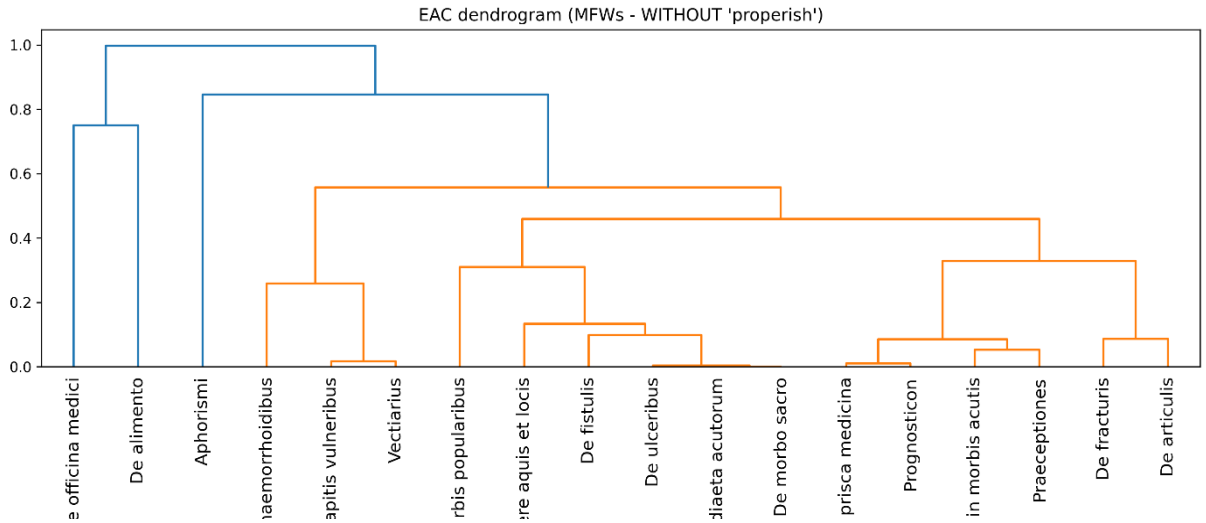
- Aphorismi (Ἀφορισμοί): το κείμενο εμφανίζεται ως μεμονωμένη και σχετικά απομονωμένη συστάδα. Η θέση του παραμένει σταθερή και στις δύο παραλλαγές (με και χωρίς φιλτράρισμα ονομαστικών στοιχείων), γεγονός που υποδηλώνει έντονη υφολογική ιδιαιτερότητα και χαμηλή εγγύτητα με τα υπόλοιπα κείμενα.
- De alimento (Περὶ τροφῆς), De officina medici (Περὶ ἰητροῦ ἐργαστηρίου): τα δύο κείμενα σχηματίζουν σταθερό διμελές cluster και στις δύο αναλύσεις. Η συστάδα παρουσιάζει υψηλή συνοχή, με μικρές μόνο μεταβολές στις αποστάσεις συγχώνευσης μεταξύ των δύο μεθόδων, χωρίς όμως αλλαγή στη σύνθεσή της. Βέβαια, πρόκειται για χαλαρή ομαδοποίηση.
- De articulis (Περὶ ἄρθρων), De fracturis (Περὶ καταγμάτων): πρόκειται για ιδιαίτερα συμπαγή συστάδα, με πρώιμη συγχώνευση στο δενδρόγραμμα. Η σταθερότητά της διατηρείται ανεξάρτητα από την παρουσία ή απουσία λέξεων με πιθανή θεματική φόρτιση, υποδηλώνοντας ισχυρή υφολογική συγγένεια.
- De capitis vulneribus (Περὶ τραυμάτων κεφαλῆς), Vectiarius / Mochlicon (Μοχλικός): τα δύο κείμενα ομαδοποιούνται με συνέπεια και στις δύο εκδοχές.
- De haemorrhoidibus (Περὶ αἱμορροΐδων), De morbis popularibus (Περὶ νούσων δημοσίων): αυτό το cluster παρουσιάζει μέτρια συνοχή και μεγαλύτερη απόσταση συγχώνευσης. Αν

και η βασική ομαδοποίηση διατηρείται, η εσωτερική της σταθερότητα είναι χαμηλότερη σε σύγκριση με άλλες συστάδες.

- Praeceptiones (Παραινέσεις), Prognosticon (Προγνωστικόν), De diaeta in morbis acutis (Περὶ διαίτης ὀξέων), De morbo sacro (Περὶ ἱερῆς νούσου), De prisca medicina (Περὶ ἀρχαίης ἰητρικῆς), De aere aquis et locis (Περὶ ἀέρων ὑδάτων τόπων): σχηματίζουν ευρύτερη και λιγότερο συμπαγή συστάδα, με διαβαθμίσεις στην εσωτερική τους εγγύτητα. Η συστάδα είναι σταθερή ως προς τη σύνθεση, αλλά εμφανίζει διαφοροποιήσεις στο ύψος και στη σειρά των συγχωνεύσεων μεταξύ των δύο παραλλαγών, γεγονός που υποδηλώνει μεγαλύτερη εσωτερική ετερογένεια.



Εικόνα 4.6: Δενδρόγραμμα για MFW EAC (με κύρια ονόματα)



Εικόνα 4.7: Δενδρόγραμμα για MFW EAC (χωρίς κύρια ονόματα)

Η ομαδοποίηση που προκύπτει από την ανάλυση με βάση τις συχνότερες λέξεις παρουσιάζει, σε γενικές γραμμές, καλή συμβατότητα με τις βασικές γραμμές της φιλολογικής αποτίμησης του Ιπποκρατικού σώματος, όπως αυτή αποτυπώνεται στη μελέτη της Craik. Ορισμένα κείμενα εμφανίζονται σταθερά απομονωμένα (Ἀφορισμοί, Περὶ τροφῆς, Περὶ ἰητροῦ ἐργαστηρίου), γεγονός που συνιστά ένδειξη έντονης υφολογικής ιδιαιτερότητας, συμβατής με τη μορφή και τη δομή τους, χωρίς να προϋποθέτει κατ' ανάγκην κοινή συγγραφική προέλευση.

Άλλες ομάδες κειμένων σχηματίζουν ιδιαίτερα συμπαγείς συστάδες, με πρώιμες συγχωνεύσεις στο δενδρόγραμμα και ελάχιστη μεταβολή μεταξύ των δύο παραλλαγών της ανάλυσης. Χαρακτηριστικά παραδείγματα αποτελούν η συστάδα που περιλαμβάνει τα Περὶ ἄρθρων και Περὶ καταγμάτων, καθώς και εκείνη που περιλαμβάνει τα Περὶ τραυμάτων κεφαλῆς και Περὶ αἱμορροΐδων. Η σταθερότητα αυτών των ομαδοποιήσεων συνάδει με την παρατήρηση της Craik ότι πρόκειται για κείμενα με κοινό τεχνικό προσανατολισμό και συγκρίσιμη γλωσσική οργάνωση, στοιχεία που αντανακλώνται με σαφήνεια σε υφολογικά χαρακτηριστικά χαμηλού επιπέδου, όπως οι συχνότητες λειτουργικών λέξεων.

Παράλληλα, διακρίνεται και μία ευρύτερη ομάδα κειμένων που ομαδοποιείται με μικρότερη εσωτερική συνοχή και με συγχωνεύσεις σε υψηλότερα επίπεδα απόστασης. Σε αυτήν εντάσσονται οι πραγματείες Παραινέσεις, Προγνωστικόν, Περὶ διαίτης ὀξέων, Περὶ ἱερῆς νούσου, Περὶ ἀρχαίης ἰητρικῆς και Περὶ ἀέρων ὑδάτων τόπων. Η εικόνα αυτή είναι, υπό μία έννοια, συμβατή με την εκτίμηση της Craik ότι πρόκειται για πραγματείες με αυξημένο βαθμό αφαιρετικότητας και θεωρητικού ή προγραμματικού χαρακτήρα, οι οποίες, αν και μοιράζονται κοινές εννοιολογικές αφετηρίες, δεν συγκροτούν ένα αυστηρά ομοιογενές υφολογικό σύνολο.

4.3 Σύγκριση των δύο μεθόδων

Η σύγκριση των αποτελεσμάτων που προκύπτουν από τις δύο αναπαραστάσεις χαρακτηριστικών, αναδεικνύει τόσο σημεία σύγκλισης όσο και ουσιώδεις διαφοροποιήσεις στον τρόπο με τον οποίο αποτυπώνεται η υφολογική συγγένεια των κειμένων.

Σε γενικές γραμμές, και οι δύο προσεγγίσεις οδηγούν σε παρόμοιες μακροδομές ομαδοποίησης, καθώς βασικές ομάδες κειμένων εμφανίζονται σταθερά και στις δύο αναλύσεις. Ιδιαίτερα οι τεχνικές και χειρουργικές πραγματείες σχηματίζουν συνεκτικές συστάδες με μικρές εσωτερικές αποστάσεις και πρώιμες συγχωνεύσεις στα δενδρογράμματα, κάτι που υποδεικνύει έντονη υφολογική ομοιογένεια ανεξάρτητα από το επίπεδο ανάλυσης των χαρακτηριστικών. Η σταθερότητα αυτή παρατηρείται τόσο στις αναπαραστάσεις με χαρακτήρες όσο και σε εκείνες που βασίζονται σε λειτουργικές λέξεις.

Οι διαφορές μεταξύ των δύο προσεγγίσεων γίνονται εμφανείς στον τρόπο με τον οποίο διαχειρίζονται κείμενα με αυξημένο βαθμό αφαιρετικότητας ή θεωρητικού προσανατολισμού. Στην περίπτωση των character n-grams, τα κείμενα αυτά τείνουν να ομαδοποιούνται με μεγαλύτερη συνοχή, γεγονός που υποδηλώνει ότι η ανάλυση σε επίπεδο χαρακτήρων συλλαμβάνει λεπτότερα μορφολογικά και ορθογραφικά μοτίβα, λιγότερο εξαρτημένα από θεματικό λεξιλόγιο. Αντίθετα, στην ανάλυση με βάση τις συχνότερες λέξεις, οι ίδιες πραγματείες παρουσιάζουν μεγαλύτερη διασπορά και συγχωνεύονται σε υψηλότερα επίπεδα απόστασης, στοιχείο που υποδηλώνει αυξημένη ευαισθησία σε διαφοροποιήσεις λεξιλογικής επιλογής.

Συνολικά, η ανάλυση με character n-grams φαίνεται να παράγει πιο συμπαγείς και σταθερές συστάδες, ιδίως σε περιπτώσεις κειμένων με θεωρητικό ή προγραμματικό χαρακτήρα, ενώ η προσέγγιση με Most Frequent Words αναδεικνύει εντονότερα τη διαφοροποίηση που προκύπτει από λεξιλογικές και

θεματικές επιλογές. Η σύγκλιση των δύο μεθόδων στα βασικά σχήματα ομαδοποίησης, σε συνδυασμό με τις αποκλίσεις τους σε λεπτομέρειες, προσφέρει μια συμπληρωματική εικόνα της υφολογικής δομής του σώματος.

Κεφάλαιο 5ο: Μελλοντικές επεκτάσεις

Η παρούσα ανάλυση αφήνει ανοιχτά περιθώρια για περαιτέρω μεθοδολογική και ερμηνευτική διεύρυνση.

Καίριας σημασίας για τη μελλοντική έρευνα κρίνεται η διεύρυνση του σώματος κειμένων, τόσο ως προς τον αριθμό των πραγματειών όσο και ως προς την ένταξη διαφορετικών εκδοχών ή παραδόσεων του ίδιου κειμένου. Η αύξηση του μεγέθους του corpus θα επέτρεπε την πιο σταθερή εκτίμηση των συστάδων και την εξέταση της ανθεκτικότητας των αποτελεσμάτων σε μεγαλύτερη ποικιλία υφολογικών δεδομένων.

Ένα πιο πλήρες σώμα κειμένων επιτρέπει ενδιαφέρουσες εναλλακτικές αναπαραστάσεις χαρακτηριστικών, ικανές ενδεχομένως να συλλάβουν διαφορετικές όψεις του ύφους, συμπληρωματικές προς εκείνες που αναδεικνύονται μέσω character n-grams και λειτουργικών λέξεων. Κάποιες από τις αναπαραστάσεις που θα μπορούσαν να χρησιμοποιηθούν αναφέρονται παρακάτω:

- Προσωδία: μία κατεύθυνση μελλοντικής έρευνας θα μπορούσε να αφορά την ανάλυση ρυθμικών χαρακτηριστικών του λόγου, κατά το πρότυπο πρόσφατων υφομετρικών μελετών σε λατινική πεζογραφία. Ενδεικτικά, οι Corbara et al. [32] προτείνουν την εξαγωγή ρυθμικών χαρακτηριστικών με βάση ακολουθίες συλλαβικής ποσότητας (μακρών και βραχέων συλλαβών), οι οποίες κωδικοποιούνται ως συμβολικές ακολουθίες και αναλύονται μέσω character n-grams συγκεκριμένου εύρους (π.χ. 3–7 συλλαβών), δείχνοντας ότι τέτοιες ρυθμικές αναπαραστάσεις βελτιώνουν με στατιστικά σημαντικό τρόπο την απόδοση μοντέλων συγγραφικής απόδοσης σε λατινική πεζογραφία, ακόμη και όταν συνδυάζονται με αυστηρά θεματικά ουδέτερα χαρακτηριστικά. Μια αντίστοιχη προσέγγιση για τα αρχαία ελληνικά θα ήταν θεωρητικά ιδιαίτερα ελκυστική. Επί του παρόντος δεν υπάρχουν διαθέσιμα υπολογιστικά εργαλεία που να υποστηρίζουν αξιόπιστη προσωδιακή ή ρυθμική ανάλυση αρχαίας ελληνικής πεζογραφίας. Υφιστάμενες εφαρμογές, όπως το Dactylo, περιορίζονται αποκλειστικά στην ανάλυση έμμετρων κειμένων και δεν είναι κατάλληλες για συνεχή πεζό λόγο. Ως εκ τούτου, η ενσωμάτωση ρυθμικών χαρακτηριστικών κρίνεται προς το παρόν μη εφαρμόσιμη και παραμένει αντικείμενο μελλοντικής διερεύνησης, υπό την προϋπόθεση ανάπτυξης κατάλληλης υπολογιστικής υποδομής.
- N-grams συντακτικών σχέσεων (syntactic n-grams): σε αντίθεση με τα παραδοσιακά n-grams, τα syntactic n-grams βασίζονται σε σχέσεις που προκύπτουν από το συντακτικό δέντρο εξάρτησης και αποτυπώνουν σταθερά μορφοσυντακτικά μοτίβα ανεξάρτητα από τη γραμμική σειρά των λέξεων. Στους Sidorov et al.[33], η μέθοδος αυτή υλοποιείται μέσω διάσχισης των dependency trees και εξαγωγής ακολουθιών σχέσεων εξάρτησης συγκεκριμένου μήκους, επιτρέποντας την ανίχνευση μη γραμμικών συντακτικών προτύπων που δεν είναι προσβάσιμα με surface-level n-grams, και έχει αποδειχθεί ιδιαίτερα αποτελεσματική σε εφαρμογές συγγραφικής απόδοσης, υπερέχοντας συχνά των απλών n-grams χαρακτήρων ή λέξεων. Αντίστοιχα, οι Ríos-Toledo et al.[12], εφαρμόζουν dependency-based syntactic n-grams και POS n-grams σε πλαίσιο επιβλεπόμενης μάθησης για την ανίχνευση διαχρονικών μεταβολών ύφους, δείχνοντας ότι οι συντακτικές σχέσεις διατηρούν υψηλή διακριτική ισχύ ακόμη και όταν εφαρμόζονται τεχνικές μείωσης διαστασιμότητας, γεγονός που αναδεικνύει τον έντονα topic-agnostic χαρακτήρα τους

- Η εφαρμογή τους στο ιπποκρατικό corpus προϋποθέτει, ωστόσο, πιο αξιόπιστα εργαλεία συντακτικής ανάλυσης για τα αρχαία ελληνικά.η συστηματική σύγκριση της EAC με άλλες μη εποπτευόμενες στρατηγικές, όπως μοντέλα βασισμένα σε πιθανοκρατικές παραδοχές ή τεχνικές πυκνότητας, ιδίως σε συνδυασμό με διαφορετικές αναπαραστάσεις χαρακτηριστικών είναι μία ακόμη ενδιαφέρουσα προοπτική. Μια τέτοια προσέγγιση θα επέτρεπε την αξιολόγηση της επίδρασης της ίδιας της μεθόδου συσταδοποίησης στα παρατηρούμενα αποτελέσματα και θα ενίσχυε τη μεθοδολογική γενικευσιμότητα των συμπερασμάτων. Συνολικά, οι παραπάνω κατευθύνσεις υποδεικνύουν ότι η παρούσα εργασία μπορεί να λειτουργήσει ως αφετηρία για μια ευρύτερη και βαθύτερη υπολογιστική διερεύνηση του Ιπποκρατικού corpus, σε στενή και γόνιμη συνομιλία με τη φιλολογική έρευνα.

Κεφάλαιο 6ο: Συμπεράσματα

Η παρούσα εργασία διερεύνησε τη δυνατότητα ανίχνευσης υφολογικών συγγενειών εντός του ιπποκρατικού corpus μέσω μη επιβλεπόμενων μεθόδων υπολογιστικής ανάλυσης κειμένου, με έμφαση στη σύγκριση διαφορετικών αναπαραστάσεων χαρακτηριστικών και στρατηγικών συσταδοποίησης. Παρά τη γνωστή ετερογένεια του corpus ως προς τη θεματική, τη χρονολόγηση και τη λειτουργία των επιμέρους πραγματειών, τα αποτελέσματα δείχνουν ότι είναι δυνατή η ανάδυση σταθερών και ερμηνεύσιμων ομαδοποιήσεων, οι οποίες παρουσιάζουν ουσιαστική συνάφεια με καθιερωμένες φιλολογικές εκτιμήσεις. Η ύπαρξη τέτοιων ομαδοποιήσεων υποδηλώνει ότι, πέρα από το θεματικό επίπεδο, τα κείμενα μοιράζονται επαναλαμβανόμενα υφολογικά μοτίβα που μπορούν να ανιχνευθούν ποσοτικά.

Η ανάλυση με βάση character n-grams, ιδίως όταν συνδυάζεται με τη μέθοδο Evidence Accumulation Clustering, αποδείχθηκε ιδιαίτερα αποτελεσματική στην ανάδειξη μακρο-υφολογικών δομών. Οι συστάδες που προέκυψαν αντανακλούν διαφορές στο επίπεδο αφαιρετικότητας, στη ρητορική οργάνωση και στη μεθοδολογική στόχευση των πραγματειών, στοιχεία που έχουν επισημανθεί και από τη φιλολογική έρευνα. Η σταθερότητα των συστάδων στο πλαίσιο του EAC ενισχύει την αξιοπιστία των αποτελεσμάτων, καθώς μειώνει την επίδραση μεμονωμένων παραμετροποιήσεων ή τυχαίων αρχικοποιήσεων και αναδεικνύει σχέσεις που επαναλαμβάνονται συστηματικά.

Αντίστοιχα, η ανάλυση με βάση λειτουργικές λέξεις (MFW) ανέδειξε συστάδες που είναι σε μεγάλο βαθμό συνεπείς τόσο μεταξύ τους όσο και σε σχέση με τα αποτελέσματα των character n-grams. Η σύγκριση των εκδοχών με και χωρίς λέξεις που λειτουργούν ως θεματικοί δείκτες έδειξε ότι η βασική δομή των ομαδοποιήσεων παραμένει σταθερή, γεγονός που υποδηλώνει ότι τα αποτελέσματα δεν καθορίζονται πρωτίστως από θεματικές συμπτώσεις, αλλά από υφολογικά χαρακτηριστικά χαμηλού επιπέδου. Η παρατήρηση αυτή ενισχύει την υπόθεση ότι οι συχνότητες λειτουργικών λέξεων μπορούν να λειτουργήσουν ως αξιόπιστοι δείκτες υφολογικής συγγένειας, ακόμη και σε μικρά και ιστορικά corpora.

Το περιορισμένο υλικό της μελέτης και η χρήση χαρακτηριστικών χαμηλού επιπέδου για την αναπαράσταση των κειμένων αφήνουν περιθώρια για μελλοντικές επεκτάσεις, τόσο ως προς το εύρος του corpus όσο και ως προς τις μεθόδους ανάλυσης. Η ενσωμάτωση επιπλέον πραγματειών ή πληρέστερων εκδόσεων των κειμένων θα μπορούσε να ενισχύσει τη στατιστική ισχύ των αποτελεσμάτων και να επιτρέψει την ανίχνευση λεπτότερων υφολογικών διαφοροποιήσεων. Παράλληλα, η διερεύνηση πιο σύνθετων χαρακτηριστικών, όπως συντακτικά πρότυπα ή ρυθμικές ιδιότητες του λόγου, θα μπορούσε να συμπληρώσει την παρούσα ανάλυση και να φωτίσει πτυχές του ύφους που δεν αποτυπώνονται πλήρως από επιφανειακές συχνότητες. Τέτοιες επεκτάσεις θα επέτρεπαν τη συστηματικότερη διερεύνηση της σχέσης ανάμεσα στη γλωσσική μορφή, τη λειτουργία των κειμένων και τη φιλολογική τους ταξινόμηση, ενισχύοντας περαιτέρω τον διάλογο ανάμεσα στην υπολογιστική και τη φιλολογική προσέγγιση του Ιπποκρατικού corpus.

Συνολικά, η εργασία καταδεικνύει ότι οι υπολογιστικές μέθοδοι δεν λειτουργούν ανταγωνιστικά προς τη φιλολογική ανάλυση, αλλά μπορούν να προσφέρουν ένα συμπληρωματικό και συστηματικό εργαλείο για τη διερεύνηση υφολογικών συγγενειών. Η σύγκριση διαφορετικών χαρακτηριστικών και μεθόδων ανέδειξε τη σημασία της επιλογής αναπαράστασης, καθώς και την αξία συνδυαστικών

προσεγγίσεων, όπως το EAC, για την ενίσχυση της σταθερότητας και της ερμηνευσιμότητας των αποτελεσμάτων. Τα ευρήματα ενισχύουν την άποψη ότι το ιπποκρατικό corpus παρουσιάζει διακριτές υφολογικές τάσεις, οι οποίες μπορούν να ανιχνευθούν με αξιόπιστο και αναπαραγώγιμο τρόπο μέσω σύγχρονων υπολογιστικών τεχνικών, ανοίγοντας τον δρόμο για περαιτέρω συνδυαστικές μελέτες φιλολογίας και υπολογιστικής ανάλυσης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] E. Craik, *The Hippocratic Corpus: Content and Context*. London: Routledge, 2014.

[2] J. Jouanna, *Hippocrates*, M. B. DeBevoise, trans. Baltimore: Johns Hopkins University Press, 2012.

[3] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, "Surveying stylometry Techniques and Applications," *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–36, Nov. 2017, doi: 10.1145/3132039.

[4] F. Mosteller and D. L. Wallace, "Inference in an authorship problem," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 275–309, Jun. 1963, doi: 10.1080/01621459.1963.10500849.

[5] M. A. Boukhaled and J.-G. Ganascia, "Using Function Words for Authorship Attribution: Bag-Of-Words vs. Sequential Rules," in *Natural Language Processing and Cognitive Science*, 2015, pp. 115–122. doi: 10.1515/9781501501289.115.

[6] D. L. Hoover, "Frequent collocations and authorial style," *Literary and Linguistic Computing*, vol. 18, no. 3, pp. 261–286, Sep. 2003, doi: 10.1093/llc/18.3.261.

[7] R. Gorman, "Morphosyntactic annotation in literary stylometry," *Information*, vol. 15, no. 4, p. 211, Apr. 2024, doi: 10.3390/info15040211.

[8] M. Koppel and J. Schler, "Authorship verification as a one-class classification problem," *Proceedings of the Twenty-first International Conference on Machine Learning*, p. 62, Jan. 2004, doi: 10.1145/1015330.1015448.

[9] O. Yavanoglu, "Intelligent authorship identification with using Turkish newspapers metadata," *2016 IEEE International Conference on Big Data*, vol. 3, pp. 1895–1900, Dec. 2016, doi: 10.1109/bigdata.2016.7840809.

[10] B. Verhoeven, I. Škrjanec, and S. Pollak, "Gender Profiling for Slovene Twitter communication: the Influence of Gender Marking, Content and Style," *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pp. 119–125, Jan. 2017, doi: 10.18653/v1/w17-1418.

[11] M. L. Brocardo, I. Traore, I. Woungang, and M. S. Obaidat, "Authorship verification using deep belief network systems," *International Journal of Communication Systems*, vol. 30, no. 12, Jan. 2017, doi: 10.1002/dac.3259.

[12] G. Ríos-Toledo, J. P. F. Posadas-Durán, G. Sidorov, and N. A. Castro-Sánchez, "Detection of changes in literary writing style using N-grams as style markers and supervised machine learning," *PLoS ONE*, vol. 17, no. 7, p. e0267590, Jul. 2022, doi: 10.1371/journal.pone.0267590.

[13] A. Caliskan-Islam *et al.*, "De-anonymizing programmers via code stylometry," *USENIX Security Symposium*, pp. 255–270, Aug. 2015, [Online]. Available: https://atc.usenix.org/system/files/conference/usenixsecurity15/sec15-paper-caliskan-islam.pdf

[14] D. Fifield, T. Follan, and E. Lunde, "Unsupervised authorship attribution," *arXiv (Cornell University)*, Mar. 2015, doi: 10.48550/arxiv.1503.07613.

[15] R. Layton, P. Watters, and R. Dazeley, "Automated unsupervised authorship analysis using evidence accumulation clustering," *Natural Language Engineering*, vol. 19, no. 1, pp. 95–120, Nov. 2011, doi: 10.1017/s1351324911000313.

[16] C. Martín-Del-Campo-Rodríguez, G. Sidorov, and I. Batyrshin, "Unsupervised authorship attribution using feature selection and weighted cosine similarity," *Journal of Intelligent & Fuzzy Systems*, vol. 42, no. 5, pp. 4357–4367, Dec. 2021, doi: 10.3233/jifs-219226.

[17] M. T. Zamir, M. A. Ayub, A. Gul, N. Ahmad, and K. Ahmad, "Stylometry Analysis of multi-authored documents for authorship and author style change detection," *arXiv (Cornell University)*, Jan. 2024, doi: 10.48550/arxiv.2401.06752.

[18] L. De Langhe, O. De Clercq, and V. Hoste, "Unsupervised authorship attribution for medieval Latin using Transformer-Based embeddings," *ACL Anthology*, May 01, 2024. https://aclanthology.org/2024.lt4hala-1.8/

[19] A. Lesky, *A History of Greek Literature*, trans. J. Willis and C. de Heer, London, UK: Methuen, 1966, ch. 10, pp. 213–230.

[20] J. Pigeaud, *La maladie de l'âme: Études sur la relation de l'âme et du corps dans la tradition médico-philosophique antique*, Paris, France: Les Belles Lettres, 1981, pp. 45–82.

[21] M. Eder, "Does size matter? Authorship attribution, small samples, big problem," *Digital Scholarship in the Humanities*, vol. 30, no. 2, pp. 167–182, Nov. 2013, doi: 10.1093/llc/fqt066.

[22] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[23] J. Burrows, "Delta: a measure of stylistic difference and a guide to likely authorship," *Literary and Linguistic Computing*, vol. 17, no. 3, pp. 267–287, 2002.

[24] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[25] C. C. Aggarwal, A. Hinneburg and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Proc. 8th Int. Conf. on Database Theory (ICDT)*, London, UK, 2001, pp. 420–434.

[26] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1983.

[27] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[28] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 1990.

[29] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, Jun. 2002, doi: 10.1198/016214502760047131.

[30] J. H. Ward Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[31] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, Portland, OR, USA, 1996, pp. 226–231.

[32] S. Corbara, A. Moreo, and F. Sebastiani, "Syllabic quantity patterns as rhythmic features for Latin authorship attribution," *Journal of the Association for Information Science and Technology*, vol. 74, no. 1, pp. 128–141, May 2022, doi: 10.1002/asi.24660.

[33] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, "Syntactic N-grams as machine learning features for natural language processing," *Expert Systems With Applications*, vol. 41, no. 3, pp. 853–860, Aug. 2013, doi: 10.1016/j.eswa.2013.08.015.

[34] D. W. Prakoso, A. Abdi, and C. Amrit, "Short text similarity measurement methods: a review," *Soft Computing*, vol. 25, no. 6, pp. 4699–4723, Jan. 2021, doi: 10.1007/s00500-020-05479-2.

[35] B. Huang, C. Chen, and K. Shu, "Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges," *ACM SIGKDD Explorations Newsletter*, vol. 26, no. 2, pp. 21–43, Jan. 2025, doi: 10.1145/3715073.3715076.

[36] M. Kocher and J. Savoy, "Distributed language representation for authorship attribution," *Digital Scholarship in the Humanities*, vol. 33, no. 2, pp. 425–441, Aug. 2017, doi: 10.1093/llc/fqx046.

[37] B. Murauer and G. Specht, "DT-Grams: Structured Dependency Grammar Stylometry for Cross-Language Authorship Attribution," *arXiv (Cornell University)*, Jun. 2021, doi: 10.48550/arxiv.2106.05677.

[38] J. Bevendorff *et al.*, “Overview of PAN 2021: Authorship verification, profiling hate speech spreaders on Twitter, and style change detection,” in *Lecture notes in computer science*, 2021, pp. 419–431. doi: 10.1007/978-3-030-85251-1_26.

[39] J. Bevendorff *et al.*, “Overview of PAN 2023: Authorship verification, Multi-Author writing style analysis, profiling cryptocurrency influencers, and trigger detection,” in *Lecture notes in computer science*, 2023, pp. 459–481. doi: 10.1007/978-3-031-42448-9_29.

[40] S. Päpcke, T. Weitin, K. Herget, A. Glawion, and U. Brandes, “Stylometric similarity in literary corpora: Non-authorship clustering and *Deutscher Novellenschatz,*” *Digital Scholarship in the Humanities*, vol. 38, no. 1, pp. 277–295, Aug. 2022, doi: https://doi.org/10.1093/llc/fqac039.

[41] M. Eder, “Visualization in stylometry: Cluster analysis using networks,” *Digital Scholarship in the Humanities*, vol. 32, no. 1, pp. 50–64, Dec. 2015, doi: 10.1093/llc/fqv061.

[42] V. B. Gorman and R. J. Gorman, “Approaching questions of text reusein Ancient Greek using ComputationalSyntactic Stylometry,” *Open Linguistics*, vol. 2, no. 1, Nov. 2016, doi: 10.1515/opli-2016-0026.

[43] Efthimios Gianitsos, T. Bolt, P. Chaudhuri, and J. P. Dexter, “Stylometric Classification of Ancient Greek Literary Texts by Genre,” *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Jan. 2019, doi: https://doi.org/10.18653/v1/w19-2507.

[44] N. I. Tripto and M. E. Ali, “The Word2VeC Graph Model for author Attribution and Genre Detection in Literary analysis,” *arXiv (Cornell University)*, Oct. 2023, doi: 10.48550/arxiv.2310.16972.

[45] M. Oakes and A. Pichler, “Computational Stylometry of Wittgenstein’s ‘Diktat für Schlick’.,” *Bergen Language and Linguistics Studies*, vol. 3, no. 1, Apr. 2013, doi: https://doi.org/10.15845/bells.v3i1.373.

[46] A. Grebennikov, E. Ivanova, M. Koryshev, and M. Solovieva, “Stylometric Methods in Comparative Text Analysis,” *Literature, Language and Computing*, pp. 103–112, 2023, doi: https://doi.org/10.1007/978-981-99-3604-5_9.

[47] S. H. H. Ding, B. C. M. Fung, F. Iqbal and W. K. Cheung, "Learning Stylometric Representations for Authorship Analysis," in IEEE Transactions on Cybernetics, vol. 49, no. 1, pp. 107-121, Jan. 2019, doi: 10.1109/TCYB.2017.2766189.

[48] T. Stanisz, J. Kwapień, and S. Drożdż, “Linguistic data mining with complex networks: A stylometric-oriented approach,” *Information Sciences*, vol. 482, pp. 301–320, May 2019, doi: https://doi.org/10.1016/j.ins.2019.01.040.

[49] U. Stańczyk, “On unsupervised and supervised discretisation in mining stylometric features,” in *Advances in intelligent systems and computing*, 2019, pp. 156–166. doi: 10.1007/978-3-030-31964-9_15.

[50] Z. Ullah and A. Mahmood, “Stylometry of Short Stories through Voyant Corpus Summary Tool: A Text Mining Study,” *Kashmir Journal of Language Research*, vol. 22, no. 1, 2019, Accessed: Jan. 21, 2026. [Online]. Available: https://kjlr.pk/index.php/kjlr/article/view/91

[51] D. Jagtap, S. Ambekar, H. Singh and N. Sharma, "An Approach to Detecting Writing Styles Based on Clustering Technique," *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, Bhopal, India, 2024, pp. 1-7, doi: 10.1109/SCEECS61402.2024.10482055.

ΠΑΡΑΡΤΗΜΑ Α: Κώδικας του pipeline

Στο παρόν παράρτημα παρατίθεται ο πλήρης κώδικας υλοποίησης, όπως αναπτύχθηκε σε περιβάλλον Jupyter Notebook.

In [1]:
```
#Cell 0: Markdown

# Hippocratic Stylometry Project
# Author: <Konstantinos Argiriadis>
```

In [2]:
```
# Cell 1.1 – Imports
import os
os.environ["OMP_NUM_THREADS"] = "1"

import os
from pathlib import Path

import requests
from lxml import etree
```

In [3]:
```
#Cell 1.2 - Directory structure

BASE_DIR = Path("data")
RAW_DIR = BASE_DIR / "raw"
CLEAN_DIR = BASE_DIR / "clean"

from pathlib import Path
import matplotlib.pyplot as plt

FIG_DIR = Path("figures")
FIG_DIR.mkdir(exist_ok=True)

def savefig(name: str, pdf=True, png=True, dpi=300):
    """
    Save current matplotlib figure in /figures as PDF (vector) and/or PNG (raster).
    Call this right BEFORE plt.show().
    """
    path_base = FIG_DIR / name
    if pdf:
        plt.savefig(path_base.with_suffix(".pdf"), bbox_inches="tight")
    if png:
        plt.savefig(path_base.with_suffix(".png"), dpi=dpi, bbox_inches="tight")


RAW_DIR.mkdir(parents=True, exist_ok=True)
CLEAN_DIR.mkdir(parents=True, exist_ok=True)

print("Files ready:")
print(RAW_DIR)
print(CLEAN_DIR)
```

In [5]:
```
#Cell 2.1a - CTS metadata: retrieve all Hippocratic works (Perseus)

CTS_BASE = "https://www.perseus.tufts.edu/hopper/CTS"
AUTHOR_URN = "urn:cts:greekLit:tlg0627"  # Hippocrates

def get_works(author_urn):
    url = f"{CTS_BASE}?request=GetCapabilities&urn={author_urn}"
    r = requests.get(url, timeout=60)
    r.raise_for_status()
    xml = etree.fromstring(r.content)

    # Namespace-agnostic XPath (robust to CTS namespace variations)
    urns = xml.xpath("//*[local-name()='work']/@urn")

    # Keep only Hippocratic URNs (tlg0627)
    urns = [u for u in urns if "tlg0627" in u]
    return sorted(set(urns))

works = get_works(AUTHOR_URN)
print(f"Found {len(works)} Hippocratic CTS work URNs")
works[:15]
```

In [6]:
```python
# Cell 2.1b — Build a map: work_id -> best full CTS URN (prefer perseus-grc1)

def work_id_from_full_urn(full_urn: str) -> str:
    # "urn:cts:greekLit:tlg0627.tlg001.perseus-grc1" -> "tlg0627.tlg001"
    tail = full_urn.split(":")[-1]          # tlg0627.tlg001.perseus-grc1
    parts = tail.split(".")
    return ".".join(parts[:2])              # tlg0627.tlg001

works_full_urns = works  # from 2.1a

WORK_URN_MAP = {}
for u in works_full_urns:
    wid = work_id_from_full_urn(u)
    # prefer a Perseus Greek version if present
    if wid not in WORK_URN_MAP:
        WORK_URN_MAP[wid] = u
    else:
        # if we already have one, replace only if new one is "better"
        current = WORK_URN_MAP[wid]
        if ("perseus-grc" in u) and ("perseus-grc" not in current):
            WORK_URN_MAP[wid] = u
        # optional: prefer grc1 over others
        if ("perseus-grc1" in u) and ("perseus-grc1" not in current):
            WORK_URN_MAP[wid] = u

HIPPOCRATIC_WORK_IDS = sorted(WORK_URN_MAP.keys())

print(f"Work IDs: {len(HIPPOCRATIC_WORK_IDS)}")
print("Example work_id -> full urn:")
for wid in HIPPOCRATIC_WORK_IDS[:10]:
    print(wid, "->", WORK_URN_MAP[wid])
```

In [7]:
```python
# Cell 2.1c — Download one work as plain text via CTS using the FULL URN

CTS_BASE = "https://www.perseus.tufts.edu/hopper/CTS"

def download_cts_plain_text(full_urn: str) -> str | None:
    """
    Tries to fetch the whole work using GetPassage with the urn only.
    If it fails, returns None.
    """
    url = f"{CTS_BASE}?request=GetPassage&urn={full_urn}"
    r = requests.get(url, timeout=90)
    if r.status_code != 200:
        return None

    # Perseus returns XML/TEI-like content; we extract text content.
    try:
        xml = etree.fromstring(r.content)
        text = " ".join(xml.xpath("//text()"))
        text = " ".join(text.split())
        return text if text.strip() else None
    except Exception:
        return None
```

In [9]:
```python
# Cell 2.3 — Safe saving with manifest (no duplicates)

import hashlib
import csv

MANIFEST = BASE_DIR / "manifest.csv"

def sha1_text(s: str) -> str:
    return hashlib.sha1(s.encode("utf-8")).hexdigest()

def load_manifest(path: Path) -> dict:
    if not path.exists():
        return {}
    rows = {}
    with open(path, "r", encoding="utf-8", newline="") as f:
        reader = csv.DictReader(f)
        for r in reader:
            rows[r["work_id"]] = r
    return rows

def append_manifest_row(path: Path, row: dict):
    file_exists = path.exists()
    with open(path, "a", encoding="utf-8", newline="") as f:
        fieldnames = ["work_id", "full_urn", "filepath", "sha1"]
        writer = csv.DictWriter(f, fieldnames=fieldnames)
        if not file_exists:
            writer.writeheader()
        writer.writerow(row)

manifest = load_manifest(MANIFEST)

downloaded = 0
skipped = 0
updated = 0
```

```python
for wid in HIPPOCRATIC_WORK_IDS:
    full_urn = WORK_URN_MAP[wid]
    print(f"Test: {wid} | URN: {full_urn}")

    text = download_cts_plain_text(full_urn)
    if not text:
        print("X Not found / empty")
        continue

    h = sha1_text(text)
    filepath = RAW_DIR / f"{wid}.txt"

    # If we already downloaded this work_id and hash matches, skip
    if wid in manifest and manifest[wid]["sha1"] == h and Path(manifest[wid]["filepath"]).exists():
        print("↷ Skip (already downloaded, same hash)")
        skipped += 1
        continue

    # If file exists but different content, keep both (safe)
    if filepath.exists():
        # save as wid__v2.txt (or v3, etc.)
        i = 2
        while True:
            alt = RAW_DIR / f"{wid}__v{i}.txt"
            if not alt.exists():
                filepath = alt
                break
            i += 1
        updated += 1

    with open(filepath, "w", encoding="utf-8") as f:
        f.write(text)

    append_manifest_row(MANIFEST, {
        "work_id": wid,
        "full_urn": full_urn,
        "filepath": str(filepath),
        "sha1": h
    })
    downloaded += 1
    print(f"✓ Saved: {filepath.name}")

print("\n--- Summary ---")
print("Downloaded:", downloaded)
print("Skipped:", skipped)
print("Saved as new version:", updated)
print("Manifest:", MANIFEST)
print("Files in raw:", len(list(RAW_DIR.glob('*.txt'))))
```

In [10]:
```
# Cell 2.4 — Titles (work metadata) from CTS GetCapabilities

import re
import csv
from datetime import datetime

CTS_BASE = "https://www.perseus.tufts.edu/hopper/CTS"
AUTHOR_URN = "urn:cts:greekLit:tlg0627"  # Hippocrates

def work_id_from_full_urn(full_urn: str) -> str:
    # "urn:cts:greekLit:tlg0627.tlg001.perseus-grc1" -> "tlg0627.tlg001"
    tail = full_urn.split(":")[-1]                # tlg0627.tlg001.perseus-grc1
    parts = tail.split(".")
    return ".".join(parts[:2])                    # tlg0627.tlg001

def fetch_capabilities_xml(author_urn: str) -> bytes:
    url = f"{CTS_BASE}?request=GetCapabilities&urn={author_urn}"
    r = requests.get(url, timeout=90)
    r.raise_for_status()
    return r.content

def extract_work_titles_from_capabilities(xml_bytes: bytes) -> dict:
    """
    Returns dict: {work_id: title}
    """
    root = etree.fromstring(xml_bytes)

    # Every <work> node (namespace-agnostic)
    work_nodes = root.xpath("//*[local-name()='work']")

    titles = {}
    for w in work_nodes:
        urn = w.get("urn") or ""
        if "tlg0627" not in urn:
            continue

        wid = work_id_from_full_urn(urn)

        # title: usually <title> or <title xml:lang="...">
        title_nodes = w.xpath(".//*[local-name()='title']")
        if title_nodes:
            title = " ".join(" ".join(title_nodes[0].itertext()).split())
        else:
            # fallback: case where there is <label> / <groupname> (rare)
            title = ""

        # keep the first "good" title
        if wid not in titles and title:
            titles[wid] = title

    return titles

cap_xml = fetch_capabilities_xml(AUTHOR_URN)
WORK_TITLES = extract_work_titles_from_capabilities(cap_xml)

print("Titles found:", len(WORK_TITLES))
# Sample for my 20 IDs:
for wid in HIPPOCRATIC_WORK_IDS:
    print(wid, "->", WORK_TITLES.get(wid, "(no title found)"))
```

In [11]:
```python
# Cell 2.5 — Save metadata.csv (+ manifest update)

METADATA = BASE_DIR / "metadata.csv"

rows = []
now = datetime.utcnow().isoformat(timespec="seconds") + "Z"

for wid in HIPPOCRATIC_WORK_IDS:
    full_urn = WORK_URN_MAP.get(wid, "")
    title = WORK_TITLES.get(wid, "")
    rows.append({
        "work_id": wid,
        "title": title,
        "full_urn": full_urn,
        "retrieved_at": now
    })

with open(METADATA, "w", encoding="utf-8", newline="") as f:
    w = csv.DictWriter(f, fieldnames=["work_id", "title", "full_urn", "retrieved_at"])
    w.writeheader()
    w.writerows(rows)

print("Wrote:", METADATA)
print("Preview:")
rows[:5]

# add/update 'title' column in manifest.csv

MANIFEST = BASE_DIR / "manifest.csv"

if MANIFEST.exists():
    import pandas as pd

    df = pd.read_csv(MANIFEST)
    if "title" not in df.columns:
        df["title"] = ""

    # map by work_id
    df["title"] = df["work_id"].map(lambda x: WORK_TITLES.get(x, ""))

    df.to_csv(MANIFEST, index=False, encoding="utf-8")
    print("Updated manifest with titles:", MANIFEST)
else:
    print("manifest.csv not found, skipped.")
```

In [12]:
```
# Cell 3.1 — Cleaning imports

import re
```

In [13]:
```
# Cell 3.2 — Cleaner function

CTS_GARBAGE_LINE_PREFIXES = (
    "Invalid URN reference:",
    "Cite requested:",
    "Passage requested:",
    "Work requested:",
    "XSLT found",
)

def clean_perseus_cts_text(raw: str) -> str:
    """
    Cleans Perseus/CTS plain-text output:
    - Removes CTS error/header lines (Invalid URN..., Cite requested..., etc.)
    - Normalizes whitespace
    - Keeps Greek diacritics, punctuation, capitalization intact
    """
    lines = raw.splitlines()

    kept = []
    for line in lines:
        s = line.strip()

        # drop empty lines early (we'll re-normalize later anyway)
        if not s:
            continue

        # drop known CTS garbage header lines
        if any(s.startswith(p) for p in CTS_GARBAGE_LINE_PREFIXES):
            continue

        # some CTS headers are long and contain multiple phrases in one line
        # keep it safe: if it contains these key phrases, drop it
        if ("Invalid URN reference" in s) or ("Cite requested" in s) or ("Passage requested" in s) or ("Work requested" in s):
            continue

        kept.append(s)

    text = "\n".join(kept)

    # Normalize whitespace:
    # - Collapse any whitespace runs to single space inside lines
    # - Keep paragraph breaks as single newline
    text = re.sub(r"[ \t\u00A0]+", " ", text)        # spaces/tabs/nbsp -> single space
    text = re.sub(r"\n{2,}", "\n", text)             # multiple newlines -> single newline
    text = text.strip()

    return text
```

In [14]:
```
# Cell 3.3 — Clean all raw files into clean/

def clean_corpus(raw_dir: Path, clean_dir: Path, overwrite: bool = True):
    raw_files = sorted(raw_dir.glob("*.txt"))
    if not raw_files:
        print("No raw .txt files found in", raw_dir)
        return

    written = 0
    skipped_empty = 0
    for fp in raw_files:
        raw = fp.read_text(encoding="utf-8", errors="replace")
        clean = clean_perseus_cts_text(raw)

        out = clean_dir / fp.name  # same filename
        if (not overwrite) and out.exists():
            continue

        if not clean:
            skipped_empty += 1
            # still write an empty file? usually better NOT to
            # out.write_text("", encoding="utf-8")
            print(f"⚠️ Empty after cleaning: {fp.name} (skipped)")
            continue

        out.write_text(clean, encoding="utf-8")
        written += 1

    print("---- Cleaning summary ----")
    print("Raw files:", len(raw_files))
    print("Written clean files:", written)
    print("Skipped (became empty):", skipped_empty)
    print("Clean dir now has:", len(list(clean_dir.glob('*.txt'))))

clean_corpus(RAW_DIR, CLEAN_DIR, overwrite=True)
```

In [15]:
```python
# Cell 3.4 — Debug one file through cleaning steps

sample_path = sorted(RAW_DIR.glob("*.txt"))[0]
raw = sample_path.read_text(encoding="utf-8", errors="ignore")

print("SAMPLE FILE:", sample_path.name)
print("RAW length:", len(raw))
print("RAW preview:\n", raw[:600])
```

In [16]:
```python
# Cell 3.5 — Cleaning function (polytonic-safe, fixes "Invalid URN..." inline)

# Greek ranges: monotonic + polytonic (Greek Extended)
GREEK_CHAR = r"[\u0370-\u03FF\u1F00-\u1FFF]"

def strip_until_greek(s: str) -> str:
    """
    If there is junk/metadata before the first Greek character, drop it.
    Works even if everything is on ONE line.
    """
    return re.sub(rf"^.*?(?={GREEK_CHAR})", "", s, flags=re.DOTALL)

def clean_hippocratic_text(text: str) -> str:
    text = text.replace("\ufeff", "")  # BOM

    # 1) If CTS error/metadata is prepended, drop everything until Greek starts
    if "Invalid URN reference:" in text[:400] or "Cite requested:" in text[:400]:
        text = strip_until_greek(text)

    # 2) Sometimes CTS phrases appear inline again; remove them safely (only up to next Greek char)
    #    (If no Greek follows, it will remove to end — but that's fine.)
    for key in ["Invalid URN reference:", "Cite requested:", "Passage requested:", "Work requested:"]:
        text = re.sub(rf"{re.escape(key)}.*?(?={GREEK_CHAR}|$)", "", text, flags=re.DOTALL)

    # 3) Whitespace normalize (keep Greek + polytonic)
    text = re.sub(r"[ \t]+", " ", text)
    text = re.sub(r"\n{3,}", "\n\n", text)

    return text.strip()
```

In [17]:
```python
# Cell 3.6 — Test cleaning on sample

cleaned = clean_hippocratic_text(raw)

print("CLEANED length:", len(cleaned))
print("CLEANED preview:\n", cleaned[:600])
```

In [18]:
```python
# Cell 3.7 — Clean whole corpus into CLEAN_DIR

def clean_corpus(raw_dir: Path, clean_dir: Path, overwrite: bool = True):
    clean_dir.mkdir(parents=True, exist_ok=True)

    raw_files = sorted(raw_dir.glob("*.txt"))
    written = 0
    skipped_empty = 0

    for p in raw_files:
        out = clean_dir / p.name

        if out.exists() and not overwrite:
            continue

        txt = p.read_text(encoding="utf-8", errors="ignore")
        txt2 = clean_hippocratic_text(txt)

        if not txt2.strip():
            print(f"⚠️ Empty after cleaning: {p.name} (skipped)")
            skipped_empty += 1
            continue

        out.write_text(txt2, encoding="utf-8")
        written += 1

    print("\n---- Cleaning summary ----")
    print("Raw files:", len(raw_files))
    print("Written clean files:", written)
    print("Skipped (became empty):", skipped_empty)
    print("Clean dir now has:", len(list(clean_dir.glob("*.txt"))))

clean_corpus(RAW_DIR, CLEAN_DIR, overwrite=True)
```

In [19]:
```python
# Cell X — Corpus overview table (title + counts + diagnostics)

import re
import pandas as pd
from pathlib import Path

BASE_DIR = Path("data")
RAW_DIR = BASE_DIR / "raw"
CLEAN_DIR = BASE_DIR / "clean"
MANIFEST = BASE_DIR / "manifest.csv"

GREEK_RE = re.compile(r"[\u0370-\u03FF\u1F00-\u1FFF]")  # Greek + Greek Extended

def greek_ratio(text: str) -> float:
    if not text:
        return 0.0
    greek = len(GREEK_RE.findall(text))
    return greek / max(len(text), 1)

def basic_stats(text: str) -> dict:
    text = text or ""
    # simple tokenization (whitespace). For stylometry we'll do a more controlled one later.
    tokens = [t for t in re.split(r"\s+", text.strip()) if t]
    words = [re.sub(r"[^\w\u0370-\u03FF\u1F00-\u1FFF]+", "", t) for t in tokens]
    words = [w for w in words if w]

    wc = len(words)
    uniq = len(set(words))
    ttr = (uniq / wc) if wc else 0.0
    avg_wlen = (sum(len(w) for w in words) / wc) if wc else 0.0

    return {
        "word_count": wc,
        "char_count": len(text),
        "unique_words": uniq,
        "ttr": ttr,
        "avg_word_len": avg_wlen,
        "greek_char_ratio": greek_ratio(text),
    }
```

```python
# Load manifest (expects at least: work_id, filepath, title)
dfm = pd.read_csv(MANIFEST) if MANIFEST.exists() else pd.DataFrame()

rows = []
for _, r in dfm.iterrows():
    work_id = r.get("work_id", "")
    title = r.get("title", "")
    # prefer clean version if exists, else raw
    clean_path = CLEAN_DIR / f"{work_id}.txt"
    raw_path = RAW_DIR / f"{work_id}.txt"

    path_used = None
    if clean_path.exists():
        path_used = clean_path
    elif raw_path.exists():
        path_used = raw_path
    else:
        # maybe you saved versions like _v2; fall back to any matching
        candidates = sorted(list(CLEAN_DIR.glob(f"{work_id}*.txt"))) + sorted(list(RAW_DIR.glob(f"{work_id}*.txt")))
        path_used = candidates[0] if candidates else None

    if path_used and path_used.exists():
        text = path_used.read_text(encoding="utf-8", errors="ignore")
        stats = basic_stats(text)
        rows.append({
            "work_id": work_id,
            "title": title,
            "file": str(path_used),
            **stats
        })
    else:
        rows.append({
            "work_id": work_id,
            "title": title,
            "file": "",
            "word_count": 0,
            "char_count": 0,
            "unique_words": 0,
            "ttr": 0.0,
            "avg_word_len": 0.0,
            "greek_char_ratio": 0.0,
        })

df = pd.DataFrame(rows)

# Helpful flags for “limited material” discussion
df["very_short"] = df["word_count"] < 500
df["likely_not_greek"] = df["greek_char_ratio"] < 0.20  # catches english/metadata-heavy texts

df = df.sort_values(["word_count"], ascending=False).reset_index(drop=True)
display(df)
print("\nSummary:")
print("Docs:", len(df))
print("Total words:", int(df["word_count"].sum()))
print("Very short (<500 words):", int(df["very_short"].sum()))
print("Likely not Greek (greek_char_ratio < 0.20):", int(df["likely_not_greek"].sum()))
```

In [20]:
```
#Cell 4.1 - Load clean corpus

import pandas as pd

CLEAN_DIR = Path("data/clean")

texts = {}
for p in sorted(CLEAN_DIR.glob("*.txt")):
    texts[p.stem] = p.read_text(encoding="utf-8", errors="ignore")

doc_ids = list(texts.keys())
documents = list(texts.values())

print(f"Loaded {len(documents)} documents")
```

```
Loaded 18 documents
```

In [21]:
```
# Cell 4.2 - n-gram feature space

from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(
    analyzer="char",
    ngram_range=(3,5),
    min_df=2,
    max_df=0.9,
    norm="l2"
)

X = vectorizer.fit_transform(documents)
print("TF-IDF matrix shape:", X.shape)
```

```
TF-IDF matrix shape: (18, 9173)
```

In [22]:
```
#Cell 4.3 - cosine similarity (doc-doc)

from sklearn.metrics.pairwise import cosine_similarity
import numpy as np

cos_sim = cosine_similarity(X)
cos_dist = 1 - cos_sim

print("Cosine distance matrix:", cos_dist.shape)
```

```
Cosine distance matrix: (18, 18)
```

In [23]:
```
#Cell 4.4 - Hierarchical clustering

from scipy.cluster.hierarchy import linkage

Z = linkage(cos_dist, method="average")
```

In [24]:
```
#Create dict
df_meta = pd.read_csv("data/metadata.csv")

print(df_meta.columns)

id2title = dict(zip(df_meta["work_id"], df_meta["title"]))
```

```
Index(['work_id', 'title', 'full_urn', 'retrieved_at'], dtype='object')
```

In [25]:
```
#Cell 4.5 - Dendrogram

import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram

plt.figure(figsize=(12, 6))
dendrogram(
    Z,
    labels=[id2title[i] for i in doc_ids],
    leaf_rotation=90
)
plt.title("Hierarchical clustering (char n-grams, cosine)")
plt.tight_layout()
savefig("char_ngrams_hier_dendrogram")
plt.show()
```

In [26]:
```
#Cell 4.6 - Heatmap (doc-doc similarity)

import seaborn as sns

plt.figure(figsize=(8, 6))
sns.heatmap(
    cos_sim,
    xticklabels=doc_ids,
    yticklabels=doc_ids,
    cmap="viridis"
)
plt.title("Cosine similarity (char n-grams)")
plt.tight_layout()
plt.show()
```

In [27]:
```
#Cell 4.7 - MDS/PCA

from sklearn.manifold import MDS
import matplotlib.pyplot as plt

mds = MDS(
    n_components=2,
    dissimilarity="precomputed",
    random_state=42
)

coords = mds.fit_transform(cos_dist)

plt.figure(figsize=(7, 6))
plt.scatter(coords[:, 0], coords[:, 1])

for i, doc_id in enumerate(doc_ids):
    plt.text(
        coords[i, 0],
        coords[i, 1],
        id2title.get(doc_id, doc_id),
        fontsize=8
    )

plt.title("MDS projection (char n-grams, cosine distance)")
plt.tight_layout()
plt.show()
```

In [28]:
```python
# Cell 5.1 - Load cleaned corpus (docs) + basic ids (EAC for n-grams)

from pathlib import Path

CLEAN_DIR = Path("data") / "clean"

paths = sorted(CLEAN_DIR.glob("*.txt"))
doc_ids = [p.stem for p in paths]      # e.g. tlg0627.tlg001
docs = [p.read_text(encoding="utf-8", errors="ignore") for p in paths]

print("Docs:", len(docs))
print("Example:", doc_ids[0], "chars:", len(docs[0]))
```

```
Docs: 18
Example: tlg0627.tlg001 chars: 1256
```

In [29]:
```python
# Cell 5.2 - Feature extraction: char n-grams TF-IDF

from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(
    analyzer="char",
    ngram_range=(3, 5),
    min_df=2,           # cut super rare n-grams
    max_df=0.95,        # cut almost similar n-grams
    sublinear_tf=True, # log(1+tf)
    norm="l2"
)

X = vectorizer.fit_transform(docs)
print("X shape:", X.shape)  # (n_docs, n_features)
```

```
X shape: (18, 9184)
```

In [30]:
```python
# Cell 5.3 - EAC core: build co-association matrix from many k-means runs

import numpy as np
from sklearn.cluster import KMeans

def eac_coassociation(
    X,
    k_values=range(2, 9),    # try k=2..8
    n_runs=300,              # can be 500 or 1000
    init="k-means++",
    n_init=10,
    random_seed=42
):
    rng = np.random.default_rng(random_seed)
    n = X.shape[0]
    C = np.zeros((n, n), dtype=np.float64)

    total = 0
    for r in range(n_runs):
        k = int(rng.choice(list(k_values)))
        seed = int(rng.integers(0, 1_000_000))

        km = KMeans(
            n_clusters=k,
            init=init,
            n_init=n_init,
            random_state=seed
        )
        labels = km.fit_predict(X)

        # update co-association: for each cluster, add 1 to all pairs inside it
        for cl in np.unique(labels):
            idx = np.where(labels == cl)[0]
            C[np.ix_(idx, idx)] += 1.0

        total += 1

    # normalize to [0,1]
    C /= total
```

```
    # make sure diagonal is 1
    np.fill_diagonal(C, 1.0)

    return C

C = eac_coassociation(X, k_values=range(2, 9), n_runs=300)
print("Co-association matrix shape:", C.shape)
print("C min/max:", float(C.min()), float(C.max()))
```

```
Co-association matrix shape: (18, 18)
C min/max: 0.013333333333333334 1.0
```

In [31]:
```
# Cell 5.4 - Final clustering on EAC distance + dendrogram

import matplotlib.pyplot as plt
from scipy.spatial.distance import squareform
from scipy.cluster.hierarchy import linkage, dendrogram

D = 1.0 - C  # distance

# linkage needs condensed distance vector
D_condensed = squareform(D, checks=False)

Z = linkage(D_condensed, method="average")  # average/complete is ok for these distances

plt.figure(figsize=(12, 6))
dendrogram(Z, labels=[id2title[i] for i in doc_ids], leaf_rotation=90)
plt.title("EAC dendrogram (char n-grams TF-IDF)")
plt.tight_layout()
savefig("char_ngrams_eac_dendrogram")
plt.show()
```

In [32]:
```
# Cell 5.5 - MDS projection for EAC distance

import matplotlib.pyplot as plt
from sklearn.manifold import MDS

# 1) EAC distance matrix
# already computed C = eac_coassociation(X, ...) earlier)
D_eac = 1.0 - C   # shape (n_docs, n_docs)

# 2) MDS in 2D using precomputed dissimilarities
mds = MDS(
    n_components=2,
    dissimilarity="precomputed",
    random_state=42
)

coords = mds.fit_transform(D_eac)

# 3) Plot
plt.figure(figsize=(7, 6))
plt.scatter(coords[:, 0], coords[:, 1])

for i, label in enumerate(doc_ids):
    plt.text(coords[i, 0], coords[i, 1], id2title[label], fontsize=8)

plt.title("MDS projection (EAC distance: 1 - co-association)")
plt.tight_layout()
savefig("char_ngrams_mds_cosine")
plt.show()
```

In [33]:
```
# Cell 5.6 - Top most similar pairs according to EAC

pairs = []
n = C.shape[0]
for i in range(n):
    for j in range(i+1, n):
        pairs.append((C[i, j], doc_ids[i], doc_ids[j]))

pairs.sort(reverse=True, key=lambda x: x[0])

print("Top 15 pairs by co-association (higher = more often clustered together):")
for score, a, b in pairs[:15]:
    print(f"{score:.3f}  {id2title[a]}  <->  {id2title[b]}")
```

In [35]:
```python
# Cell 6.1 - Imports + paths (function words / MFW pipeline)

from pathlib import Path
import re
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# assuming you already have these (adjust if needed)
BASE_DIR = Path("data")
CLEAN_DIR = BASE_DIR / "clean"   # where your 18 clean greek files are
# If you kept raw too:
RAW_DIR = BASE_DIR / "raw"
```

In [36]:
```python
# Cell 6.2 - Load docs (from CLEAN_DIR) + basic tokenization (Greek-aware)

GREEK_WORD_RE = re.compile(r"[Ͱ-ϡἀ-ᾼ῀-ῌῐ-ΐῠ-Ῥῲ-ῼ]+", re.UNICODE)  # broad Greek ranges incl polytonic

def load_docs_from_dir(text_dir: Path):
    paths = sorted(text_dir.glob("*.txt"))
    doc_ids = [p.stem for p in paths]
    docs = [p.read_text(encoding="utf-8", errors="ignore") for p in paths]
    return doc_ids, docs, paths

def word_tokenize_greek(text: str):
    # Lowercase keeps "style" signal; also helps reduce false proper-noun detection if edition uses caps
    text = text.replace("\ufeff", "").lower()
    return GREEK_WORD_RE.findall(text)

doc_ids, docs, paths = load_docs_from_dir(CLEAN_DIR)
docs_tokens = [word_tokenize_greek(t) for t in docs]
doc_lengths = np.array([len(toks) for toks in docs_tokens], dtype=int)

print("Docs:", len(doc_ids))
print("Min/Max words:", int(doc_lengths.min()), int(doc_lengths.max()))
print("Example doc:", doc_ids[0], "words:", int(doc_lengths[0]))
```

```
Docs: 18
Min/Max words: 30 1137
Example doc: tlg0627.tlg001 words: 217
```

In [37]:
```python
# Cell 6.3 - Build MFW list (Most Frequent Words) data-driven from whole corpus

from collections import Counter

all_tokens = [tok for toks in docs_tokens for tok in toks]
freq = Counter(all_tokens)

# Hyperparameters you can tweak:
TOP_N_MFW = 250          # size of function-words-ish vocabulary (try 150–500)
MIN_GLOBAL_FREQ = 5      # drop ultra-rare tokens

mfw_all = [w for (w, c) in freq.most_common() if c >= MIN_GLOBAL_FREQ][:TOP_N_MFW]

print("Total unique tokens:", len(freq))
print("MFW list size:", len(mfw_all))
print("Top 30 MFW:", mfw_all[:30])
```

```
Total unique tokens: 1941
MFW list size: 116
Top 30 MFW: ['καὶ', 'δὲ', 'τὸ', 'τῶν', 'τοῦ', 'τε', 'τὰ', 'γὰρ', 'μὲν', 'ἢ', 'ἐς', 'ἐν', 'τὴν', 'τῆς', 'ἡ', 'ἂν', 'οἱ', 'ὡς', 'τι', 'ὑπὸ', 'μὴ', 'τῷ', 'πρὸς', 'ἐκ', 'περὶ', 'οὐ', 'εἶναι', 'τοῖσι', 'τὰς', 'εἰ']
```

In [38]:
```
# Cell 6.4 - Heuristic filter for "likely proper nouns" (optional)
# NOTE: In many Ancient Greek editions everything is lowercase -> this heuristic may remove almost nothing.
# We'll use two heuristics:
#   A) If original text had capitals: token starts with Greek uppercase
#   B) Token is "spiky": appears concentrated in very few docs (topic-like) -> likely content word / name

GREEK_UPPER_RE = re.compile(r"^[A-ΩϚϜ]")  # rough: Greek uppercase letters (incl archaic forms)

def tokens_from_original_case(text: str):
    # keep original case for detecting uppercase-initial tokens
    text = text.replace("\ufeff", "")
    return GREEK_WORD_RE.findall(text)

docs_tokens_case = [tokens_from_original_case(t) for t in docs]  # same docs but with original case
case_vocab_caps = set()
for toks in docs_tokens_case:
    for tok in toks:
        if GREEK_UPPER_RE.match(tok):
            case_vocab_caps.add(tok.lower())  # map to lowercase form used in features

# Build doc frequency (in how many docs a token appears)
doc_sets = [set(toks) for toks in docs_tokens]
df_counts = Counter()
for s in doc_sets:
    for tok in s:
        df_counts[tok] += 1

def is_spiky(tok: str, min_df: int = 3):
    # if appears in very few docs, it's more content-like (names/technical terms)
    return df_counts.get(tok, 0) < min_df
```

```
def filter_mfw_remove_properish(mfw_list, remove_caps=True, remove_spiky=True, min_df=3):
    out = []
    for w in mfw_list:
        if remove_caps and w in case_vocab_caps:
            continue
        if remove_spiky and is_spiky(w, min_df=min_df):
            continue
        out.append(w)
    return out

mfw_no_properish = filter_mfw_remove_properish(
    mfw_all,
    remove_caps=True,
    remove_spiky=True,
    min_df=3
)

print("MFW with possible proper nouns:", len(mfw_all))
print("MFW without 'properish' (caps + spiky):", len(mfw_no_properish))
print("Removed:", len(set(mfw_all) - set(mfw_no_properish)))
print("Top 30 filtered:", mfw_no_properish[:30])
```

```
MFW with possible proper nouns: 116
MFW without 'properish' (caps + spiky): 102
Removed: 14
Top 30 filtered: ['καὶ', 'δὲ', 'τὸ', 'τῶν', 'τοῦ', 'τε', 'τὰ', 'γὰρ', 'μὲν', 'ἢ', 'ἐς', 'ἐν', 'τὴν', 'τῆς', 'ἡ', 'ἂν', 'οἱ', 'ὡς', 'τι', 'ὑπὸ', 'μὴ', 'τῷ', 'πρὸς', 'ἐκ',
'περὶ', 'οὐ', 'εἶναι', 'τοῖσι', 'τάς', 'εἰ']
```

In [39]:
```
# Cell 6.5 - Build feature matrix for MFWs: relative frequencies (doc x vocab)

def build_relfreq_matrix(docs_tokens, vocab):
    vocab_index = {w:i for i,w in enumerate(vocab)}
    X = np.zeros((len(docs_tokens), len(vocab)), dtype=float)

    for i, toks in enumerate(docs_tokens):
        if len(toks) == 0:
            continue
        counts = Counter(toks)
        for w, j in vocab_index.items():
            X[i, j] = counts.get(w, 0)

    # relative frequencies (normalize by document length)
    lengths = np.array([len(t) for t in docs_tokens], dtype=float)
    lengths[lengths == 0] = 1.0
    X = X / lengths[:, None]
    return X

X_mfw_with = build_relfreq_matrix(docs_tokens, mfw_all)
X_mfw_without = build_relfreq_matrix(docs_tokens, mfw_no_properish)

print("X_mfw_with shape:", X_mfw_with.shape)
print("X_mfw_without shape:", X_mfw_without.shape)
```

```
X_mfw_with shape: (18, 116)
X_mfw_without shape: (18, 102)
```

In [40]:
```
# Cell 6.6 - Run EAC for function-words features (WITH possible proper nouns)

C_with = eac_coassociation(X_mfw_with, k_values=range(2, 9), n_runs=300, random_seed=42)
D_with = 1.0 - C_with

print("Co-association (with) shape:", C_with.shape, "min/max:", float(C_with.min()), float(C_with.max()))
```

```
Co-association (with) shape: (18, 18) min/max: 0.0 1.0
```

In [41]:
```
# Cell 6.7 - Dendrogram (WITH possible proper nouns)

from scipy.spatial.distance import squareform
from scipy.cluster.hierarchy import linkage, dendrogram

Z_with = linkage(squareform(D_with, checks=False), method="average")

plt.figure(figsize=(12, 6))
dendrogram(Z_with, labels=[id2title[i] for i in doc_ids], leaf_rotation=90)
plt.title("EAC dendrogram (MFWs - WITH possible proper nouns)")
plt.tight_layout()
savefig("mfw_with_eac_dendrogram")
plt.show()
```

In [42]:
```
# Cell 6.8 - Top similar pairs (WITH possible proper nouns)

pairs = []
n = C_with.shape[0]
for i in range(n):
    for j in range(i+1, n):
        pairs.append((C_with[i, j], doc_ids[i], doc_ids[j]))

pairs.sort(reverse=True, key=lambda x: x[0])

print("Top 15 pairs by co-association (WITH):")
for score, a, b in pairs[:15]:
    print(f"{score:.3f}  {id2title[a]}  <->  {id2title[b]}")
```

In [43]:
```
# Cell 6.9 - Run EAC for function-words features (WITHOUT 'properish')

C_without = eac_coassociation(X_mfw_without, k_values=range(2, 9), n_runs=300, random_seed=42)
D_without = 1.0 - C_without

print("Co-association (without) shape:", C_without.shape, "min/max:", float(C_without.min()), float(C_without.max()))
```

```
Co-association (without) shape: (18, 18) min/max: 0.0 1.0
```

In [44]:
```
# Cell 6.10 - Dendrogram (WITHOUT 'properish')

Z_without = linkage(squareform(D_without, checks=False), method="average")

plt.figure(figsize=(12, 6))
dendrogram(Z_without, labels=[id2title[i] for i in doc_ids], leaf_rotation=90)
plt.title("EAC dendrogram (MFWs - WITHOUT 'properish')")
plt.tight_layout()
savefig("mfw_without_eac_dendrogram")
plt.show()
```

In [45]:
```
# Cell 6.11 - Top similar pairs (WITHOUT 'properish')

pairs = []
n = C_without.shape[0]
for i in range(n):
    for j in range(i+1, n):
        pairs.append((C_without[i, j], doc_ids[i], doc_ids[j]))

pairs.sort(reverse=True, key=lambda x: x[0])

print("Top 15 pairs by co-association (WITHOUT):")
for score, a, b in pairs[:15]:
    print(f"{score:.3f}  {id2title[a]}  <->  {id2title[b]}")
```

In [46]:
```
# Cell 6.12 - Compare "WITH vs WITHOUT" )

# correlation between the two co-association matrices (upper triangle)
def upper_triangle_values(M):
    vals = []
    n = M.shape[0]
    for i in range(n):
        for j in range(i+1, n):
            vals.append(M[i, j])
    return np.array(vals, dtype=float)

v_with = upper_triangle_values(C_with)
v_without = upper_triangle_values(C_without)

corr = np.corrcoef(v_with, v_without)[0, 1]
print("Correlation of co-association (WITH vs WITHOUT):", float(corr))
```

```
Correlation of co-association (WITH vs WITHOUT): 0.9944357601102484
```

In [48]:
```
# Cell 6.13b — MDS for EAC (MFW WITH / WITHOUT) — 2 figures total, saved

import numpy as np
import matplotlib.pyplot as plt
from sklearn.manifold import MDS

# clean old figures
plt.close("all")

def plot_mds_from_C(Cmat, labels, title, fname, random_state=42):
    # EAC distance
    D = 1.0 - Cmat

    mds = MDS(
        n_components=2,
        dissimilarity="precomputed",
        random_state=random_state
    )
    coords = mds.fit_transform(D)

    fig = plt.figure(figsize=(8, 6))
    plt.scatter(coords[:, 0], coords[:, 1])

    for i, lab in enumerate(labels):
        plt.text(coords[i, 0], coords[i, 1], lab, fontsize=8)

    plt.title(title)
    plt.tight_layout()

    savefig(fname)

    plt.show()
    plt.close(fig)
```

```
labels = [id2title.get(doc_id, doc_id) for doc_id in doc_ids]

plot_mds_from_C(
    C_with,
    labels,
    "MDS (EAC distance) — MFW WITH possible proper nouns",
    "mfw_with_eac_mds"
)

plot_mds_from_C(
    C_without,
    labels,
    "MDS (EAC distance) — MFW WITHOUT 'properish'",
    "mfw_without_eac_mds"
)
```

In [49]:
```
#Cell 6.14
import numpy as np

def upper_triangle_values(M):
    vals = []
    n = M.shape[0]
    for i in range(n):
        for j in range(i+1, n):
            vals.append(M[i, j])
    return np.array(vals, dtype=float)

# C from char-n-grams EAC
v_char = upper_triangle_values(C)

# pick one of the MFW versions:
v_mfw_with = upper_triangle_values(C_with)
v_mfw_without = upper_triangle_values(C_without)

corr_char_vs_mfw_with = np.corrcoef(v_char, v_mfw_with)[0, 1]
corr_char_vs_mfw_without = np.corrcoef(v_char, v_mfw_without)[0, 1]

print("Corr(EAC co-assoc): char vs MFW_WITH     =", float(corr_char_vs_mfw_with))
print("Corr(EAC co-assoc): char vs MFW_WITHOUT  =", float(corr_char_vs_mfw_without))
```

```
Corr(EAC co-assoc): char vs MFW_WITH     = 0.4597819422681269
Corr(EAC co-assoc): char vs MFW_WITHOUT  = 0.4382898827489151
```