Εθνικό Μετσόβιο Πολυτεχνείο

Δ.Π.Μ.Σ.'Μαθηματική Προτυποποίηση στις Σύγχρονες Τεχνολογίες και την Οικονομία'

Εντροπική Ανάλυση Φυσικής Γλώσσας

Μεταπτυχιακή εργασία
Μαρία Καλημέρη

Επιβλέπων Καθηγητής:
Ανδρέας Σταφυλοπάτης
...

Σεπτέμβριος 5, 2011

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή, στο πλαίσιο της μαθηματικής και υπολογιστικής μελέτης της φυσικής γλώσσας, διεξάγουμε και υλοποιούμε μία μεθοδολογία εντροπικής ανάλυσης γραπτών κειμένων σε μία αναπαράσταση χρονοσειρών μήκους λέξεων. Η κλασσική εντροπία του Shannon και η γενίκευση αυτής με την μορφή των $n-gram$ εντροπιών βρίσκονται να είναι μεγέθη ευαίσθητα στην αναγνώριση του είδους της γλώσσας (ελληνικά, αγγλικά, ολλανδικά κ.τ.λ.) και του είδους του κειμένου (πολιτικά και οικονομικά άρθρα, αθλητικά νέα, λογοτεχνία) για την εν λόγω αναπαράσταση. Η διαφορά στις τιμές των εντροπιών αποδίδεται στην ομοιομορφία και την παρουσία πλατό στις κατανομές πιθανότητας των μηκών λέξεων αλλά και στις διαφορετικές συσχετίσεις μεταξύ των μηκών γειτονικών λέξεων στις υπό μελέτη χρονοσειρές. Με την σειρά της, η παρουσία των πλατό στις κατανομές πιθανότητας αντανακλά βασικές γλωσσολογικές ιδιότητες των διαφόρων γλωσσών, όπως τον πλούτο της κλιτικής μορφολογίας και την παραγωγικότητα της γλώσσας (μέσω μηχανισμών όπως είναι η παραγωγή, η σύνθεση και η σύμμειξη (blending)).

ABSTRACT

In the present work and in the context of mathematical and computational study of natural language, we carry out an entropic analysis of natural language texts in a word-length representation. Shannon's entropy and its generalization in the form of $n-gram$ entropy are found to be characteristic of the language (english, greek, finnish e.t.c.) as well as of the text genre (political and economical news, sports and literature). This is attributed to changes in the probability distribution of the lengths of single words (specifically the crucial role of the uniformity of probabilities of having words with length between five and ten) and the different word-length correlations in the studied symbolic series. On its behalf, the presence of the plateaus in the probability distributions reflects basic linguistic properties of the languages such as richness of inflectional morphology and productivity of a language through mechanisms like agglutination and synthesis of words.

ΕΥΧΑΡΙΣΤΙΕΣ

Είμαι ευγνώμων στους δασκάλους μου Βασσίλη Κωνσταντούδη και Χάρη Παπαγεωργίου για την ανεκτίμητη προσφορά, καθοδήγηση και υποστήριξή τους, κάνοντας την συνεργασία μας μία από τις πολυτιμότερες εμπειρίες.

Ένα ευχαριστώ στους Κωνσταντίνο Παπαδημητρίου, Κώστα Καραμάνο και τους καθηγητές μου Φώτη Διάκονο και Ανδρέα Σταφυλοπάτη για την συνεργασία τους και τα πολύτιμα σχόλιά τους και κυρίως τους γονείς μου για τη συνεχή συμπαράσταση τους που εκτείνεται πολύ πιο πέρα από τα όρια κάθε εργασίας.

Περιεχόμενα

1 ΕΙΣΑΓΩΓΗ

1.1 Μαθηματική και πληροφορική ανάλυση φυσικής γλώσσας

Ο Noam Chomsky δικαιολογεί την ανάγκη για μελέτη της φυσικής γλώσσας λέγοντας ότι είναι η εν μέρει μελέτη της ανθρώπινης φύσης όπως αυτή εκφράζεται μέσα στον ανθρώπινο εγκέφαλο. Κατά τον ίδιο, ένα από τα θεμελιώδη χαρακτηριστικά του ανθρώπου είναι η δημιουργικότητά του. Αρκεί να παρατηρήσει κανείς ότι (αν και η ιδέα έχει ξαναδιατυπωθεί) αυτή ακριβώς η πρόταση, μπορεί να μην έχει διατυπωθεί ξανά με αυτόν τον τρόπο. Για την ακρίβεια υπάρχει, και θα συνεχίζει να υπάρχει, ένα άπειρο πλήθος, σωστών φυσικά, προτάσεων που δεν έχουν ακόμα διατυπωθεί. Ο κάθε άνθρωπος, λοιπόν, δεν έχει στη διάθεσή του μία λίστα προτάσεων στην οποία μπορεί να ανατρέχει για να ελέγξει κάθε φορά την ορθότητα της φράσης που ακούει. Κατά το Chomsky, ο άνθρωπος κατέχει υποσυνείδητα ένα αφηρημένο σύστημα γνώσης που σχετίζεται με την γλώσσα. Το σύστημα αυτό αφορά σε κανόνες συντακτικούς, σημασιολογικούς και σημειολογικούς. Εντούτοις όμως, πώς μαθαίνουμε τους κανόνες αυτούς, πριν ακόμα τους διδαχθούμε ως μαθητές; Ίσως λοιπόν να υπάρχει μία ενδογενής συγγένεια του ανθρώπου εγκεφάλου με τη γλώσσα [6], [7].

Η αντίληψη αυτή έχει οδηγήσει επιστήμονες από πολύ διαφορετικούς μεταξύ τους κλάδους, όπως γλωσσολογία, γνωστική φιλοσοφία, ψυχολογία, μαθηματικά, πληροφορική, φυσική και άλλα, να εμπλακούν με την μελέτη της φυσικής γλώσσας. Μπορούμε να ξεχωρίσουμε δύο διαφορετικές κεντρικές οδούς στην ανάλυση του προφορικού και γραπτού λόγου, η προσέγγιση της θεωρητικής γλωσσολογίας κατά Chomsky και η προσέγγιση μέσω της θεωρίας πληροφορίας κατά Shannon. Στην εργασία αυτή, κινούμαστε πάνω στην δεύτερη οδό, αυτή της υπολογιστικής και μαθηματικής έρευνας πάνω στην ανθρώπινη γλώσσα.

Η σύγχρονη εξέλιξη των ηλεκτρονικών υπολογιστών στην ταχύτητα και την αποθήκευση πληροφορίας, έχει κυριολεκτικά απογειώσει την προσπάθεια διατύπωσης μοντέλων της φυσικής γλώσσας. Πέρα από την μελέτη της ίδιας της φύσης του ανθρώπινου εγκεφάλου, ένας πρακτικός σκοπός αυτού είναι η επίτευξη μίας βέλτιστης αλληλεπίδρασης μεταξύ ανθρώπου και ηλεκτρονικού υπολογιστή (Natural Language Processing, υπολογιστική γλωσσολογία). Έχει σημειωθεί σημαντική πρόοδος σε θέματα όπως αναγνώριση λόγου και ανάκτηση πληροφορίας από υπολογιστές καθώς και μηχανική μετάφραση και ανάλυση της γλώσσας. Ιστορικά η προσπάθεια ξεκίνησε από τον Alan Turing το 1950, με το άρθρο "Computing Machinery and Intelligence" όπου πρότεινε το λεγόμενο Turing test ως κριτήριο νοημοσύνης για έναν ηλεκτρονικό υπολογιστή: Ένας υπολογιστής είναι «ευφυής» όταν μπορεί να εμπλακεί σε διάλογο με έναν ανθρώπινο κριτή χωρίς ο δεύτερος να μπορεί να αποφανθεί για το αν μιλάει σε ανθρώπινο ον ή σε μηχανή [33]. Στον τομέα της μηχανικής μετάφρασης, η πρώτη μηχανή στατιστικής μετάφρασης κατασκευάστηκε τη δεκαετία του 1980, αν και οι προσπάθειες είχαν ξεκινήσει από το 1954 με το πείραμα του Georgetown. Στη συνέχεια η συγκεκριμένη περιοχή έρευνας σημείωσε μεγάλη πρόοδο με τη διατύπωση των πρώτων αλγορίθμων μηχανικής μάθησης (Μέχρι το 1980, όλα τα συστήματα επεξεργασίας φυσικού λόγου είχαν κατασκευαστεί ώστε να υπακούουν

σε ένα, πολύπλοκο βέβαια, σύνολο κανόνων). Τώρα πια, οι περισσότεροι τέτοιου είδους αλγόριθμοι είναι βασισμένοι σε μη επιβλεπόμενη ή ήμι-επιβλεπόμενη μάθηση, δηλαδή την αναγνώριση δομών από σύνολα δεδομένων που δεν έχουν υποστεί ανθρώπινη επεξεργασία.

Από μαθηματικής σκοπιάς, είναι αβέβαιο κατά πόσο μπορεί να διατυπωθεί μία πλήρης και συνεπής μαθηματική θεωρία για την φυσική γλώσσα, αν και ο ισχυρισμός έχει άλλοτε διατυπωθεί. Εντούτοις τα μαθηματικά είναι το άλφα και το ωμέγα στην προσπάθεια της ποσοτικοποίησης της πληροφορίας που μεταδίδεται μέσω του λόγου και της αναγνώρισης δομών μέσα σε αυτόν. Αν και ο σκοπός ενός ομιλητή ή συγγραφέα μπορεί να μελετηθεί μόνο θεωρητικά ή ανακριβώς, οι ελπίδες για μία οποιαδήποτε χειροπιαστή και ακριβή ανάλυση μπορούν να βρίσκονται μόνο στα φυσικά γεγονότα της ομιλίας και της γραφής. Άλλωστε με τον ένα ή τον άλλο τρόπο, τα επιμέρους συστατικά της γλώσσας (γράμμα, λέξη, πρόταση, κ.τ.λ.) δεν παύουν να είναι συμπαγής οντότητες με εσωτερική δομή [13]. Η υπεράσπιση της στενής συσχέτισης μεταξύ των γραμματικών και θεωρό-πληροφοριακών αρχών έγινε πρώτα από τον Zellig Harris το 1951, στο βιβλίο του "Structural linguistics" [12].

Για μερικούς συγγραφείς, η ανθρώπινη φυσική γλώσσα δεν διαφέρει ουσιαστικά από την γλώσσα των μαθηματικών. Η πρώτη προσπαθεί να διατυπώσει αρχές με την βοήθεια λέξεων, ενώ η δεύτερη με την βοήθεια συμβόλων. Ίσως βέβαια να υπάρχει μία αντίφαση στην προσπάθεια των μαθηματικών να εξηγήσουν την ανθρώπινη γλώσσα όταν αυτή είναι αναπόσπαστο κομμάτι της ύπαρξής τους, αλλά ίσως πάλι αυτό να μπαίνει στη σφαίρα της φιλοσοφίας.

Ιστορικά, ο κλάδος της μαθηματικής θεωρίας της πληροφορίας εισήχθη το 1948 με μία εργασία σταθμό από τον Claude Shannon, "A mathematical Theory of Communication" [27]. Ο Shannon προσέγγισε την στατιστική δομή της Αγγλικής γλώσσας σε ένα γραπτό κείμενο με χρησιμοποιώντας ένα πολύ απλό μαθηματικό μοντέλο, μία αλυσίδα Markov. Τώρα πια αλυσίδες Markov χρησιμοποιούνται κατά κόρων στην αναγνώριση του προφορικού και γραπτού λόγου, την ανάκτηση πληροφορίας, την συμπίεση δεδομένων και άλλα. Στην εργασία του αυτή, ο Shannon έκανε πρώτη φορά λόγο για τη μονάδα μέτρησης της πληροφορίας, το δυαδικό ψηφίο (binary digit ή αλλιώς το γνωστό μας bit) και συνέδεσε την πληροφορία με την αβεβαιότητα εισάγοντας την γνωστή θεωρό-πληροφοριακή εντροπία για την οποία κάνουμε λόγο παρακάτω. Έκτοτε η εντροπία έχει εμπλακεί αρκετά στην ανάλυση του γραπτού και προφορικού λόγου (σχετικές αναφορές βρίσκονται στο δεύτερο μέρος της εργασίας).

Ένας άλλος κλάδος των μαθηματικών (ή ίσως της θεωρίας των πολύπλοκων συστημάτων) που έχει σχετικά πρόσφατα συνεισφέρει στη μελέτη του φυσικού λόγου είναι η θεωρία των πολύπλοκων δικτύων. Υπάρχουν δύο διαφορετικές προσεγγίσεις σε αυτό το πεδίο. Η μία είναι η μελέτη των δικτύων ενός συγκεκριμένου γλωσσολογικού συστήματος (λεκτικά δίκτυα βασισμένα πάνω σε διαφορετικές σχέσεις ανάμεσα στις λέξεις, όπως συντακτικές ή σημασιολογικές) και η άλλη είναι τα κοινωνικά δίκτυα (δηλαδή δίκτυα χρηστών γλώσσας και ο ρόλος τους στην εξέλιξή της). Υπάρχει μία μεγάλη βιβλιογραφία στον τομέα αυτό η οποία όμως παραλείπεται εδώ αφού η δική μας προσέγγιση στο ζήτημα είναι εντροπικής φύσης.

1.2 Η γλώσσα ως πολύπλοκο σύστημα

Αν και κάποιος θα μπορούσε να ισχυριστεί ότι πολύπλοκα συστήματα μελετώνται από τον άνθρωπο εδώ και χιλιετίες, η μοντέρνα επιστημονική αντιμετώπιση της πολυπλοκότητας έχει αρχίσει να ανθίζει την τελευταία εικοσαετία. Ας κοιτάξουμε λοιπόν την φυσική γλώσσα μέσα από την σκοπιά της πολυπλοκότητας, αφού πρώτα κάνουμε μία ανασκόπηση στο «τι είναι» ένα πολύπλοκο σύστημα.

Στην παρούσα χρονική στιγμή δεν έχει δοθεί ένας κοινά αποδεκτός, αυστηρός ορισμός για το τι είναι ένα πολύπλοκο σύστημα. Συνηθίζουμε να καλούμε πολύπλοκο, ένα σύστημα όταν η συμπεριφορά ή οι ιδιότητές του δεν μπορούν να γίνουν άμεσα αντιληπτές από τις ιδιότητες των επιμέρους συστατικών του.
Πιο συγκεκριμένα, ή ίσως πιο ξεκάθαρα, έχουν διατυπωθεί κάποιες κοινές, ευρέως αποδεκτές ιδιότητες, που σχετίζονται με την πολυπλοκότητα [2].

1. *Τα επιμέρους συστατικά ενός πολύπλοκου συστήματος είναι αλληλοεξαρτώμενα και ως εκ τούτου ένα πολύπλοκο σύστημα είναι ικανό να παρουσιάσει μία ιδιόμορφη ανακύπτουσα συμπεριφορά (emergence).* Ένα χαρακτηριστικό παράδειγμα είναι αυτό του ανθρώπινου σώματος και των ζωτικών του οργάνων. Αν αφαιρέσουμε τυχαία ένα από αυτά, το πιθανότερο είναι να επιδράσουμε δραστικά στην λειτουργία και ίσως και την ύπαρξη του όλου συστήματος. Ένα παράδειγμα ανακύπτουσας συμπεριφοράς για το παραπάνω πολύπλοκο σύστημα είναι το περπάτημα.

2. *Ένα πολύπλοκο σύστημα έχει συνήθως δομή σε διαφορετικές κλίμακες.* Χρησιμοποιώντας και πάλι το παραπάνω παράδειγμα του ανθρώπινου σώματος η πρώτη κλίμακα θα μπορούσε να είναι το κεφάλι, ο κορμός και τα μέλη, μία δεύτερη κλίμακα αυτή των ζωτικών οργάνων (μύες, οστά, νεύρα κ.τ.λ.) ενώ μία τρίτη κλίμακα αυτή των κυττάρων (πυρήνας, μιτοχόνδρια, κυτόπλασμα, κ.τ.λ.). Κανείς, μπορεί να συνεχίσει σε μία τέταρτη ή πέμπτη κλίμακα κ.τ.λ.

3. *Τα επιμέρους συστατικά ενός πολύπλοκου συστήματος αλληλεπιδρούν μεταξύ τους με τρόπο μη γραμμικό.*
   Σε αυτό το σημείο θα μπορούσε να σχολιαστεί ότι αν και η μη γραμμικότητα, αρκετές φορές παραπέμπει ενστικτωδώς στην έννοια του χάους, η αλήθεια είναι ότι ένα χαοτικό σύστημα δεν είναι εν γένει πολύπλοκο. Κάθε άλλο μάλιστα. Για να γίνει πιο εύκολα αντιληπτό κάτι τέτοιο αρκεί κάποιος να παρατηρήσει ότι για το χάος αρκούν λίγοι βαθμοί ελευθερίας ενώ για την πολυπλοκότητα όχι.

Οι παραπάνω είναι ίσως οι χαρακτηριστικότερες ιδιότητες που καθιστούν ένα σύστημα πολύπλοκο αν και έχουν βεβαίως διατυπωθεί πολλές άλλες.

Η ανθρώπινη γλώσσα είναι ένα πολύπλοκο σύστημα από τη σκοπιά της ίδιας της της φύσης ως κατασκεύασμα της ανθρώπινης νόησης. Πρόκειται για μία φυσιολογική, νευρολογική, ψυχολογική και κοινωνιολογική οντότητα [4].
Από γλωσσολογικής άποψης, η πολυπλοκότητα αντανακλάται στη φωνητική, τη μορφολογία,

το συντακτικό, τη σημασιολογία, το λεξικό, τη σημειολογία κ.τ.λ.
Αν ανακαλέσουμε τις παραπάνω ιδιότητες που χαρακτηρίζουν ένα πολύπλοκο σύστημα, στο επίπεδο αυτό, μπορούμε να αναγνωρίσουμε την αλληλοεξάρτηση των επιμέρους συστατικών της γλώσσας. Αν αφαιρέσουμε τυχαία μία λέξη από μία πρόταση ή μία πρόταση από μία παράγραφο, είναι πιθανό να επιδράσουμε λιγότερο ή περισσότερο στο νόημα του κειμένου.
Μπορούμε επίσης, σαφώς, να δούμε την δομή του λόγου σε διαφορετικές κλίμακες. Για παράδειγμα, ξεκινώντας από την λέξη, τη μικρότερη μονάδα που περιέχει νόημα, μπορούμε να πάμε στην πρόταση, την παράγραφο, την ενότητα κ.τ.λ.
Η μη γραμμικότητα των επιμέρους συστατικών της γλώσσας στο γλωσσολογικό επίπεδο, είναι μία έννοια λίγο πιο γενική και ανεικονική. Μία άποψη αυτής γίνεται αντιληπτή από τους γλωσσολόγους ως τις συσχετίσεις μεγάλου μήκους σε μία πρόταση ή ένα κείμενο. Για παράδειγμα, αν σε μία πρόταση ενός λογοτεχνικού βιβλίου γίνεται λόγος για ένα χαρακτηριστικό του ήρωα της ιστορίας, ο αναγνώστης μπορεί να το αντιληφθεί, ακόμα και αν στη πρόταση αυτή δεν αναφέρεται ρητά το όνομα του υποκειμένου.

1.3 Σκοπός και κίνητρα της εργασίας

Σχεδόν σε όλες τις περιπτώσεις μελέτης κάποιου πολύπλοκου συστήματος, είναι αδύνατον να ξεκινήσουμε από υπάρχοντες φυσικούς νόμους ή πρώτες αρχές. Έχουμε ανάγκη μία στατιστική μέθοδο. Σε αυτή την εργασία και στο πλαίσιο συμφιλίωσης της επιστήμης των πολύπλοκων συστημάτων με την μελέτη της φυσικής γλώσσας, εφαρμόζουμε ένα από τα γνωστότερα εργαλεία της θεωρίας πληροφορίας, την εντροπία, σε μία αναπαράσταση μηκών λέξεων γραπτών και προφορικών κειμένων. Αναζητούμε διαφοροποιήσεις στις τιμές του μεγέθους αυτού, μεταξύ διαφορετικών γλωσσών (Ελληνικά, Αγγλικά, Γερμανικά, κ.τ.λ.) και διαφορετικών ειδών κειμένου (λογοτεχνία, πολιτική, οικονομικά, αθλητικά).

Ο σκοπός μας είναι πολλαπλός. Σε ειδικότερο πλαίσιο, η εντροπία μας βοηθάει να αναγνωρίσουμε συσχετίσεις στα διαφορετικά επίπεδα του υπό μελέτη συστήματος: Πώς αντανακλούν οι αναπαραστάσεις μήκους λέξεων το είδος της γλώσσας και το είδος του κειμένου. Σε γενικότερο πλαίσιο, ο ακριβής υπολογισμός της εντροπίας γραπτών κειμένων έχει μεγάλη σπουδαιότητα, αφενός για την ίδια την θεωρία της πληροφορίας αφού μπορεί να καθοδηγήσει την συμπίεση των δεδομένων στην διαδικασία της επικοινωνίας τους και αφετέρου στην μοντελοποίηση της γλώσσας αφού ένα σωστό μοντέλο θα πρέπει να έχει ίδια εντροπία με την ίδια την γλώσσα. Βέβαια, εφόσον παραμένει άγνωστη η ακριβής κατανομή πιθανοτήτων των συμβόλων - ή στην προκειμένη περίπτωση των μηκών των λέξεων - μίας γλώσσας, ο στατιστικός υπολογισμός της εντροπίας κάποιων κειμένων δε μπορεί παρά να είναι ένα (όχι βέλτιστο) άνω φράγμα για την πραγματική τιμή της και για την ώρα θα πρέπει να αρκεστούμε σε αυτό [3]. Σε κάθε περίπτωση, η μελέτη μίας τυχαίας διαδικασίας όπως αυτή της φυσικής γλώσσας, κατά τον Shannon, δε μπορεί παρά να μας οδηγήσει ένα βήμα πιο μπροστά στη κατανόηση της φύσης άλλων τυχαίων διαδικασιών ζώντων οργανισμών [17].

2 ΜΕΘΟΔΟΛΟΓΙΑ: ΣΥΜΒΟΛΟΣΕΙΡΕΣ ΚΑΙ ΕΝΤΡΟΠΙΚΗ ΑΝΑΛΥΣΗ

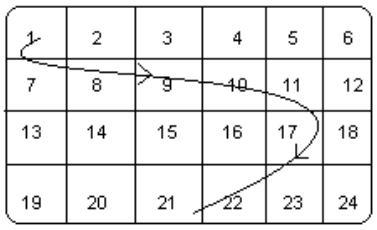
2.1 Έννοιες και ορισμοί

2.1.1 Χρονοσειρές - συμβολοσειρές

Οι επιστήμονες έχουν την τάση να αναζητούν για μοτίβα και να εδραιώνουν συγγένειες ή συνάφειες για τα φαινόμενα γύρω τους. Για να πετύχουν κάτι τέτοιο συνήθως ξεκινούν από την απλή μέτρηση μίας ή περισσότερων μεταβλητών για το σύστημά τους, που μεταβάλλονται χρονικά (ή μερικές φορές χωρικά). Έτσι, καταλήγουν με μία ή περισσότερες χρονοσειρές (ή χωροσειρές), όπως λέγονται, που ενθυλακώνουν την δυναμική του εν λόγω συστήματος στο χρόνο (ή το χώρο αντίστοιχα). Στη συνέχεια, προκειμένου για την εξαγωγή συμπερασμάτων, οι μέθοδοι ανάλυσης των χρονοσειρών εκτείνονται σε ένα πολύ ευρύ φάσμα με χρήση εργαλείων από την φυσική, τη στατιστική ή της θεωρία της πληροφορίας. Παραδείγματος χάρη, μέτρα πολυπλοκότητας και εντροπίες (όπως μερικά από αυτά που αναφέρονται παρακάτω), κλασική ανάλυση Fourier, model-based μέθοδοι, κ.α.
Σε αδρές γραμμές, μία χρονοσειρά είναι προφανώς μία συμβολοσειρά. Εντούτοις, συνηθίζουμε να κάνουμε χρήση του όρου 'συμβολοσειρά' όταν από την αρχική χρονοσειρά έχουμε περάσει, μέσω μίας συμβολικής αναπαράστασης σε μία πιο αδρομερή περιγραφή της δυναμικής της υπό μελέτης μεταβλητής.

Συμβολική Δυναμική
Κατά τη μελέτη, λοιπόν, ενός συστήματος, το ζητούμενο είναι να καταφέρουμε να εξάγουμε κάποια από τα χαρακτηριστικά του, με απώτερο σκοπό την ανακάλυψη ενδεχόμενων νόμων που το διέπουν.
Ένα από τα προβλήματα των πειραματικών της επιστήμης είναι να βρουν τη χρυσή τομή ανάμεσα στην πεπερασμένη ακρίβεια με την οποία αναπόφευκτα γίνονται οι παρατηρήσεις των φυσικών φαινομένων, και την αυστηρότητα με την οποία πρέπει να διεξαχθούν τα συμπεράσματα.
Για την εν μέρει επίλυση του προβλήματος αυτού, εισήχθη η συμβολική δυναμική. Είναι μια μέθοδος αδρομερούς περιγραφής (coarse-graining) του φασικού χώρου των υπό μελέτη μεταβλητών, όπως λέγεται, η οποία βασίζεται στην πεπερασμένης ακρίβειας παρατήρηση και μέτρηση, ενώ ταυτόχρονα παρέχει, ανάλογα με το πρόβλημα, αρκούντως αυστηρά συμπεράσματα.
Για παράδειγμα αν η απάντηση σε ένα ερώτημα μπορεί να περιοριστεί σε ένα «ναι» ή ένα «όχι» τότε μια μέτρηση χωρίς πολύ επεξεργασία μπορεί να δώσει ακριβή απάντηση.
Η επιτυχία εξαρτάται από την φύση του προβλήματος και από τον τρόπο με τον οποίο επεξεργαζόμαστε τα δεδομένα. Αυτή είναι και η ουσία της συμβολικής δυναμικής.
Ιδέες πάνω σε αυτό είχαν ήδη δοθεί από τον C. Shannon το 1951 στην εργασία του με τίτλο "Prediction and Entropy of Printed English" [28]. Έκτοτε έχουν εφαρμοστεί σε μία ευρεία γκάμα θεμάτων μεταξύ των οποίων οι βιοσειρές και άλλοι φορείς πληροφορίας.

Ας δούμε όμως ένα παράδειγμα.
Έστω ότι κατατμίζουμε το φασικό χώρο ενός δυναμικού συστήματος συνεχούς χρόνου, σε μη αλληλεπικαλυπτόμενες κυψελίδες. Δίνουμε ένα μοναδικό σύμβολο (- «όνομα») στην κάθε κυψελίδα, όπως φαίνεται στο σχήμα παρακάτω (Σχήμα 1), και αφήνουμε το σύστημα να εξελιχθεί στο χρόνο. Σημειώνουμε κάθε φορά το σύμβολο της κυψελίδας από την οποία πέρασε η



Σχήμα 1: Παράδειγμα αδρομερούς περιγραφής της εξέλιξης μίας τροχιάς

τροχιά. Η ακολουθία των συμβόλων στο τέλος ονομάζεται συμβολική τροχιά. Στο παράδειγμα του σχήματος η τροχιά σε όρους συμβολικής δυναμικής είναι

$$1 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 17 \quad 23 \quad 22 \quad 21$$

Με τη μέθοδο αυτή χάνουμε αρκετή πληροφορία, αλλά κάποιες ουσιώδης ιδιότητες του συστήματος διατηρούνται. Δύο από αυτές, για παράδειγμα, είναι η περιοδικότητα και η χαοτικότητα. Στη μελέτη των δύο αυτών χαρακτηριστικών η συμβολική δυναμική αποδεικνύεται χρησιμότατη. Αν επιπλέον ληφθεί υπόψη και η γεωμετρία του συστήματος (π. χ. τυχόν συμμετρίες) , είναι δυνατόν με μία στοιχειώδη διαμέριση να πετύχουμε πολύ περισσότερα.

2.1.2 Εντροπίες και άλλα μέτρα πολυπλοκότητας στην ανάλυση συμβολοσειρών

Έχουμε κάνει ήδη ένα πρώτο λόγο για την θεωρία της πληροφορίας, η οποία έχει λάβει μεγάλη προσοχή τα τελευταία χρόνια, κυρίως εξαιτίας της ανερχόμενης αυτής επιστήμης των πολύπλοκων συστημάτων. Είναι εκείνος ο κλάδος των εφαρμοσμένων μαθηματικών που αποσκοπεί στην ποσοτικοποίηση της πληροφορίας που σχετίζεται με ένα σύστημα. Εργαλεία από την θεωρία της πληροφορίας, όπως η εντροπία του Shannon, έχουν βοηθήσει ουσιαστικά στην περιγραφή και την ποσοτικοποίηση της δυναμικής ενός πολύπλοκου συστήματος μέσω της αποτίμησης του «ποσού» της τάξης ή αταξίας σε αυτό. Η εντροπία, λοιπόν είναι ένα χαρακτηριστικό μέτρο πολυπλοκότητας που συνήθως εκφράζεται ως ο μέσος αριθμός ψηφίων που απαιτείται για την

συμπίεση, αποθήκευση ή επικοινωνία δεδομένων.
Ιστορικά ωστόσο, η ανάμειξη της θεωρίας της πληροφορίας με την έννοια της εντροπίας και την πολυπλοκότητα ξεκίνησε σχεδόν 90 χρόνια πριν μέσω της ίδιας της προσπάθειας να μετρηθεί ο ρυθμός που παίρνουμε πληροφορία από μια συμβολοσειρά:

- Nyquist 1924: ο ρυθμός της πληροφορίας κάποια χρονική στιγμή αντιστοιχεί στον λογάριθμο του αριθμού των συμβόλων που πρέπει να μεταδίδουμε κάθε χρονική στιγμή [20].
- Hartley 1928: η πληροφορία είναι ο λογάριθμος του μεγέθους του λεξικού [14].
- Shannon 1948 [27].
- Solomonoff 1964: η πιθανότητα Solomonoff με δεδομένη Turing Machine Σ είναι η πιθανότητα $\Sigma(X) = S$ για τυχαίο X. Αυτή η πιθανότητα δεν είναι εν γένει υπολογίσιμη [30], [31].
- Lempel-Ziv 1976 [36].
- Pincus 1991 [23].
- Titchener 1998 [35].

Η εντροπία του Shannon
Η εντροπία του Shannon είναι μία μαθηματική ποσοτικοποίηση του βαθμού της αβεβαιότητας ενός πειράματος με περισσότερα από ένα πιθανά αποτελέσματα [27], [28], [8].
Έστω ένας δειγματικός χώρος X με n το πλήθος γεγονότα $w_1, w_2, ..., w_n$, οι πιθανότητες των οποίων είναι $p_1, p_2, ..., p_3$ αντίστοιχα. Δεδομένου ότι μας ενδιαφέρουν μόνο οι αριθμητικές τιμές των πιθανοτήτων, ο δειγματικός αυτός χώρος $(X; p_1, p_2, ..., p_n)$ μπορεί να συμβολιστεί απλούστερα ως $(p_1, p_2, ..., p_n)$.
Επιπλέον ισχύει η σχέση

$$\sum_{i=1}^{n} p_i = 1, \quad p_i \geq 0, \quad i = 1, 2, ..., n$$

Θέλουμε να ορίσουμε μία μη αρνητική συνάρτηση H η οποία ορίζεται σε όλους τους δειγματικούς χώρους $(p_1, p_2, ..., p_n)$ για κάθε θετικό ακέραιο n. Η τιμή της H στον $(p_1, p_2, ..., p_n)$ συμβολίζεται με $H(p_1, p_2, ..., p_n)$. Αυτός ο αριθμός θα χρησιμοποιείται για να χαρακτηρίσει μαθηματικά το βαθμό της αβεβαιότητας του αποτελέσματος των γεγονότων $w_1, w_2, ..., w_n$ με πιθανότητες $p_1, p_2, ..., p_3$ αντίστοιχα.
Η $H(p_1, p_2, ..., p_n)$ θα πρέπει να πληρεί τις ακόλουθες προϋποθέσεις:

- Για κάθε θετικό ακέραιο $n > 0$, η H είναι συνεχής συνάρτηση των μεταβλητών $p_1, p_2, ..., p_3$.
- Αν $p_i = \frac{1}{n}$ για κάθε $i = 1, 2, ..., n$, τότε η αντίστοιχη $H(\frac{1}{n}, ..., \frac{1}{n})$ είναι μονότονα αύξουσα ως προς n.

- Αν ένα πείραμα μπορεί να εκτελεστεί σε μικρότερα διαδοχικά πειράματα, τότε η αρχική τιμή της H είναι το σταθμισμένο άθροισμα των αντίστοιχων H των επιμέρους πειραμάτων.

Κάθε πραγματική μη αρνητική συνάρτηση H που ικανοποιεί τις παραπάνω προϋποθέσεις πρέπει να είναι της μορφής

$$H(p_1, p_2, ..., p_3) = -K \sum_{i=1}^{n} p_i ln p_i,$$

όπου K θετική σταθερά. (Απόδειξη [8])
Από την παραπάνω έκφραση της H αν $p_i = 1$ για κάποιο i, τότε $H(p_1, p_2, ..., p_n) = 0$. Αναμενόμενο, αφού αν $p_i = 1$ για κάποιο i τότε το συγκεκριμένο γεγονός συμβαίνει πάντα και άρα η αβεβαιότητα είναι μηδέν.

Ορισμός: Η ποσότητα

$$H(p_1, p_2, ..., p_3) = -\sum_{i=1}^{n} p_i ln p_i,$$

καλείται εντροπία του Shannon και αφορά στο δειγματικό χώρο $(p_1, p_2, ..., p_3)$.
Εφόσον η εντροπία είναι μία μαθηματική έκφραση αβεβαιότητας, περιμένουμε ότι όταν $p_1 = ... = p_n = \frac{1}{n}$ η τιμή της H γίνεται μέγιστη.

Ας επιστρέψουμε τώρα στην έννοια των χρονοσειρών και συμβολοσειρών.

Έστω ότι s είναι μία ακολουθία μήκους N αποτελούμενη από σύμβολα ενός πεπερασμένου αλφαβήτου συνολικού αριθμού γραμμάτων λ. Υποακολουθίες n γραμμάτων θα καλούνται με τον όρο n-λέξεις ή $n - blocks$ ή $n - grams$ (Παρακάτω, οι τρεις αυτοί όροι ταυτίζονται). Υποθέτοντας την ύπαρξη στασιμότητας, κάθε $n-gram\ i$ αναμένεται να προκύψει με πιθανότητα $p_i^{(n)}$, σε κάθε αυθαίρετο σημείο της ακολουθίας. Συμφωνούμε ότι σαρώνοντας την ακολουθία από τα αριστερά προς τα δεξιά, το διάβασμα μπορεί να γίνεται με δύο διαφορετικούς τρόπους.

1. Κατά τμήματα (lumping process), όπου το πρώτο γράμμα κάθε $n-gram$ ξεκινάει αμέσως μετά το τελευταίο γράμμα του προηγούμενου, όπως φαίνεται στο σχήμα παρακάτω.

$$... \underbrace{s_1 ... s_n}_{x_1} \underbrace{s_{n+1} ... s_{2n}}_{x_2} ... \underbrace{s_{jn+1} ... s_{(j+1)n}}_{x_{j+1}} ...$$

2. «Γλιστρώντας» πάνω στην ακολουθία με κάθε γράμμα της να αποτελεί το πρώτο γράμμα κάθε $n - gram$ (gliding process), όπως φαίνεται παρακάτω.

$$... \underbrace{s_1 ... s_n}_{x_1} s_{n+1} ... s_{j-1+n} ... \quad , \quad ... s_1 \underbrace{s_2 ... s_{n+1}}_{x_2} ... s_{j-1+n} ... \quad , \quad ...$$

Αν και στην εργασία αυτή δεν εμπλεκόμαστε με αυστηρώς [10] χαοτικά συστήματα, για λόγους πληρότητας αναφέρουμε ότι η κατά τμήματα ανάγνωση της συμβολοσειράς (lumping), σε αντίθεση

με την κατά συνεχή τρόπο ανάγνωση (gliding) έχει αποδειχθεί ότι είναι περισσότερο ευαίσθητη όσον αφορά τις αλγοριθμικές και εργοδικές ιδιότητες των ασθενώς χαοτικών συστημάτων [15]. Στην παρούσα εφαρμογή χρησιμοποιούμε την gliding process.

Οι παρακάτω ποσότητες χαρακτηρίζουν την πληροφορία που περιέχεται στη συμβολοσειρά.

1. **Τμηματική Εντροπία (Block Entropy or $n-gram$ Entropy)**
   Το μέγεθος αυτό επεκτείνει τον ορισμό του Shannon από την εντροπία μίας κατάστασης στην εντροπία διαδοχικών καταστάσεων.
   Οι $n-gram$ εντροπία λέξεων μήκους n, δίνεται από τον τύπο:

$$\Phi_n = -\sum_{(s_1,...,s_n)} p^{(n)}(s_1,...,s_n)\ln(p^{(n)}(s_1,...s_n))$$

   Η άθροιση γίνεται πάνω σε όλα τα μοναδικά $n-grams$, ενώ

$$p^{(n)}(s_1,...,s_n) = \frac{\text{No. of n-grams } (s_1,...,s_n), \text{encountered when lumping or gliding}}{\text{Total No. of n-grams}}$$

   είναι η πιθανότητα να συναντήσουμε ένα συγκεκριμένο $n-gram$. Προφανώς η μέγιστη τιμή της εντροπίας προκύπτει για την συμβολοσειρά στην οποία όλα τα $n-grams$ προκύπτουν με ίση πιθανότητα. Η Φ_n μετρά την ποσότητα της πληροφορίας που περιέχεται σε μία λέξη μήκους n, ή ισοδύναμα τη μέση πληροφορία απαραίτητη για την πρόβλεψη μίας υποακολουθίας μήκους n.

2. **Δεσμευμένη Εντροπία**
   Ως συνέπεια των παραπάνω μπορούμε να ορίσουμε την δεσμευμένη εντροπία h_n ως την μέση απαραίτητη πληροφορία για να προβλέψουμε το επόμενο σύμβολο, δεδομένων των προηγούμενων n συμβόλων μέσω του τύπου

$$h_n = \Phi_{n+1} - \Phi_n$$

   Ο παραπάνω τύπος συμπληρώνεται από την αρχική συνθήκη $h_0 := \Phi_1$. Η δεσμευμένη εντροπία h_n ικανοποιεί την ανισότητα $h_{n+1} \leq h_n$. Αν η h_n, μειώνεται εκθετικά ως προς n δεν μπορούμε να κάνουμε πρόβλεψη του επόμενου συμβόλου ακόμα και για σχετικά μικρά μήκη ακολουθίας n, που σημαίνει συσχετίσεις μικρού μήκους μεταξύ των συμβόλων (short-range correlations).

3. **Εντροπία Πηγής (Source Entropy)**
   Μία άλλη ποσότητα, ιδιαίτερου ενδιαφέροντος, είναι η εντροπία της πηγής (source entropy) που ορίζεται ως το όριο της δεσμευμένης εντροπίας για μεγάλα n.

$$h := lim_{n\to\infty} h_n = lim_{n\to\infty} \frac{\Phi_n}{n}$$

Η οριακή αυτή εντροπία, που είναι το διακριτό ανάλογο της Kolmogorov-Sinai (K-S) entropy , είναι η μέση απαραίτητη πληροφορία για την πρόβλεψη του επόμενου συμβόλου δεδομένης της πλήρους ιστορίας του συστήματος.
Σημαντικό εδώ είναι το γεγονός ότι μία θετική εντροπία Kolmogorov-Sinai συνεπάγεται την ύπαρξη ενός θετικού εκθέτη Lyapunov και άρα είναι ένας σημαντικός δείκτης χάους. Επίσης η ταχύτητα σύγκλισης των διαφορικών εντροπιών στο όριο τους h, μπορεί να θεωρηθεί και ένα μέτρο συσχέτισης.

4. **Προσεγγιστική Εντροπία (Approximate Entropy)**
Η προσεγγιστική εντροπία, διαφοροποιείται ουσιαστικά από τις παραπάνω, με την έννοια ότι συνήθως δεν απαιτείται συμβολική δυναμική για τον υπολογισμό της, «δρα» κατευθείαν στην αρχική χρονοσειρά. Το φυσικό της νόημα μπορεί να διατυπωθεί ως εξής. Όσο περισσότερα επαναληπτικά πρότυπα περιέχονται στη χρονοσειρά τόσο μικρότερη η τιμή της προσεγγιστικής εντροπίας.
Ο αλγόριθμος υπολογισμού της περιγράφεται παρακάτω.

Δεδομένης της ακολουθίας s μήκους N, θεωρούμε $n-grams$ 'γλιστρώντας ' πάνω στην s (gliding process).

$$x_i := (s_i, s_{i+1}, ..., s_{i+n-1})$$

όπου $i = 1, ..., N-n+1$.
Με σκοπό να υπολογίσουμε τη συχνότητα με την οποία τα $n-grams$ αυτά επαναλαμβάνονται με μία «ανεκτικότητα» r (tolerance), μέσα στο σύνολο των δεδομένων, ορίζουμε ως απόσταση δύο $n-grams$ x_i και x_j ως

$$d(x_i, x_j) = max|s_{i+k} - s_{j+k}|$$

όπου $0 \leq k < n$.
Ακολούθως, δύο $n-grams$ x_i και x_j είναι *παρεμφερή*, δηλαδή όμοια κάτω από μία ανεκτικότητα r, αν

$$d(x_i, x_j) < r$$

Ορίζουμε τη ποσότητα C_i^n, ως τη συχνότητα με την οποία προκύπτουν όμοια $n-grams$ κάτω από την ανεκτικότητα r.

$$C_i^n(r) = \frac{\text{No. of j such that } d(x_i, x_j) < r}{N-n+1}$$

όπου $j \leq N-n+1$.
Στη συνέχεια παίρνουμε τη μέση τιμή του νεπέριου λογαρίθμου της παραπάνω ποσότητας για να ορίσουμε επιπλέον

$$F^n(r) = \sum_i \frac{ln(C_i^n(r))}{(N-n+1)}$$

όπου $i = 1, ..., N - n + 1$.
Τελικά η προσεγγιστική εντροπία σε ορίζεται ως εξής

$$ApEn(n, r, N) = F^n(r) - F^{n+1}(r)$$

Η διάσταση n και η ανεκτικότητα r είναι δύο πολύ κρίσιμες για το αποτέλεσμα της $ApEn(n, r, N)$ παράμετροι, εντούτοις γίνονται ακόμα προσπάθειες διατύπωσης μεθόδων για την βελτιστοποίησή τους. Πολύ μικρές τιμές r δίνουν πολύ φτωχές εκτιμήσεις δεσμευμένων πιθανοτήτων ενώ πολύ μεγάλες τιμές r αποκλείουν πολύ πληροφορία για το σύστημα. Καταρχήν όμως, η μέγιστη $ApEn$ είναι εκείνη που αναδεικνύει καλύτερα την πολυπλοκότητα του συστήματος. Αναφορικά, όταν ο Pincus [23] πρότεινε την προσεγγιστική εντροπία, την εφάρμοσε για τρία ευρέως χρησιμοποιούμενα θεωρητικά μοντέλα, τις απεικονίσεις Rossler, Henon και λογιστική, όπου συμπέρανε ότι το φάσμα τιμών $0.1 * std < r < 0.2 * std$ και $n = 2$ παρέχει μία καλή στατιστική ακρίβεια [23]. Έκτοτε, οι τιμές αυτές έχουν χρησιμοποιηθεί αρκετα στη βιβλιογραφία. Το 2009, οι Chon et. al. [18] απέδειξαν ότι η μέγιστη προσεγγιστική εντροπία δεν βρίσκεται για όλα τα φυσικά συστήματα στο εύρος που πρότεινε ο Pincus και έδωσαν μία μέθοδο υπολογισμού του βέλτιστου r για ένα συγκεκριμένο n και ένα οποιοδήποτε σήμα.

Η προσεγγιστική εντροπία, σε αδρές γραμμές, μπορεί να διατυπωθεί ως τη σχετική διάδοση των επαναλαμβανόμενων προτύπων μήκους n όπως αυτά συγκρίνονται με πρότυπα μήκους $n + 1$. Συχνά διατυπώνεται στη βιβλιογραφία και μία άλλη αντίληψη αυτής της ποσότητας. Μικρές τιμές προσεγγιστικής εντροπίας υπονοούν τη δυνατότητα σύμπτυξης των δεδομένων χωρίς την απώλεια ουσιαστικής πληροφορίας (high compressibility).
Το μέγεθος αυτό προτάθηκε από τον S. Pincus, το 1990 [23], [11].

Η κλασσική (Shannon) και η αλγοριθμική θεωρία της πληροφορίας (Kolmogorov) δεν είναι αρκετές να περιγράψουν φυσικά συστήματα καθώς αναφέρονται σε πιθανότητες και υποθέσεις, κάτι που στην φύση δεν υπάρχει. Στην φύση συναντάμε παντού πεπερασμένες χορδές όπως π.χ. σε γραπτά κείμενα, στο DNA κ.α. Παρουσιάστηκε λοιπόν η ανάγκη να οριστούν νέα μεγέθη που να μπορούν να υπολογίζουν εύκολα την πολυπλοκότητα και την πληροφορία μιας χορδής (Lempel, Ziv, Pincus, Titchener). Παρακάτω, για ιστορικούς λόγους αναφέρουμε μερικά από αυτά.

5. **T-Complexity**
O Mark Titchener (1998) πρότεινε μια νέα μέθοδο υπολογισμού της πολυπλοκότητας μιας χορδής που σχετίζεται πολύ με εκείνη που είχε προταθεί από τους Lempel και Ziv [36]. Εδώ όμως χρησιμοποιείται ο λεγόμενος recursive hierarchical pattern copying algorithm (RHPC). Η T-Complexity δηλαδή είναι ο αριθμός των βημάτων που απαιτούνται για να παραχθεί η χορδή από το αλφάβητο της. Η τιμή της είναι μοναδική και επομένως οδηγεί σε έναν μοναδικό χαρακτηρισμό της συγκεκριμένης χορδής. Σχετίζεται με το μέγεθος της χορδής αλλά και την πληροφορία που φέρει αυτή.

Ο αλγόριθμος:

Έστω σειρά $x(n)$, μήκους n και αλφαβήτου A. Αυτή σαρώνεται και παράγει τις λέξεις $p_i \in A^+$ και τους αντίστοιχους εκθέτες (copy-exponents) $k_i \in N^+$, $i = 1, 2, ..., q$ με $q \in N^+$:

$$x = p_q^{k_q} p_{q-1}^{k_{q-1}} ... p_i^{k_i} ... p_1^{k_1} \alpha_0$$

Για κάθε λέξη, πρέπει:

$$p_i = p_{i-1}^{m_{i,i-1}} p_{i-2}^{m_{i,i-2}} ... p_j^{m_{i,j}} ... p_1^{m_{i,1}} \alpha_i$$

με $\alpha_i \in A$ και $0 \leq m_{i,j} \leq k_j$
Η T-complexity ορίζεται με την βοήθεια των k_i:

$$C_T(x(n)) = \sum_i^q ln(k_i + 1)$$

Οι μονάδες της T-Complexity είναι τα taugs (t-augmentation steps) που είναι «ο αριθμός των βημάτων που χρειάζονται για να παραχθεί η χορδή. Μπορεί κάποιος να παρατηρήσει ότι για μια χορδή που έχει μόνο ένα σύμβολο, η C_T γίνεται ελάχιστη, $ln(n)$, οπότε:

$$C_T(x(n)) \geq ln(n)$$

Το άνω φράγμα για $n > n_0$ είναι:

$$C_T(x(n)) \leq li \lfloor ln(2ln(\#A^n)) \rfloor$$

όπου $li(z) \equiv \int_0^z \frac{du}{ln(u)}$ είναι η "logarithmic integral function". Στην πράξη δηλαδή ο RHPC αλγόριθμος βασίζεται στο ότι σαρώνουμε την χορδή από αριστερά προς τα δεξιά αλλά διαλέγουμε τις λέξεις από δεξιά προς τα αριστερά (αυτό θα φανεί αργότερα).

6. **T-Information**
   Η T-information ορίζεται ως το "inverse logarithmic integral" της T-Complexity διαιρεμένο με μια σταθερά κανονικοποίησης $ln2$:

$$I_T(x(n)) = li^-1(\frac{C_T(x(n))}{ln2})$$

   Μονάδες μέτρησης είναι τα nats.
   Για $n \to \infty$, $I_T(x(n)) \leq ln(\#A^n)$. Παρατηρούμε πάλι ότι το δεξί μέλος είναι η μέγιστη εντροπία χορδής n κατά Shannon.

7. **T-Entropy**
   Η T-Entropy είναι ο μέσος ρυθμός T-Information ανά σύμβολο.

$$\overline{h_T}(x(n)) = \frac{I_T(x(n))}{n}$$

   Μονάδες μέτρησης είναι τα nats/symbol.
   Gia $n \to \infty$, $\overline{h_T}(x(n)) \leq ln(\#A)$.

Παράδειγμα:
Θα υπολογίσουμε την T-Complexity, την T-Information και την T-Entropy για μία χορδή x:

- Έστω $x =$`1011010100010′ με αλφάβητο $A = 0, 1$.
  Παρατηρούμε ότι $a_0 =$`0′ και άρα $p_1 =$`1′. Αφού το p_1 δεν επαναλαμβάνεται αμέσως μετά, τότε $k_1 = 1$.

- Στην συνέχεια σαρώνουμε την χορδή από αριστερά προς τα δεξιά ομαδοποιώντας το p_1 μαζί με το αμέσως επόμενο σύμβολο που το ακολουθεί. Αυτό φαίνεται παρακάτω με την υπογράμμιση:

  $$x = \underline{10} \quad \underline{11} \quad \underline{0} \quad \underline{10} \quad \underline{10} \quad \underline{0} \quad \underline{0} \quad \underline{10}$$

  Παρατηρούμε τώρα ότι $p2 =$`0′ και αφού επαναλαμβάνεται 2 φορές

  $$x = (\underline{10} \quad \underline{11} \quad \underline{0} \quad \underline{10} \quad \underline{10} \quad \overbrace{\underline{0} \quad \underline{0}}^{2} \quad \underline{10})$$

  τότε $k_2 = 2$.

- Τώρα σαρώνουμε την νέα, ομαδοποιημένη από πριν χορδή, από αριστερά. Τώρα $p2 =$`0′ και το ομαδοποιούμε με το αμέσως επόμενο σύμβολο. Επειδή όμως $k_2 = 2$ αν το επόμενο σύμβολο είναι p_2 τότε το ομαδοποιούμε και με το μεθεπόμενο. Τελικά θα έχουμε:

  $$x = \underline{\underline{10}} \quad \underline{\underline{11}} \quad \underline{\underline{0}} \quad \underline{10} \quad \underline{\underline{10}} \quad \underline{\underline{0}} \quad \underline{0} \quad \underline{10}$$

- Η διαδικασία αυτή επαναλαμβάνεται μέχρι να έχουμε πλέον όλη την χορδή ομαδοποιημένη. Αν συνεχίσουμε, δηλαδή, τον αλγόριθμο βρίσκουμε:
  $p_3 =$`10′ με $k_3 = 1$,

  $$x = \underline{\underline{\underline{10}}} \quad \underline{\underline{\underline{11}}} \quad \underline{\underline{\underline{0}}} \quad \underline{\underline{10}} \quad \underline{\underline{\underline{10}}} \quad \underline{\underline{\underline{0}}} \quad \underline{\underline{0}} \quad \underline{\underline{10}}$$

  $p_4 =$`010′ με $k_4 = 1$,

  $$x = \underline{\underline{\underline{10}}} \quad \underline{\underline{\underline{11}}} \quad \underline{\underline{\underline{0}}} \quad \underline{\underline{\underline{10}}} \quad \underline{\underline{\underline{10}}} \quad \underline{\underline{\underline{0}}} \quad \underline{\underline{\underline{0}}} \quad \underline{\underline{\underline{10}}}$$

  $p_5 =$`1010′ με $k_5 = 1$,

  $$x = \underline{\underline{\underline{10}}} \quad \underline{\underline{\underline{11}}} \quad \underline{\underline{\underline{0}}} \quad \underline{\underline{\underline{10}}} \quad \underline{\underline{\underline{10}}} \quad \underline{\underline{\underline{0}}} \quad \underline{\underline{\underline{0}}} \quad \underline{\underline{\underline{10}}}$$

- Τελικά έχουμε από την εξίσωση $C_T(x(n)) = \sum_i^q ln(k_i + 1)$:

  $$C_T(x(n)) = \sum_{i=1}^{5} ln(k_i + 1) = ln(\prod_{i=1}^{5}(k_i + 1)) = ln \quad 48 \approx 3.87$$

- Άρα $I_T(x) \approx 8.7 \quad nats$ και $\overline{h_T}(x) \approx \frac{8.7}{13} \approx 0.669 \quad nats/symbol$

2.2 Συμβολοσειρές και εντροπίες στη ανάλυση φυσικής γλώσσας

2.2.1 Βιβλιογραφική ανασκόπηση

Η φυσική γλώσσα είναι το σημαντικότερο εργαλείο του ανθρώπου για επικοινωνία και πληροφορία. Η θεωρό-πληροφοριακή προσέγγιση στην ανάλυση της λοιπόν, δε θα μπορούσε παρά να είναι απαραίτητη.

Ο πιο άμεσος υπολογισμός της εντροπίας για μία γλώσσα μπορεί να γίνει αν δεχτεί κανείς την ίδια την ακολουθία γραμμάτων ενός κειμένου ως συμβολοσειρά και υπολογίσει τις $n-gram$ εντροπίες με βάση τις πιθανότητες εμφάνισης των γραμμάτων μέσα στο κείμενο. Πρώτη φορά αυτό έγινε το 1951 για την αγγλική γλώσσα από τον C. Shannon [28] και ακολούθησαν άλλοι συγγραφείς για διαφορετικές γλώσσες. Ενδεικτικά, το 1968 οι P. Balasurhamanan & G. Siromoney [1] υπολόγισαν, την εντροπία της γλώσσας των σύγχρονων Tulungu (μία γλώσσα που ομιλείται από 70 εκ. ινδούς και ανήκει στην οικογένεια των Δραβιδικών γλωσσών). Συνέλεξαν 50000 γράμματα από τέσσερις κατηγορίες κειμένων: μυθιστορήματα, μικρές ιστορίες, θεατρικά και 'άλλα' (βιογραφίες, ιστορικά και ετερόκλητα δοκίμια) και από την συχνότητα των γραμμάτων απεφάνθησαν ότι η $n-gram$ εντροπία, για $n = 1$, δεν αποτελεί χαρακτηριστικό του είδους του κειμένου αλλά ούτε και του είδους της γλώσσας, τουλάχιστον μεταξύ διαφορετικών γλωσσών της Ινδίας, όπως είχε αποδείξει έξι χρόνια πριν ο Ramakrishna [24].

Το 1993 οι L.Levitin & Z. Reingold [17], έδωσαν μία νέα μέθοδο για την εκτίμηση του κάτω και άνω φράγματος της εντροπίας του γραπτού λόγου, αναδιατυπώνοντας της μέθοδο του "guessing" που είχε προτείνει ο Shannon το 1951 (Η μέθοδος του "guessing" είναι η εκτίμηση της κατανομής πιθανότητας των λέξεων της αγγλικής γλώσσας μέσω της ποσοτικοποίησης της δυνατότητας ενός ανθρώπου να προβλέψει τον επόμενο χαρακτήρα σε ένα κείμενο) [28] .

Τρία χρόνια αργότερα, οι Rateitschak et. al. [25], πατώντας πάνω στις ίδιες εντροπικές ιδέες, διατύπωσαν ένα μοντέλο γεννήτριας συμβόλων που ήταν σε θέση να αναπαραστήσει τις μεγάλου μήκους συσχετίσεις μεταξύ προτάσεων σε γραπτά κείμενα όπως τις είχε εκτιμήσει ο Ebeling το 1995 [9].

Συσχετίσεις μεγάλου μήκους όμως, επαλήθευσαν και οι Montemurro & Pury [19] για την αγγλική γλώσσα ενώ οι Bhan et. al. [5] για την κορεατική, υπολογίζοντας τον εκθέτη Hurst λογοτεχνικών κειμένων. Ακολούθησαν οι Sahin et. al. για την τουρκική γλώσσα κάνοντας χρήση της μεθόδου detrended fluctuation analysis [10].

Μία ενδιαφέρουσα, από άποψη υπολογιστικής γλωσσολογίας, εργασία, ήταν αυτή των Behr et. al., όπου υπολόγισαν την εντροπία ίδιων κειμένων μεταφρασμένων σε εννέα διαφορετικές γλώσσες και συμπιεσμένων με την τεχνική της πρόβλεψης με μερική ταύτιση (PPM compression). Οι συγγραφείς κατέληξαν στο ότι το μέγεθος του κειμένου μετά την συμπίεση και η τιμή της εντροπίας αυτού είναι ανεξάρτητες της γλώσσας. Το αποτέλεσμα αυτό τους επέτρεψε να προτείνουν τη τεχνική της PPM compression για ανίχνευση κακής μηχανικής μετάφρασης [3].

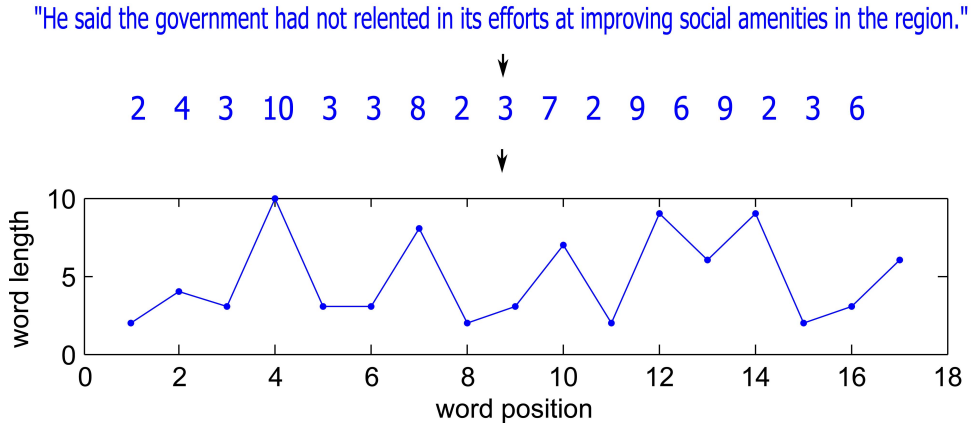
Αν και ένα κείμενο φυσικού γραπτού λόγου σε αντιπαραβολή με μία χρονοσειρά αριθμών να είναι σε πρώτη άποψη δύο ανοίκειες έννοιες, δεν θέλει πολύ φαντασία προκειμένου για κάποιον ο οποίος θέλει να αναλύσει μαθηματικά ή υπολογιστικά ένα κείμενο, να βρει ένα τρόπο απεικόνισης του σε μία χρονοσειρά.
Πράγματι, το 2006 οι Kosmidis et. al. [16], στην εργασία τους με τίτλο "Language time series analysis", αναπαρέστησαν ένα γλωσσικό κείμενο με τη μορφή χρονοσειράς αριθμών με δύο διαφορετικούς τρόπους. Ο πρώτος ήταν με μεταβλητή το μήκος μιας λέξης και ο δεύτερος με μεταβλητή τη συχνότητας μίας λέξης, ενώ ο «χρόνος» και στις δύο περιπτώσεις ήταν η θέση της λέξης αυτής μέσα στο κείμενο. Χρησιμοποιώντας τις μεθόδους detrended fluctuation analysis (DFA) και Grassberger-Proccacia analysis (GP), και μέσω τις σύγκρισης με τυχαία σήματα, οι συγγραφείς κατάφεραν να αναδείξουν την γλωσσολογική πολυπλοκότητα των αντίστοιχων χρονοσειρών, παρά το τεράστιο όγκο πληροφορίας που αποκλείει η αναπαράστασή τους.

Λίγα χρόνια αργότερα, οι Papadimitriou et. al. [21] βρήκαν τις Shannon και Kolmogorov εντροπίες ευαίσθητες μεταξύ διαφορετικών γλωσσών (Ελληνικών και Αγγλικών) και διαφορετικών ειδών κειμένου (λογοτεχνία, οικονομικά, πολιτικά και αθλητικά) κάνοντας χρήση μιας δυαδικής αναπαράστασης του γραπτού λόγου. Επιπλέον, συνέκριναν τις εντροπίες αυτές με τις αντίστοιχες στοχαστικών συμβολοσειρών προκειμένου να εξετάσουν την επίδραση συσχετίσεων μικρού μήκους σε αυτές.

Ιστορικά, έχουν επιχειρηθεί και άλλοι τρόποι δημιουργίας γλωσσολογικών χρονοσειρών όπως η παραμετροποίηση των λέξεων εκχωρώντας συγκεκριμένες αριθμητικές τιμές σε κάθε γράμμα [10]. Κάτι τέτοιο μπορεί να γίνει για παράδειγμα με χρήση του κώδικα ASCII ή του αύξοντα αριθμού ενός γράμματος μέσα στο αλφάβητο.

2.2.2 Η μέθοδός μας

Στην παρούσα εργασία, κινούμαστε μεθοδολογικά εμπνευσμένοι από τους Kosmidis et. al. και Papadimitriou et. al. [21]. Αναπαριστούμε τα κείμενά μας με την μορφή χρονοσειρών μήκους λέξεων με τον εξής τρόπο.
Απεικονίζουμε το κάθε απόσπασμα σε μία συμβολοσειρά ακεραίων, με κάθε ακέραιο να αναπαριστά το μήκος της αντίστοιχης λέξης (καλούμε εδώ τις συμβολοσειρές και χρονοσειρές, όπου και για εμάς εδώ ο «χρόνος» είναι η θέση της κάθε λέξης στο κείμενο). Μία τέτοια χρονοσειρά εμφανίζει προφανώς ελάχιστο ίσο με 1 και μέγιστο ίσο με το μήκος της μεγαλύτερης λέξης μέσα στο κείμενο. Ένα παράδειγμα της διαδικασίας μπορεί να γίνει αντιληπτό με τη βοήθεια του παρακάτω σχήματος (Σχήμα 2). Ο ιδιαίτερος χαρακτήρας των χρονοσειρών αυτών, μας απαλλάσσει από την ανάγκη μίας περαιτέρω συμβολικής δυναμικής. Η δική μας ανάλυση πάνω στις χρονοσειρές αυτές είναι εντροπικής φύσης.



Σχήμα 2: Απεικόνιση μίας πρότασης σε συμβολοσειρά μήκους λέξεων

3 ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Η ανάλυση έχει γίνει για δύο διαφορετικά σώματα κειμένων τα οποία περιγράφονται παρακάτω.

3.1 Πρώτο σώμα κειμένων - Ειδησεογραφικά άρθρα και λογοτεχνία

Αυτό το σώμα κειμένων αποτελείται από 5.4M λέξεων και περιλαμβάνει δύο γλώσσες, ελληνικά και αγγλικά. Τα δύο βασικά είδη κειμένων είναι ειδησεογραφικά άρθρα (πολιτικά, οικονομικά και αθλητικά) και λογοτεχνία. Τα άρθρα είναι κείμενα δανεισμένα από το διαδίκτυο, που ανασύρθηκαν από μεγάλες, ελληνικές και αγγλικές ειδησεογραφικές ιστοσελίδες και έχουν γραφεί από διαφορετικούς συγγραφείς την περίοδο 1η Ιανουαρίου με 31η Μαρτίου του 2008. Το σύνολο των διαδικτυακών δεδομένων περιλαμβάνει εκτός από τα ίδια τα κείμενα και ένα πλήθος μεταδεδομένων όπως ημερομηνία και ώρα δημοσίευσης, ημερομηνία και ώρα τροποποίησης, όνομα συγγραφέα, τίτλο άρθρου, URL διασύνδεση, ετικέτες και κατηγορίες κειμένου. Τα μεταδεδομένα αυτά έχουν σχηματοποιηθεί σε XML μορφή.
Τα λογοτεχνικά κείμενα είναι αποσπάσματα από ηλεκτρονικά βιβλία στην αγγλική γλώσσα ή από βιβλία του σώματος κειμένων του Ινστιτούτου Επεξεργασίας του Λόγου (ΙΕΛ)[1]. (Πίνακας 1)

[1] http://hnc.ilsp.gr/en/

	# of English words	# of Greek words	Total # of words
Web - Politics	0.84M	0.76M	1.6M
Web - Economy	0.84M	0.74M	1.4M
Web - Sports	0.89M	0.79M	1.7M
Literary works	0.42M	0.25M	0.7M
Total	2.99M	2.54M	5.4M

Πίνακας 1: Ταξινόμηση του πρώτου σώματος των κειμένων σε γλώσσες και είδη

3.2 Δεύτερο σώμα κειμένων - Κείμενα ευρωπαϊκού κοινοβουλίου

Εδώ η ανάλυση γίνεται για κείμενα πρακτικών του Ευρωπαϊκού κοινοβουλίου επεκτείνοντας έτσι την μελέτη σε περισσότερες των δύο γλωσσών. Πρόκειται για ένα τεράστιο παράλληλο σώμα κειμένων σε 11 γλώσσες που έχει ανασυρθεί από την ιστοσελίδα του Ευρωπαϊκού κοινοβουλίου από τον Philipp Koehn (University of Edinburgh)[2] [32]. Τα πρακτικά ξεκινούν από την 1η Ιανουαρίου του 2000.
Στην παρούσα ανάλυση χρησιμοποιήθηκαν 10 διαφορετικές γλώσσες (ολλανδικά, αγγλικά, φινλανδικά, γαλλικά, γερμανικά, ελληνικά, ιταλικά, πορτογαλικά, ισπανικά και σουηδικά) και περίπου δεκαπέντε εκατομμύρια λέξεις ανά γλώσσα (Πίνακας 2).

4 ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Πρώτο σώμα κειμένων - Ειδησεογραφικά άρθρα και λογοτεχνία

Έχοντας στη διάθεσή μας το πρώτο σώμα κειμένων (ειδησεογραφικά άρθρα και λογοτεχνικά κείμενα), κατασκευάζουμε τις χρονοσειρές μήκους λέξεων με τον τρόπο που περιγράφηκε στην ενότητα 2.2.2. Όλες οι χρονοσειρές που αντιστοιχούν στο ίδιο είδος κειμένου, συνδέονται αλυσιδωτά ώστε να προκύψει μία χρονοσειρά ανά κατηγορία κειμένων. Με αυτό τον τρόπο επιδιώκεται η εξάλειψη της πόλωσης των κειμένων που μπορεί να οφείλεται στην ιδιαιτερότητα α) διαφορετικών συγγραφέων και β) της ιστοσελίδας από την οποία λήφθηκε το κείμενο όπως και στο θέμα το οποίο πραγματεύεται το κάθε κείμενο.

[2]http://opus.lingfil.uu.se/Europarl3.php

Language	# of words
Dutch	16.23M
English	15.98M
Finnish	10.15M
French	17.55M
German	14.68M
Greek	15.85M
Italian	15.57M
Portuguese	16.06M
Spanish	16.22M
Swedish	14.23M
Total	152.52M

Πίνακας 2: Ταξινόμηση του παράλληλου σώματος των κειμένων του ευρωκοινοβουλίου σε γλώσσες

Ξεκινάμε την ανάλυσή μας υπολογίζοντας την προσεγγιστική εντροπία (approximate entropy). Υπενθυμίζουμε ότι αυτή ορίζεται ως εξής:

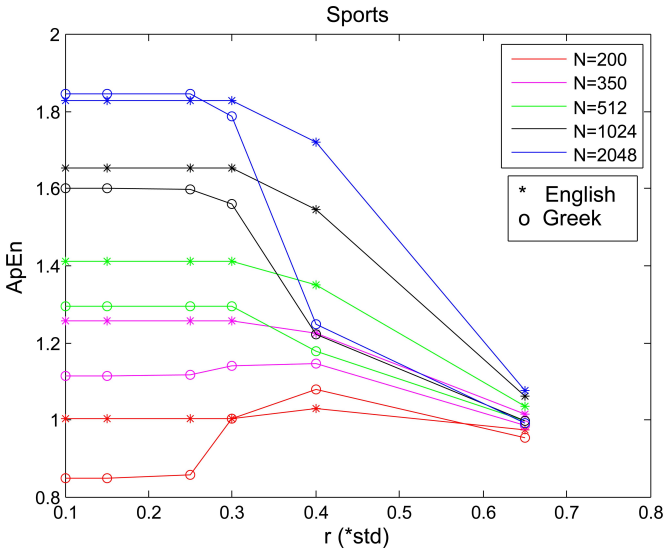
$$ApEn(n, r, N) = F^n(r) - F^{n+1}(r)$$

όπου

$$F^n(r) = \sum_i \frac{ln(C_i^n(r))}{(N-n+1)}, \quad C_i^n(r) = \frac{\text{No. of j such that } d(x_i, x_j) < r}{N-n+1}$$

με $i = 1, ..., N-n+1$ και r το κριτήριο «ανεκτικότητας» με την οποία κάποιος ορίζει δύο $n-grams$ ως όμοια. Εδώ, για λόγους καλύτερης στατιστικής, υπολογίζουμε τις εντροπίες σε υπακολουθίες μήκους N με το τελικό αποτέλεσμα να είναι η μέση τιμή για όλες τις υπακολουθίες που ανήκουν στο ίδιο είδος κειμένου. Συμφωνώντας με την βιβλιογραφία, επιλέγουμε $n = 2$ και δοκιμάζουμε διάφορες τιμές για την ανεκτικότητα r, ορίζοντάς τήν πάντα ως διαφορετικά ποσοστά της τυπικής απόκλισης της κάθε υπακολουθίας μήκους N. Κατά την προσπάθεια υπολογισμού της $ApEn$ ανακαλύπτουμε ότι οι σχετικές διαφορές μεταξύ των γλωσσών δεν παραμένουν σταθερές αλλά εξαρτώνται με έναν ιδιόρρυθμο τρόπο από τα N και r. Κάτι τέτοιο μπορεί να φανεί στο Σχήμα 3 όπου σχεδιάζεται, ενδεικτικά για τα αθλητικά άρθρα, η τιμή της προσεγγιστικής εντροπίας ως προς r και για διάφορα μήκη N.

Είναι σαφές ότι δεν μπορούν να προκύψουν στέρεα συμπεράσματα, μέχρι να παρατηρήσουμε ότι από μία τιμή της r και κάτω (περίπου $r < 0.28 * std$) και λόγω του ακεραίου των τιμών των χρονοσειρών, ο αλγόριθμος συγκρίνει πανομοιότυπα και όχι «όμοια» διανύσματα μήκους n και



Σχήμα 3: Πρώτο σώμα κειμένων: $ApEn$ για τα αθλητικά άρθρα, για διαφορετικές τιμές της ανεκτικότητας r και διάφορα μήκη υποακολουθίας N.

$n+1$. Αποφασίζουμε να μειώσουμε το, υπό μία έννοια, συστηματικό αυτό σφάλμα περιορίζοντας την τιμή της r σε $0.25 * std$. Στον Πίνακα 3 παρουσιάζονται τα αποτελέσματα για $N = 1000$.

Βλέπουμε ότι οι τιμές της $ApEn$ για τα ελληνικά ειδησεογραφικά άρθρα (πολιτικά, οικονομικά και αθλητικά) είναι ελάχιστα μικρότερες από τις αντίστοιχες των αγγλικών. Ανακαλούμε την φυσική σημασία αυτής της εντροπίας· μικρές τιμές $ApEn$ υπονοούν την δυνατότητα σύμπτυξης των δεδομένων χωρίς την απώλεια ουσιαστικής πληροφορίας (high compressibility). Κάτι τέτοιο θα μπορούσε να σημαίνει ότι μία πληροφορία στην γλώσσα που έχει σταθερά μικρότερη τιμή $ApEn$ αναπτύσσεται με ένα 'πλεονάζοντα' τρόπο σε σχέση με την άλλη γλώσσα, πράγμα που μπορεί να συνεπάγεται πλουσιότερο λεξιλόγιο ή μεγαλύτερη ποικιλία στη δομή του λόγου. Εν τούτοις τα ελληνικά λογοτεχνικά κείμενα εμφανίζουν μεγαλύτερη εντροπία από τα αντίστοιχα αγγλικά, αφήνοντας μας με το συμπέρασμα ότι η $ApEn$ υπολογισμένη σε αυτή την αναπαράσταση γραπτών κειμένων δεν μπορεί να αποτελεί χαρακτηριστικό του είδους της γλώσσας.

Αποφασίζουμε να επιστρέψουμε σε μία κλασικότερη αντιμετώπιση των χρονοσειρών μας, αυτή των block εντροπιών, ή αλλιώς εδώ $n-gram$ εντροπίες, όπως ορίστηκαν στην ενότητα 2.1.2. Υπενθυμίζοντας

$$\Phi_n = -\sum_{(s_1,\ldots,s_n)} p^{(n)}(s_1,\ldots,s_n)\ln(p^{(n)}(s_1,\ldots s_n))$$

Η άθροιση γίνεται πάνω σε όλα τα μοναδικά $n-grams$, ενώ

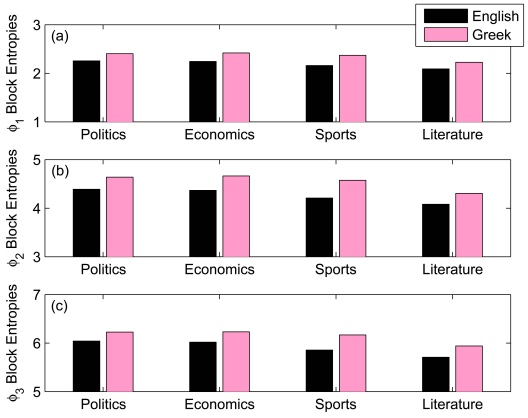
$$p^{(n)}(s_1,\ldots,s_n) = \frac{\text{No. of n-grams } (s_1,\ldots,s_n), \text{encountered when lumping or gliding}}{\text{Total No. of n-grams}}$$

	ApEn	
	English	Greek
Politics	1.6600	1.5981
Economics	1.6606	1.5780
Sports	1.6529	1.5998
Literature	1.6284	1.6345

Πίνακας 3: Τιμές $ApEn$ για $r = 0.25$ και $N = 1000$

είναι η πιθανότητα να συναντήσουμε ένα συγκεκριμένο $n - gram$.

Ξεκινάμε υπολογίζοντας αυτές τις ποσότητες για $n = 1, 2$ και 3 σε υπακολουθίες των 1000 λέξεων. Υπενθυμίζουμε ότι οι $n - gram$ εντροπίες για $n = 1$ δίνουν την κλασική εντροπία του Shannon. Η τελική τιμή της εντροπίας για κάθε συμβολοσειρά, προκύπτει και εδώ ως η μέση τιμή των εντροπιών για όλες τις υπο-ακολουθίες.
Παρακάτω φαίνονται τα αποτελέσματα (Σχήμα 4).

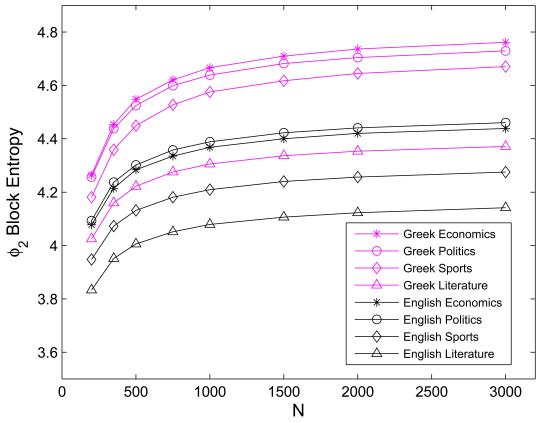


Σχήμα 4: Πρώτο σώμα κειμένων: Ιστογράμματα των unigram Φ_1 (a), bigram Φ_2 (b) και trigram Φ_3 (c), για τα ελληνικά και τα αγγλικά και για όλα τα είδη κειμένων.

Βλέπουμε ότι οι μέσες τιμές των Φ_1, Φ_2 και Φ_3 εντροπιών είναι εδώ σταθερά μεγαλύτερες για τα ελληνικά από ό,τι για τα αγγλικά. Επίσης, εδώ μπορούμε να ξεχωρίσουμε με βεβαιότητα, ότι τα οικονομικού και πολιτικού περιεχομένου κείμενα παρουσιάζουν μεγαλύτερες τιμές εντροπίας από τα κείμενα αθλητικού περιεχομένου τα οποία ακολουθούνται με τη σειρά τους από τα λο-

γοτεχνικά.

Προκειμένου να εξετάσουμε την εξάρτηση των Φ_n από το μήκος της υπο-ακολουθίας N, κάνουμε τους ίδιους υπολογισμούς για διαφορετικά μήκη N ($250 < N < 3000$). Στο Σχήμα 5 φαίνονται τα σχετικά αποτελέσματα για τις τιμές της Φ_2, για την περίπτωση αγγλικών και ελληνικών και για όλα τα είδη κειμένων. Σχολιάζουμε ότι ενώ η εντροπία εμφανίζει εξάρτηση από το N, ο

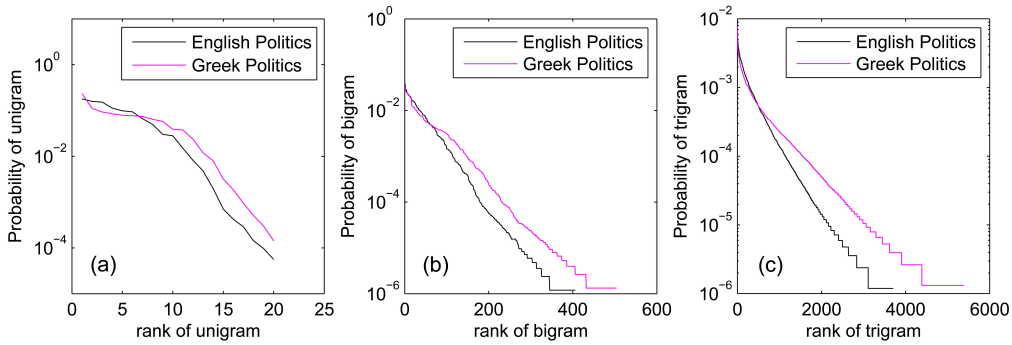


Σχήμα 5: Πρώτο σώμα κειμένων: Φ_2 εντροπία ως προς το μήκος N της υποακολουθίας.

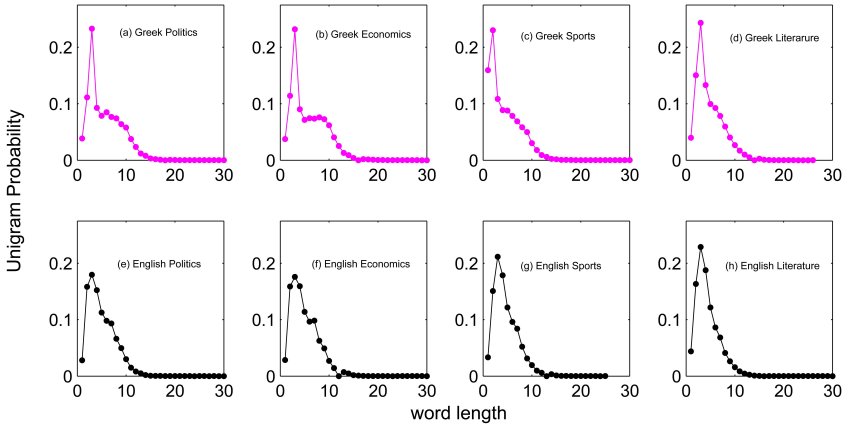
τρόπος με τον οποίο αυτή μεταβάλλεται παραμένει ανεξάρτητος του είδους του κειμένου. Τα συμπεράσματα αυτά επαληθεύονται και για τις υπόλοιπες των περιπτώσεων.

Ως γνωστών οι $n-gram$ εντροπίες αντανακλούν τη συμπεριφορά των κατανομών πιθανότητας του υπό μελέτη μεγέθους και μάλιστα όσο πιο ομοιόμορφη η κατανομή τόσο πιο μεγάλες και οι τιμές της εντροπίας. Αναμένουμε λοιπόν, μετά από την παραπάνω ανάλυση, να έχουμε περισότερη ομοιομορφία στις κατανομές πιθανότητας των $n-grams$ για τα ελληνικά κείμενα. Το Σχήμα 6, απεικονίζει την κατανομή πιθανότητας των unigrams, bigrams και trigrams ($n = 1, 2$ και 3 αντίστοιχα) ως προς την αύξουσα σειρά τους (όπως αυτή ορίζεται ως προς τη πιθανότητα εμφάνισής τους) για την περίπτωση των ελληνικών και αγγλικών πολιτικών κειμένων. Ανάλογα αποτελέσματα δίνουν και τα υπόλοιπα είδη κειμένων για την ελληνική και αγγλική γλώσσα.

Αν εστιάσουμε στα unigrams, Σχήμα 6 (a), βλέπουμε ότι η πιθανότητα μειώνεται αργά, εμφανίζοντας ένα πλατό για τα πρώτα σε κατάταξη unigrams, ενώ συνεχίζει με μία εκθετική μείωση στα δεύτερα. Η συμπεριφορά αυτή είναι κοινή και για τις δύο γλώσσες, εν τούτοις το πλατό είναι πολύ περισσότερο εμφανές για την περίπτωση των ελληνικών. Μπορούμε έτσι να εκτιμήσουμε ότι αυτό το πλατό είναι ο κυριότερος λόγος για τις μεγαλύτερες τιμές της Φ_1 στη γλώσσα αυτή. Σε αυτό το σημείο και προκειμένου να εντοπίσουμε αν πρόκειται για κάποια συγκεκριμένα μήκη λέξεων στα οποία οφείλουμε αυτή τη διαφορά στις τιμές της Φ_1, μελετάμε τη κατανομή πιθανότητας των unigram ως προς το μήκος των λέξεων (Σχήμα 7).



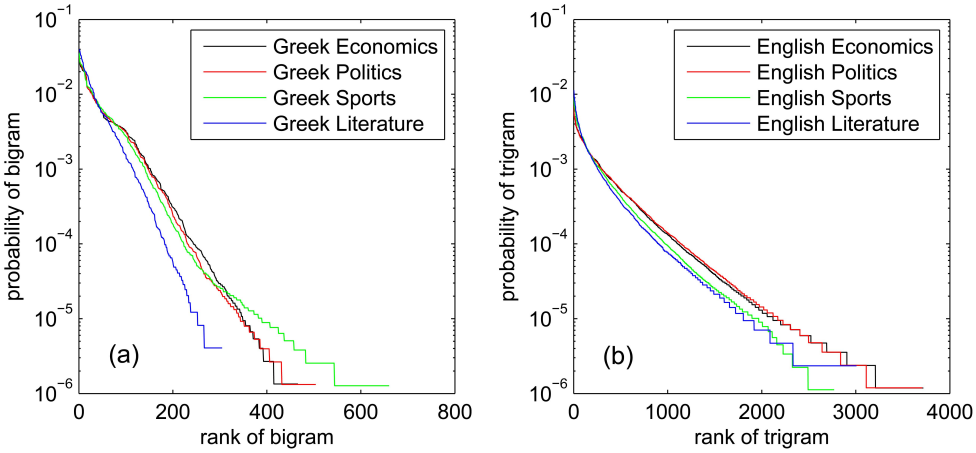
Σχήμα 6: Πρώτο σώμα κειμένων: Κατανομή πιθανότητας των $n-grams$ ως προς τη αύξουσα σειρά τους (όπως αυτή ορίζεται ως προς τη πιθανότητα εμφάνισής τους).



Σχήμα 7: Πρώτο σώμα κειμένων: Κατανομή πιθανότητας των unigrams ως προς το μήκος της λέξης.

Από το σχήμα αυτό γίνεται φανερή η συνεισφορά στην τιμή της Φ_1 εντροπίας των λέξεων με μήκη 5-10 καθώς υπάρχει σαφής διαπλάτυνση των κατανομών στα μήκη αυτά. Παρατηρούμε μάλιστα ότι το μήκος αυτού του πλατό είναι κατά αναλογία και με τις τιμές των Φ_1 όσον αφορά το είδος του κειμένου, διευρυμένο για τα οικονομικά και πολίτικά ενώ σταδιακά συρρικνώνεται όπως παίρναμε από τα αθλητικά προς τη λογοτεχνία. Επιστρέφοντας ξανά στο Σχήμα 6(b), παρατηρούμε ότι το αντίστοιχο πλατό για την περίπτωση των αγγλικων bigrams έχει εντελώς εξαφανιστεί, και αιτιολογούμε την διαφορά στις τιμές των Φ_2 εντροπιών από την αργότερη πτώση της κατανομής των ελληνικών bigrams. Προχωρώντας στο Σχήμα 6(c) ανακαλύπτουμε ότι το πλατό έχει τώρα εξαφανιστεί και για τις δύο γλώσσες ενώ η πτώση των κατανομών είναι αργότερη από εκθετική. Η διαφορά στις τιμές της Φ_3 για τα ελληνικά και τα αγγλικά μπορεί να αιτιολογηθεί από την γρηγορότερη (αργότερη) πτώση της κατανομής στις μικρές (μεγάλες) θέσεις των trigrams. Συμπερασματικά, σε αυτό το σημείο της ανάλυσης μπορούμε να πούμε ότι όταν η κατανομή εμφανίζει πλατό, η τιμή της εντροπίας καθορίζεται από το πλάτος αυτού. Στην

απουσία του πλατό, η εντροπία εξαρτάται από το πόσο γρήγορα πέφτει η κατανομή (Σχήμα 8).



Σχήμα 8: Πρώτο σώμα κειμένων: Κατανομή πιθανότητας των ελληνικών bigrams και αγγλικών trigrams ως προς τη αύξουσα σειρά τους (όπως αυτή ορίζεται ως προς τη πιθανότητα εμφάνισής τους) για διαφορετικά είδη κειμένων.

Έως τώρα, είδαμε ότι η συμπεριφορά των Φ_2 και Φ_3 για την περίπτωση των αγγλικών και ελληνικών κειμένων, είναι ανάλογη με αυτή της Φ_1 για τα ίδια κείμενα. Σαφώς λοιπόν, η Φ_1 εντροπία που ανταναλκλά την κανονικότητα της κατανομής πιθανότητας των unigrams, επηρεάζει τις τιμές των Φ_2 και Φ_3. Αλλά σε τι βαθμό παίζει ρόλο η συσχέτιση των ψηφίων (μήκη λέξεων) μέσα σε ένα bigram ή trigram για τις τιμές αυτών;
Προχωράμε, λοιπόν, σε αναδιάταξη των χρονοσειρών με τυχαίο τρόπο (shuffling), και εκτελούμε την ίδια επεξεργασία προκειμένου να μπορέσουμε να αναγνωρίσουμε την επίδραση των συσχετίσεων μεταξύ μηκών γειτονικών λέξεων στις τιμές των Φ_2 και Φ_3. Τονίζουμε ότι μία τυχαία αναδιαταγμένη (shuffled) συμβολοσειρά δεν εμπεριέχει συσχετίσεις των θέσεων των λέξεων στην αναπαράσταση του μήκους τους και έχει την ίδια κατανομή unigrams με τη σειρά από την οποία έχει προέλθει. Οι τιμές των $n-gram$ εντροπιών για τις τυχαία αναδιαταγμένες συμβολοσειρές φαίνονται στον Πίνακα 4.

Από τον πίνακα 4 διακρίνουμε μία μικρή αλλά σταθερή υπεροχή στις τιμές της εντροπίας των τυχαίων συμβολοσειρών. Το γεγονός ότι η διαφορά είναι μικρή καταδεικνύει ότι η βασική διαφορά μεταξύ των δύο γλωσσών οφείλεται στις κατατομές των μηκών των λέξεων (unigrams). Με άλλα λόγια δηλαδή η Φ_1 είναι υπεύθυνη για τις διαφορές των εντροπιών Φ_2 και Φ_3 των δύο γλωσσών. Εντούτοις, η σταθερά μεγαλύτερη τιμή των Φ_2 και Φ_3 των τυχαίων συμβολοσειρών μαρτυρά ότι υπάρχει και μία μικρή εξάρτηση από τις συσχετίσεις μεταξύ των μηκών γειτονικών λέξεων.

Σε αυτό το σημείο αξίζει να γίνει μία αντιπαράθεση των αποτελεσμάτων με εκείνα των Papadimitriou et. al. [21]. Όπως έχει αναφερθεί παραπάνω, η δική τους συμβολική αναπαράσταση των ίδιων κειμένων ήταν δυαδικής μορφής. Πιο συγκεκριμένα κάθε γράμμα απεικονίστηκε στον

	PolE	PolG	EcoE	EcoG	SpoE	SpoG	LitE	LitG
Φ_2	4.389	4.639	4.366	4.664	4.209	4.575	4.079	4.306
Φ_2(shuffled)	4.438	4.707	4.416	4.735	4.254	4.650	4.149	4.375
Φ_3	6.041	6.227	6.018	6.232	5.856	6.166	5.708	5.940
Φ_3(shuffled)	6.115	6.301	6.095	6.320	5.923	6.262	5.807	6.021

Πίνακας 4: Φ_2 και Φ_3 εντροπίες για τις αρχικές συμβολοσειρές μήκους λέξεων καθώς και για τις τυχαία αναδιαταγμένες (Shuffled).

ψηφίο 1, ενώ τα κενά και σημεία στίξης στο ψηφίο 0. Επιπλέον όλα τα διαδοχικά μηδενικά συμπτύχθηκαν σε ένα μοναδικό μηδενικό (με άλλα λόγια η προκύπτουσα συμβολοσειρά αποτελούταν από άσσους και μηδενικά με τον περιορισμό της μη ύπαρξης συνεχόμενων μηδενικών). Με αυτό τον τρόπο το μήκος των συνεχόμενων 1 αναπαριστούσε το μήκος της κάθε λέξης και η συζήτησή τους βασίστηκε επίσης σε κατανομές μήκους λέξεων και συσχετίσεις μικρού μήκους μέσα στις συμβολοσειρές. Το ενδιαφέρον σημείο όμως είναι η τιμή των εντροπιών σε σχέση με τον είδος της γλώσσας και του κειμένου. Για την δική μας αναπαράσταση, τα ελληνικά εμφανίζουν μεγαλύτερη εντροπία από τα αγγλικά ενώ τα πολιτικά και οικονομικά παίρνουν μεγαλύτερες τιμές εντροπίας από τα αθλητικά τα οποία ακολουθούνται με τη σειρά τους από τα λογοτεχνικά. Για όλες τις περιπτώσεις (ειδος γλώσσας και κειμένου) τα αποτελέσματα των Papadimitriou et. al. είναι αντεστραμμένα σε σχέση με τα δικά μας· τα αγγλικά έχουν μεγαλύτερη εντροπία από τα ελληνικά ενώ η ίδια αντιστροφή ισχύει και για τις σχέσεις των εντροπιών των διαφόρων ειδών κειμένου.
Αν και στη προκειμένη περίπτωση, η συμμετρία στην αντιστροφή των εντροπιών, δείχνει μία συνέπεια για τα αποτελέσματα των δύο εργασιών, θα πρέπει γενικότερα να τονιστεί η επίδραση της συμβολικής αναπαράστασης ενός συστήματος στην τιμή της εντροπίας του υπό μελέτη μεγέθους.

Συμπερασματικά, για το πρώτο σώμα κειμένων θα μπορούσαμε να τονίσουμε τη σημασία των λέξεων με μήκη 5-10 γράμματα στην τελική τιμή της εντροπίας. Για κείμενα ίδιου είδους η πυκνότητα πιθανότητας των μηκών λέξεων γύρω από τις τιμές 5-10 είναι περισσότερο πεπλατυσμένη στα ελληνικά από ότι στα αγγλικά. Κάτι τέτοιο αντανακλά βασικές γλωσσολογικές ιδιότητες της ελληνικής γλώσσας, το πλούτο της κλιτικής μορφολογίας (το σύνολο των διαφορετικών μορφών που μπορεί να λάβει ένα ουσιαστικό ή ένα ρήμα) και τα διαφορετικά μέρη του λόγου που προκύπτουν από ένα μόνο λήμμα (ρήμα, επίρρημα, επίθετο κ.τ.λ.) Με τους δικούς μας όρους, το προηγούμενο διατυπώνεται ως εξής: από μία ρίζα παράγονται πολλές λέξεις με διαφορετικά μήκη. Ένα παράδειγμα σε αντιπαραβολή με τα αγγλικά, δίνεται παρακάτω:

ουσιαστικό: παίχτης

παίχτης = player
παίχτη = (of the) player

λήμμα: παίζ-
παίζω = play
έπαιζα = played
εμπαίζω = deceive
παιγνίδι = game
εμπαιγμός = deception
παιγνιώδης = jocular
εμπαικτικός = mocking

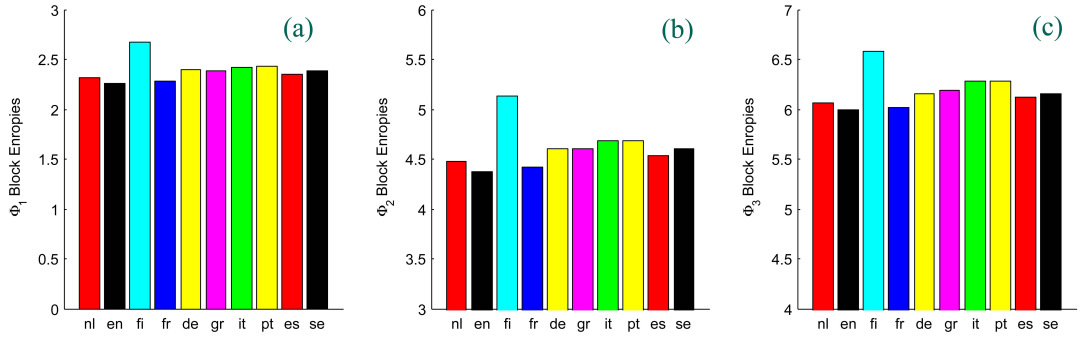
Βλέπουμε στο παραπάνω παράδειγμα ότι οι περισσότερες ελληνικές λέξεις διαφορετικού μήκους που προέρχονται από το ίδιο λήμμα, αντιστοιχούν σε αγγλικές λέξεις που δεν έχουν κοινή ρίζα. Μπορούμε εύκολα να υποθέσουμε ότι αυτό έχει αντίκτυπο για μήκη λέξεων από 5 έως 10 γράμματα αν και η σαφής επαλήθευση του μένει για μελλοντική εργασία. Έτσι λοιπόν, αν κρατήσουμε το ίδιο λεξιλόγιο για τις δύο γλώσσες, τα ελληνικά παρουσιάζουν ένα έντονο πλατό στην κατανομή πιθανότητας των μηκών λέξεων γύρω από τα εν λόγω μήκη. Ας συμπληρώσουμε ότι για την ίδια γλώσσα το είδος του κειμένου καθορίζει το εύρος του πλατό, με αυτό να μειώνεται σταδιακά όπως προχωράμε από τα πολιτικά και τα οικονομικά στα αθλητικά και λογοτεχνικά κείμενα.

4.2 Δεύτερο σώμα κειμένων - Κείμενα ευρωπαϊκού κοινοβουλίου

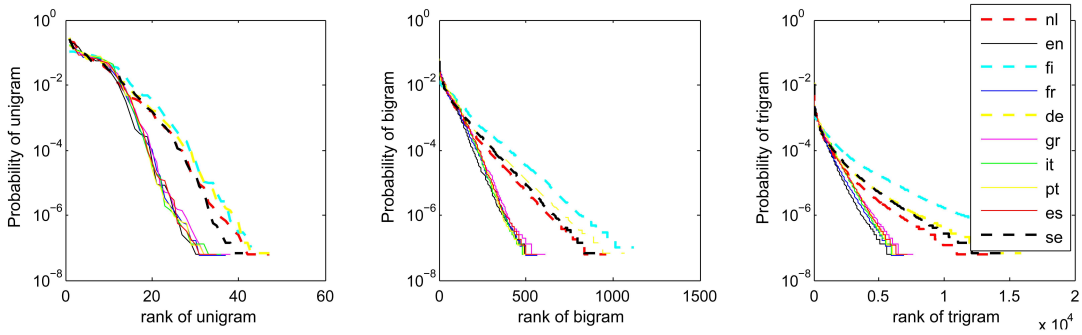
Στην ανάλυση των κειμένων του ευρωκοινοβουλίου αγνοούμε για ευνόητους λόγους την προσεγγιστική εντροπία και προχωρούμε κατευθείαν στον υπολογισμό των Φ_1, Φ_2 και Φ_3 εντροπιών. Στο Σχήμα 9 παρουσιάζεται η διάταξη των τιμών των τριών εντροπιών για την κάθε γλώσσα.

Και για τις τρείς περιπτώσεις την μεγαλύτερη εντροπία παρουσιάζουν τα φινλανδικά με μεγάλη διαφορά από τις υπόλοιπες γλώσσες, ενώ την μικρότερη τα αγγλικά και γαλλικά. Ας σχολιάσουμε την καλύτερη διακριτικότητα της Φ_2 καθώς φαίνεται να αναδεικνύει καλύτερα τις σχετικές διαφορές μεταξύ των γλωσσών (κάτι που μπορεί να οφείλεται στο μέγεθος συσχετίσεων μεταξύ μηκών γειτονικών λέξεων).

Ανατρέχουμε και εδώ στην κατανομή πιθανότητας των $n-grams$ (για $n = 1, 2$ και 3) ως προς την αύξουσα σειρά τους (όπως αυτή ορίζεται ως προς τη πιθανότητα εμφάνισής τους) (Σχήμα 10) και την κατανομή πιθανότητας ως προς το μήκος της λέξης (Σχήμα 11).
Εκτός από την αναμενόμενη εμφανή ιδιομορφία της φινλανδικής γλώσσας, ανακαλύπτουμε και ένα επιπλέον χαρακτηριστικό που δεν είχε γίνει φανερό από τις τιμές τις εντροπίας· στο Σχήμα 10 φαίνεται μία σαφής οργάνωση των γλωσσών σε 3 ομάδες (ειδικά αν παρατηρήσει κανείς την κατανομή των trigrams). Οι 3 αυτές ομάδες ξεχωρίζουν από τον διαφορετικό ρυθμό πτώσης της



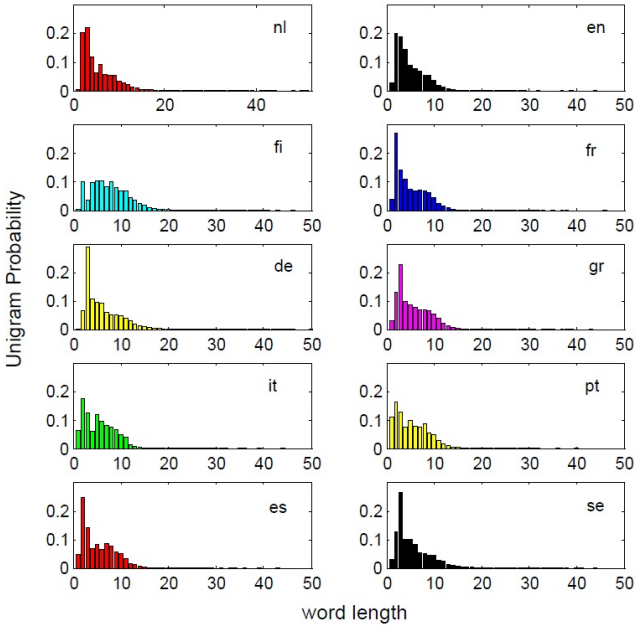
Σχήμα 9: Δεύτερο σώμα κειμένων: Ιστογράμματα των unigram Φ_1 (a), bigram Φ_2 (b) και trigram Φ_3 (c), για όλα τα κείμενα του ευρωκοινοβουλίου.



Σχήμα 10: Δεύτερο σώμα κειμένων: Κατανομή πιθανότητας των $n-grams$ ως προς τη αύξουσα σειρά τους (όπως αυτή ορίζεται ως προς τη πιθανότητα εμφάνισής τους).

κατανομής στα σπανιότερα μήκη λέξεων (ή unigrams και trigrams). Από το Σχήμα 11 γίνεται φανερό ότι τα μικρότερα μήκη λέξεων είναι και τα συνηθέστερα (με εξαίρεση τα φινλανδικά) ενώ τα μεγαλύτερα μήκη λέξεων τα σπανιότερα. Στη πρώτη ομάδα, με την αργότερη πτώση της κατανομής στα μεγάλα μήκη λέξεων ανήκει η φινλανδική γλώσσα (Σχήμα 10). Η δεύτερη ομάδα με την ενδιάμεση σε ρυθμό πτώση της κατανομής περιλαμβάνει τις λεγόμενες γερμανικές γλώσσες εκτός της αγγλικής (ολλανδικά, γερμανικά και σουηδικά) ενώ η τρίτη ομάδα με την γρηγορότερη πτώση της κατανομής πιθανότητας περιλαμβάνει όλες τις υπόλοιπες (αγγλικά, γαλλικά, ελληνικά, ιταλικά, πορτογαλικά και ισπανικά).

Αυτή η δεύτερη ανάλυση μεταξύ πολλών διαφορετικών γλωσσών, μας βοηθάει να εκτιμήσουμε την συνεισφορά των πλατό στην τιμή της εντροπίας. Ας πάρουμε για παράδειγμα την ολλανδική και την ελληνική γλώσσα. Τα ολλανδικά έχουν μικρότερη τιμή εντροπίας από τα ελληνικά αν και η πτώση της κατανομή τους στην περίπτωση των bigrams και trigrams γίνεται με αργότερο ρυθμό. Αυτό οφείλεται στο εντονότερο πλατό των ελληνικών στις πρώτες θέσεις των unigrams. Διαφορετικά, η εντροπία είναι ευαίσθητη κυρίως στο ισοπίθανο της κατανομής για τις λέξεις με



Σχήμα 11: Δεύτερο σώμα κειμένων: Κατανομή πιθανότητας των unigrams ως προς το μήκος της λέξης.

τη μεγαλύτερη πιθανότητα οι οποίες είναι, με εξαίρεση τα φινλανδικά, λέξεις με μικρό μήκος. Λόγω αυτού μπορούμε με βεβαιότητα να πούμε ότι η μεγάλη διαφορά στην τιμή της εντροπίας των φινλανδικών οφείλεται στο εντονότερο πλατό των συχνότερων unigrams της γλώσσας αυτής σε σύγκριση με όλες τις υπόλοιπες.

Η φινλανδική γλώσσα παρουσιάζει ένα χαρακτηριστικό το οποίο δεν υπάρχει σε καμία από τις υπόλοιπες γλώσσες που μελετώνται εδώ: τη συγκολλητικότητα. Οι περισσότερες φινλανδικές λέξεις δημιουργούνται από την συγκόλληση διαφορετικών μορφημάτων.
Ένα παράδειγμα δίνεται παρακάτω[3]:

työ = work
työllä = by work
työllänsä = by his work
työllänsäkään = even by his work
työllistyttämättömyydelläänsäkään = even by his unability to get employed

Μπορούμε να δούμε πώς το γεγονός της παραγωγής λέξεων με διαφορετικά μήκη γύρω από το μήκος του αρχικού λήμματος (στη προκειμένη περίπτωση *työ* που σημαίνει *εργασία*) αντανακλάται στην τιμή της εντροπίας.

Στη συνέχεια προχωρούμε σε μία τυχαία αναδιάταξη των χρονοσειρών και υπολογισμό των

[3]Ευχαριστίες στο φίλο Matti Raasakka για τα χρήσιμα σχόλιά του πάνω στη φινλανδική γλώσσα

Φ_2 εντροπιών αυτών, προκειμένου να επαληθεύσουμε τα συμπεράσματα του πρώτου σώματος κειμένων (Πίνακας 5).

Language	Φ_2 for real series	Φ_2 for shuffled series
Dutch	4.4752	4.5168
English	4.3743	4.4434
Finnish	5.1337	5.1710
French	4.4139	4.4914
German	4.6042	4.6422
Greek	4.6007	4.6767
Italian	4.6817	4.7600
Portuguese	4.6881	4.7775
Spanish	4.5382	4.6082
Swedish	4.6035	4.6436

Πίνακας 5: Φ_2 εντροπίες για τις αρχικές συμβολοσειρές μηκών λέξεων καθώς και για τις τυχαία αναδιαταγμένες (Shuffled).

Τα συμπεράσματα είναι σαφή και μόνο για την περίπτωση των Φ_2, αφού λόγω περιορισμένης υπολογιστικής ισχύος οι υπολογισμοί παραλείπονται για τις Φ_3. Η σταθερά μεγαλύτερη τιμή των Φ_2 των τυχαίων συμβολοσειρών επιβεβαιώνει μία μικρή εξάρτηση αυτών από τις συσχετίσεις μεταξύ των μηκών γειτονικών λέξεων αν και η ουσιαστική συνεισφορά προέρχεται από τις κατανομές μηκών λέξεων (unigrams).

Συμπερασματικά, οι Φ_n εντροπίες είναι χαρακτηριστικά της γλώσσας και επηρεάζονται κυριώς από την κατανομή των σχετικά συχνών μηκών λέξεων (λέξεις με μικρά μήκη). Η κατανομή των σπανιότερων μηκών είναι πιο ευαίσθητη στην ομαδοποίηση και άρα προέλευση της γλώσσας. Τέλος σε όλες τις γλώσσες υπάρχει μικρή επίδραση των συσχετίσεων μεταξύ των μηκών των γετονικών λέξεων στην Φ_2. Το μέγεθος της επίδρασης μπορεί να συσχετισθεί με την οικογένεια της γλώσσας.

5 ΣΥΜΠΕΡΑΣΜΑΤΑ - ΠΡΟΟΠΤΙΚΕΣ

Στην εργασία αυτή, στο πλαίσιο της μαθηματικής και υπολογιστικής μελέτης της φυσικής γλώσσας, διεξήγαμε και υλοποιήσαμε μία μεθοδολογία εντροπικής ανάλυσης γραπτών κειμένων. Η χρήση χρονοσειρών μήκους λέξεων αποτέλεσε την αναπαράσταση του σώματος κειμένων. Η

κλασσική εντροπία του Shannon και η γενίκευση αυτής με την μορφή των $n-gram$ εντροπιών είναι μεγέθη ευαίσθητα στην αναγνώριση του είδους της γλώσσας (ελληνικά, αγγλικά, ολλανδικά κ.τ.λ.) και του είδους του κειμένου (πολιτικά και οικονομικά άρθρα, αθλητικά νέα, λογοτεχνία) για την εν λόγω αναπαράσταση. Η διαφορά στις τιμές των εντροπιών αποδίδεται στην ομοιομορφία και την παρουσία πλατό στις κατανομές πιθανότητας των μηκών λέξεων αλλά και στις διαφορετικές συσχετίσεις μεταξύ των μηκών γειτονικών λέξεων στις υπό μελέτη χρονοσειρές. Με την σειρά της, η παρουσία των πλατό στις κατανομές πιθανότητας αντανακλά βασικές γλωσσολογικές ιδιότητες των διαφόρων γλωσσών, όπως τον πλούτο της κλιτικής μορφολογίας και την παραγωγικότητα της γλώσσας (μέσω μηχανισμών όπως είναι η παραγωγή, η σύνθεση και η σύμμειξη (blending)).

Η μικρής αυτής κλίμακας μελέτη δεν επιτρέπει την εξαγωγή γενικευμένων συμπερασμάτων. Παρά ταύτα, αξιοσημείωτη είναι η ανάδειξη της πολυπλοκότητας της ανθρώπινης φυσικής γλώσσας ακόμα και μέσα από μία τόσο απλή αναπαράσταση όπως αυτή των συμβολοσειρών μήκους λέξεων. Στο πλαίσιο αυτό, η εργασία αυτή αποβλέπει στο να συμβάλλει στη συλλογική και πολύπλευρη μελέτη του πολύπλοκου αυτού συστήματος τόσο ως προς την πρακτική κατεύθυνση της βελτιστοποίησης της αλληλεπίδρασης ανθρώπου και μηχανής αλλά και ως προς την φιλοσοφική, ίσως, κατεύθυνση της μελέτης της ανθρώπινης γλώσσας ως κατασκεύασμα της ανθρώπινης νόησης.

Τα ευρήματα της εργασίας αυτής, θέτουν την ανάγκη περαιτέρω έρευνας προς δύο κατευθύνσεις: Η πρώτη εστιάζει στην εντροπική ανάλυση και πιό συγκεκριμένα στην κατανόηση και γενίκευση της εξάρτησης της εντροπίας από την αναπαράσταση ενός συστήματος που προέκυψε από τη σύγκριση των αποτελεσμάτων αυτής της εργασίας με αυτά των Papadimitriou et al. [21]. Η δεύτερη έχει περισσότερο γλωσσολογικό προσανατολισμό και θα μπορούσε να ξεκινήσει με μία πιο συστηματική διερεύνηση των δυνατοτήτων αυτόματης ταξινόμησης κειμένων ως προς την υπογλώσσα και το είδος τους με βάση τις τμηματικές εντροπίες και να συνεχίσει με τον εμπλουτισμό της αναπαράστασης των κειμένων με συντακτικά χαρακτηριστικά πέρα του μήκους των λέξεων και την εντροπική ανάλυση των παραγομένων συμβολοσειρών. Τέλος, αξίζει να διερευνηθεί περαιτέρω η παρουσία των μικρής και μακράς εμβέλειας συσχετίσεων σε ένα κείμενο και της επίπτωσής τους στις εντροπίες των συμβολοσειρών που τις αναπαριστούν.

Αναφορές

[1] P. Balasurhamanan & G. Siromoney [1968], "A Note on Entropy of Telugu Prose", *Information and Control*, **13**, 281-285.

[2] M. Baranger "Chaos, Complexity, and Entropy. A physics talk for non-physicists".

[3] F.H. Behr Jr., V. Fossum, M. Mitzenmaacher & D. Xiao [2003], "Estimating and Comparing Entropy across Written Natural Languages Using PPM Compression", *In Proceedings of the IEEE Data Compression Conference.*

[4] G. Bel-Enguix & M. D. Jimenez-Lopez [2010], "Language as a Complex System", *Cambridge Scholars*.

[5] J. Bhan, S. Kim, J. Kim, Y. Kwon, S. Yang & K. Lee [2005]"Long-range Correlations in Korean Literary Corpora", *Chaos, Solitons & Fractals*, **29**, 69-81.

[6] N. Chomsky [1957] "Syntactic structures", *The Hague: Mouton*.

[7] N. Chomsky [1986] "Knowledge of language: Its nature, Origin and Use", *New York: Praeger Publishers*.

[8] J. Ding & A. Zhou [2009] "Statistical properties of Deterministic Systems", *Tsinghua University Press, Springer*.

[9] W. Ebeling & A. Neiman [1995] "Long-range Correlations Between Letters and Sentences in Texts", *Physica A*, **215**, 233-241.

[10] A. Hacinliyan, M. Erentürk, & G. Sahin [2010] "Possible Chaotic Structures in the Turkish Language with Time Series Analysis", *Unifying Themes in Complex Systems*, **370**, 618-625.

[11] R. Hornero, M. Aboy, D. Abasolo, J. McNames & B. Goldstein [2005] "Interpretation of Approximate Entropy: Analysis of Intracranial Pressure Approximate Entropy During Acute Intracranial Hypertension", *IEEE Transactions on Biomedical engineering*, **52**, 1671-1680.

[12] Z. Harris [1951] "Structural linguistics", *The University of Chicago Press*.

[13] Z. Harris [1982] "Mathematical Analysis of Language", *Studies in Logic and the Foundations of Mathematics*, **104**, 623-637.

[14] R. Hartley [1928] "Transmission of Information", *Bell System Technical Journal*, **7**, 535-563.

[15] K. Karamanos [2001] "Entropy Analysis of Substitutive Sequences Revisited", *Journal of Physics A: Mathematical and General*, **34**, 9231-9241.

[16] K. Kosmidis, A. Kalampokis & P. Argyrakis [2006] "Language Time Series Analysis", *Physica A*, **370**, 808-816.

[17] L.B. Levitin & Z. Reingold [1993] "Entropy of Natural Languages: Theory and Experiment", *Chaos, Solitons and Fractals*, **4**, 709-743.

[18] S. Lu, X. Chen, J. Kanters, I. Solomon & K. H. Chon [2009] "Approximate Entropy for All Signals", *IEEE Eng. Med. Biol. Mag.*, **28**, 18-23.

[19] M. A. Montemurro & P. A. Pury [2002] "Long-range Fractal Correlations in Literary Corpora" *Fractals*, **10**, 451.

[20] H. Nyquist [1924] "Certain Factors Affecting Telegraph Speed", *Bell System Technical Journal*, **3**, 324-346.

[21] C. Papadimitriou, K. Karamanos, F.K. Diakonos, V. Constantoudis & H. Papageorgiou [2010]"Entropy Analysis of Natural Language Written Texts", *Physica A*, **389**, 3260-3266.

[22] F. Pereira [2000]"Formal Grammar and Information Theory: Together Again?", *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, **358**, 1239-1253.

[23] S. Pincus [1991] "Approximate Entropy as a Measure of System Complexity", *PNAS*, **88**, 2297-2301.

[24] B.S. Ramakrishna, [1963], 'Some Aspects of Relative Efficiencies of Indian Languages", *Indian Institute of Science, Bangalore.*

[25] K. Rateitschak, W. Ebeling & J. Freund [1996] "Nonlinear Dynamical Model for Texts", *Europhys. Lett.*, **35**, 401-406.

[26] G. Sahin, M. Erentürk & A. Hacinliyan [2009] "Detrended Fluctuation Analysis in Natural Languages Using Non-corpus Parametrization", *Chaos, Solitons & Fractals*, **41**, 198-205.

[27] C.E. Shannon [1948],"A Mathematical Theory of Communication", *Bell Syst. Tech. J.*, **27**, 379-423.

[28] C.E. Shannon [1951]"Prediction and Entropy of Printed English", *Bell Syst. Tech. J.*, **30**, 50-64.

[29] M. Shroeder [1992] "Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise", *W.H. Freeman & Company.*

[30] R. Solomonoff [1964] "A Formal Theory of Inductive Inference, Part I", *Information and Control*, **7, No 1**, 1-22.

[31] R. Solomonoff [1964] "A Formal Theory of Inductive Inference, Part II", *Information and Control*, **7, No 2**, 224-254.

[32] J. Tiedemann [2009] "News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces", *Recent Advances in Natural Language Processing*, **V**, 237-248.

[33] A. M. Turing [1950] "Computing Machinery and Intelligence", *Mind*, **59**, 433-460.

[34] C. Tsallis [1988] "Possible Generalization of Boltzmann-Gibbs Statistics", *Journal of Statistical Physics*, **52**, 479-487.

[35] M.R. Titchener [1998] "Deterministic Computation of Complexity, Information and Entropy", *IEEE International Symposium on Information Theory*, 326.

[36] J. Ziv & A. Lempel [1976], "On the Complexity of Finite Sequences", *IEEE Trans. Inf. Theory*, **IT-22**, 75-81.