ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΗΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΜΗΜΑ ΑΡΧΕΙΟΝΟΜΙΑΣ, ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ ΚΑΙ ΜΟΥΣΕΙΟΛΟΓΙΑΣ

Ποσοτική Γλωσσολογία στα Μικρά Κείμενα

Διδακτορική Διατριβή Αλίκης Σύλβιας Πουλημένου

Κέρκυρα, 2015

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΗΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΜΗΜΑ ΑΡΧΕΙΟΝΟΜΙΑΣ, ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ ΚΑΙ ΜΟΥΣΕΙΟΛΟΓΙΑΣ

Ποσοτική Γλωσσολογία στα Μικρά Κείμενα

Διδακτορική Διατριβή Αλίκης Σύλβιας Πουλημένου

Επόπτης:
Μάριος Πούλος:
Αν. Καθηγητής Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Μουσειολογίας Ιονίου Πανεπιστημίου
Συμβουλευτική Επιτροπή:
Σώζων Παπαβλασόπουλος: *Επ. Καθηγητής Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Μουσειολογίας Ιονίου Πανεπιστημίου*
Σοφία Στάμου: *Λέκτορας Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Μουσειολογίας Ιονίου Πανεπιστημίου*

Λοιπά Μέλη Εξεταστικής Επιτροπής:
Θεόδωρος Παππάς: *Καθηγητής Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Μουσειολογίας Ιονίου Πανεπιστημίου*
Ευστάθιος Σταματάτος: *Αν. Καθηγητής Τμήματος Μηχανικών Πληροφοριακών & Επικοινωνιακών Συστημάτων Πανεπιστημίου Αιγαίου*
Ιωάννης Παπαδάκης: *Επ. Καθηγητής Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Μουσειολογίας Ιονίου Πανεπιστημίου*
Κάτια Λήδα Κερμανίδου: *Λέκτορας Τμήματος Πληροφορικής Ιονίου Πανεπιστημίου*

Δήλωση: Δηλώνω υπεύθυνα ότι το παρόν κείμενο αποτελεί προϊόν προσωπικής μελέτης και εργασίας και πώς όλες οι πηγές που χρησιμοποιήθηκαν για τη συγγραφή της δηλώνονται σαφώς είτε στις παραπομπές είτε στο βιβλιογραφικό κατάλογο.

Υπογραφή
_____

ΕΥΧΑΡΙΣΤΙΕΣ

Για την υλοποίηση και εκπόνηση της διδακτορικής μου διατριβής θα ήθελα να ευχαριστήσω όσους με βοήθησαν σε αυτό το δύσκολο έργο.

Πρώτα από όλα θα ήθελα να ευχαριστήσω θερμά τον Επόπτη της διατριβής μου κ. Μάριο Πούλο χωρίς την καθοδήγηση του οποίου η διατριβή αυτή δεν θα ήταν δυνατό να διεξαχθεί. Όλα αυτά τα χρόνια η επιστημονική του ειδίκευση και η ψυχολογική του υποστήριξη, το αμείωτο ενδιαφέρον του για την επιστημονική έρευνα και η βοήθεια του σε όλη την πορεία αυτών των ετών ήταν πολύτιμη.

Στη συνέχεια θα ήθελα να ευχαριστήσω ιδιαίτερα τα υπόλοιπα μέλη της Συμβουλευτικής μου Επιτροπής, κ. Σώζων Παπαβλασσόπουλο και κα. Σοφία Στάμου. Η προσφορά τους στην ολοκλήρωση της διατριβής μου ήταν σημαντικότατη μέσω των επιστημονικών τους γνώσεων, παρατηρήσεων και προτάσεων.

Τέλος θα ήθελα να ευχαριστήσω θερμά όλη την οικογένεια μου. Το σύζυγό μου, τους γονείς μου και τον αδελφό μου. Η κατανόηση τους και η αμέριστη συμπαράσταση τους όλα αυτά τα χρόνια ήταν η δύναμη μου και τους είμαι ευγνώμων.

Πουλημένου Αλίκη Σύλβια

Κέρκυρα, 2015

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΚΟΝΕΣ

ΠΙΝΑΚΕΣ

ΠΕΡΙΛΗΨΗ

Η παρούσα διατριβή αφορά τη μελέτη ποσοτικών γλωσσολογικών ζητημάτων που αφορούν τα μικρά κείμενα. Το ερευνητικό υπόβαθρο για την υλοποίηση της διατριβής αφορά στην τομή τριών επιστημονικών κλάδων, αυτών της Ανάκτησης Πληροφορίας, Υπολογιστικής Γλωσσολογίας και Ποσοτικής Γλωσσολογίας.

Στην διατριβή αρχικά μελετήθηκε σε βάθος η σχετική βιβλιογραφία που αφορά τους παραπάνω κλάδους, οι επιστημονικές τάσεις αλλά και τα προβλήματα που τους αφορούν. Στην συνέχεια μέσω στατιστικών εργαλείων πραγματοποιήθηκαν εκτεταμένα πειράματα, τα οποία απέδειξαν έναν καινοτόμο γλωσσολογικό νόμο ποσοτικής φύσεως που αφορά τα μικρά κείμενα. Για να γίνει αυτό συγκεράστηκαν θεωρίες και τεχνικές και από τους τρεις επιστημονικούς κλάδους, οι οποίοι αναλύονται στα κεφάλαια της διατριβής.

Πιο συγκεκριμένα μέσω της παρούσας έρευνας στο πλαίσιο εκπόνησης της διατριβής είναι δυνατή η περιγραφή της δομής μιας πρότασης με βάση συγκεκριμένες προϋποθέσεις ώστε να οριστεί η συνεκτικότητά της μέσω του μήκους της πρότασης χωρίς να διαταραχθεί ο επικοινωνιακός της στόχος.

ABSTRACT

This thesis concerns the study of quantitative phenomena regarding short texts. At first the necessary research background of the thesis consists of the union of three main scientific areas which are Information Retrieval, Computational Linguistics and Quantitative Linguistics.

Firstly, all relevant bibliography considering scientific trends and problems concerning the above scientific areas were studied in depth. Afterwards via statistical tools extensive experiments were realized, which proved an innovating quantitative linguistics law that concerns short texts. For the utilization of the above, theories and techniques from all three scientific areas were combined and are analyzed through the chapters of this thesis.

In detail through this research, it is possible to describe the structure of a sentence and defined its cohesion without disturbing its communicative role, under certain conditions.

ΚΕΦΑΛΑΙΟ 1ο

ΕΙΣΑΓΩΓΗ

1.1 **Ανάκτηση Πληροφορίας και σύγχρονα προβλήματα**

Καθώς ο όγκος της πληροφορίας στο διαδίκτυο ολοένα και αυξάνεται τόσο μεγαλώνει η ανάγκη **για σωστή διαχείριση της αποθηκευμένης πληροφορίας** ώστε να είναι προσβάσιμη με τον καλύτερο δυνατό τρόπο ανάλογα με τις εκάστοτε πληροφοριακές ανάγκες του χρήστη. Η παραγωγή πληροφοριών αυξάνεται όλο και περισσότερο (Moens 2006) αγγίζοντας τα 167 terabytes πληροφορίας στον επιφανειακό ιστό (surface web) και στο βαθύ ιστό (deep web) υπολογίζονται 400 φορές μεγαλύτερα μεγέθη. Με τον όρο βαθύ ιστό γίνεται αναφορά στο τμήμα των πληροφοριών που βρίσκονται στο διαδίκτυο και ενώ είναι προσβάσιμες για τον χρήστη, είναι μη ανιχνεύσιμες για τις μηχανές αναζήτησης.

Λαμβάνοντας υπόψη την αδόμητη φύση των πληροφοριών, η αναζήτηση συναφούς πληροφορίας αποτελεί μεγάλο εγχείρημα για τον χρήστη. Η καταγραφή της **πληροφορίας** και η **κατάλληλη επεξεργασία** της αποτελεί βασικό προαπαιτούμενο προκειμένου να μπορεί να ανακτηθεί. Ουσιαστικά η πληροφορία που δεν έχει επεξεργαστεί κατάλληλα ώστε να καθίσταται ανακτήσιμη μπορεί να **θεωρηθεί πραγματικά χαμένη,** καθώς χωρίς την επεξεργασία αυτή είναι αδύνατη η αναγνώριση της μέσα σε τεράστιες συλλογές ή όγκο πληροφοριών.

Η ανάγκη αυτή παρουσιάστηκε σχεδόν **απαρχής γέννησης της γραφής** και παραγωγής του γραπτού λόγου. Σύμφωνα με την βιβλιογραφία (Singhal 2001), οι Σουμέριοι ηδη από το 3000 π. Χ. είχαν αντιληφθεί την σημασία οργάνωσης και πρόσβασης των αρχείων τους, χρησιμοποιούσαν ειδικές περιοχές για την αποθήκευση πήλινων πινακίδων με σφηνοειδή γραφή και ανέπτυξαν ειδικούς τρόπους ταξινόμησης ώστε να μπορούν να αναγνωρίσουν τις πινακίδες και αντίστοιχα το περιεχόμενο τους. Αργότερα **οι αρχαίοι Έλληνες και οι Ρωμαίοι** (Ceri et al. 2013) κατέγραφαν πληροφορίες σε πάπυρους καθώς επίσης και περιλήψεις για το περιεχόμενο τους, ενώ από τον 2^{ο} αιώνα π.Χ. για πρώτη φορά εμφανίζονται σε ελληνικούς πάπυρους οι πίνακες περιεχομένων. Οι ανάγκες για κατάλληλη οργάνωση και διαχείριση της πληροφορίας αυξάνεται καθώς η ιστορία προχωρά και ειδικότερα **καθώς αναπτύσσεται η παραγωγή του γραπτού λόγου**. Κομβικά σημεία αποτελούν η ανακάλυψη του χαρτιού και η ανακάλυψη της τυπογραφίας με αποκορύφωμα την εφεύρεση του ηλεκτρονικού υπολογιστή και του διαδικτύου, όπου πλέον **η παραγωγή πληροφοριών είναι τεράστια.**

Η **ανάκτηση της πληροφορίας** αποτελεί πολύ βασικό ζήτημα ειδικότερα **για τον χώρο των βιβλιοθηκών**, που βασική τους λειτουργία είναι η συγκέντρωση πνευματικής παραγωγής, οργάνωση του υλικού και επεξεργασία του ώστε να είναι προσβάσιμο. Έτσι όπως αναφέρεται και στη βιβλιογραφία (Onwuchekwa 2011) η ανάκτηση της πληροφορίας κατά της **δεκαετία του '50** θεωρήθηκε πολύ σημαντική πρόοδος για την οργάνωση. Σύμφωνα με τον Van Rijsbergen (1979), το ζήτημα οργάνωσης της πληροφορίας και ανάκτησης της, στα περιβάλλοντα τον βιβλιοθηκών, αντιμετωπίστηκε με την **ενσωμάτωση βασικών βιβλιοθηκονομικών καθηκόντων στους ηλεκτρονικούς υπολογιστές, μέσω συστημάτων ανάκτησης πληροφορίας (ΣΑΠ)**.

Όμως, όπως αναφέρεται στη βιβλιογραφία (Blair και Kimbrough, 2002) η ανάκτηση πληροφοριών δεν περιορίζεται μόνο στο χώρο των βιβλιοθηκών, αλλά με την **εμφάνιση του διαδικτύου** και της διαρκούς αύξησης της πληροφορίας, ο σχεδιασμός ΣΑΠ αντιμετωπίζει πολλές προκλήσεις για την βελτίωση της αποτελεσματικότητας της αναζήτησης, όπως την εύρεση κατάλληλου τρόπου αναπαράστασης των εγγράφων προκειμένου να μπορεί ο χρήστης να έχει πρόσβαση στο σημασιολογικό περιεχόμενο τους. Έτσι ο Van Rijsbergen (1979) αναφέρει πως στην αποθήκευση και ανάκτηση της πληροφορίας δόθηκε ιδιαίτερη έμφαση κατά **την δεκαετία του '40, πέρα από τον χώρο των βιβλιοθηκών**. Έτσι δημιουργούνται τα συστήματα ανάκτησης ηλεκτρονικών υπολογιστών. Πιο συγκεκριμένα η ιδέα για αποθήκευση πληροφοριών και η αυτόματη πρόσβαση σε αυτές (Singhal 2001) γεννήθηκε το 1945 από ένα άρθρο του Vannevar Bush και υλοποιήθηκε κατά τη διάρκεια της δεκαετίας του '50, ενώ το '60 πραγματοποιείται η ανάπτυξη του συστήματος **SMART από τον Gerard Salton και την επιστημονική ομάδα του στο Cornell University** το οποίο είχε σκοπό την βελτίωση της ποιότητας αναζήτησης. Ταυτόχρονα όπως αναφέρεται στη βιβλιογραφία (Singhal 2001) πραγματοποιούνται τα **τεστ Cranfield**, από τον Cyril Cleverdon και την επιστημονική του ομάδα στο College of Aeronautics από τα οποία προέκυψε η μεθοδολογία αξιολόγησης ΣΑΠ.

Κατά τις δεκαετίες '70 – '80, δίνεται έμφαση στην έρευνα των ΣΑΠ και πραγματοποιούνται απόπειρες βελτίωσης των ηδη υπαρχόντων συστημάτων. Όπως αναφέρεται στη βιβλιογραφία (Singhal 2001), η έρευνα και οι βελτιώσεις που πραγματοποιούνται αφορούν **στην ανάκτηση εγγράφων** και περιορίζονται σε συλλογές μικρής κλίμακας καθώς δεν υπήρχαν ιδιαίτερα μεγάλες συλλογές κειμένων, μέχρις ότου ξεκίνησαν το 1992 τα **Text Retrieval Conferences (TREC),** υπό την

χορηγία του National Institute of Standards and Technology (NIST) από την κυβέρνηση των ΗΠΑ για έρευνα σχετική με την ανάκτηση της πληροφορίας σε μεγάλες συλλογές. Τελειώνοντας την ιστορική αναδρομή της ανάκτησης της πληροφορίας, έχει ήδη αναφερθεί πως η εφεύρεση των υπολογιστών και ο παγκόσμιος ιστός έπαιξε καταλυτικό ρόλο για την αποθήκευση, πρόσβαση και αναζήτηση πληροφορίας σε συλλογές εγγράφων (Ceri et al. 2013).

Έτσι λοιπόν με την βασική προϋπόθεση της αποθήκευσης και κατάλληλης επεξεργασίας της κάθε πληροφορίας ώστε να είναι ανακτήσιμη για να μπορέσει να καλύψει της πληροφοριακές ανάγκες του χρήστη κατά τη διάρκεια μιας αναζήτησης του για συναφή πληροφορία, δημιουργήθηκε ο επιστημονικός κλάδος της **Ανάκτηση Πληροφορίας (ΑΠ)**. Πιο συγκεκριμένα σύμφωνα με τους Baeza-Yates και Ribeiro-Neto (2011) «***η ανάκτηση πληροφοριών ασχολείται με την αναπαράσταση, αποθήκευση, οργάνωση και πρόσβαση πληροφοριακών τεκμηρίων (items), όπως έγγραφα, ιστοσελίδες, online κατάλογοι, δομημένα και ημι-δομημένα αρχεία, αντικείμενα πολυμέσων. Η αναπαράσταση και οργάνωση των πληροφοριακών τεκμηρίων θα πρέπει να είναι τέτοια ώστε να παρέχει στους χρήστες εύκολη πρόσβαση στην πληροφορία που τους ενδιαφέρει***». Οι Manning, Raghavan και Schutze (2008) ορίζουν την ΑΠ ως εξής: «*η αναζήτηση και εύρεση υλικού (συνήθως εγγράφων) αδόμητης φύσης (συνήθως κείμενο) που ικανοποιεί μια πληροφοριακή ανάγκη μέσα από μεγάλες συλλογές (συνήθως αποθηκευμένες σε υπολογιστές)*». Από τον παραπάνω ορισμό, ο όρος **«έγγραφο»** αφορά το ίδιο το υλικό ενώ ο όρος **«κείμενο»** χαρακτηρίζει τη φύση του υλικού. Συνεπώς **οι όροι αυτοί θα χρησιμοποιούνται με τις παραπάνω σημασίες στην διατριβή**.

Έναν ακόμη ορισμό ΑΠ βρίσκουμε στους Grossman και Frieder (2000): «*ως η αρχή εύρεσης* ***συναφών*** *εγγράφων σε αντίθεση με την απλή αντιστοίχιση λεξιλογικών μοτίβων σε ένα ερώτημα*», όπου εισάγεται και η έννοια **της συνάφειας (relevance)**. Συναφές (Croft et al. 2009) θεωρείται ένα έγγραφο το οποίο περιέχει την πληροφορία την οποία αναζητούσε ο χρήστης κατά την υποβολή του ερωτήματος του στο ΣΑΠ όπως π.χ. συχνά είναι μια μηχανή αναζήτησης. Τα ΣΑΠ έχουν ως στόχο την ανάκτηση εγγράφων από μια συλλογή, τα οποία έγγραφα θα ικανοποιούν τις ανάγκες του χρήστη. Η συνάφεια των αποτελεσμάτων σε σχέση με το ερώτημα του χρήστη θα καθορίσει το πόσο επιτυχημένο και αποτελεσματικό είναι το ΣΑΠ. Χαρακτηριστικά της συνάφειας (Ceri 2013) είναι **η υποκειμενικότητά της**, καθώς έχει να κάνει με την κρίση του χρήστη ως προς το αν ικανοποιείται η πληροφοριακή του ανάγκη από

τα αποτελέσματα που επιστρέφει το ΣΑΠ και η **δυναμική της φύση στο χωροχρόνο** και το ότι είναι **πολύπλευρη** καθώς καθορίζεται τόσο από το περιεχόμενο των ανακτηθέντων αποτελεσμάτων όσο και από ζητήματα που έχουν να κάνουν με την πηγή της πληροφορίας όπως η αυθεντικότητα ή η ειδικότητα. Σύμφωνα με τον Ingwersen (1992) **το πρόβλημα της συνάφειας επικεντρώνεται στην εύρεση της κατάλληλης πληροφορίας σε σχέση με το ερώτημα του χρήστη**. Έτσι η πληροφορία που βρίσκεται σε ένα κείμενο συγκρίνοντας το με άλλα κείμενα μπορεί να είναι πιο συναφής ως προς συγκεκριμένη απαίτηση πληροφορίας του χρήστη. Ακόμη η σημασία που έχει ένα κείμενο εξαρτάται από τις πληροφοριακές απαιτήσεις και μπορεί να αλλάζει καθώς αλλάζουν και αυτές. Οι Croft et al. (2009) **διακρίνουν το φαινόμενο της συνάφειας σε τοπική συνάφεια και συνάφεια χρήστη**. Ως τοπική συνάφεια περιγράφουν το φαινόμενο κατά το οποίο ένα έγγραφο και το ερώτημα του χρήστη έχουν το ίδιο θέμα ενω η συνάφεια χρήστη έγκειται στις προσωπικές απαιτήσεις του χρήστη, οι οποίες καθιστούν το έγγραφο συναφές ή όχι.

Η ΑΠ αποτελείται από δύο μεγάλες διαδικασίες: την **ευρετηρίαση** και την **αναζήτηση ερωτήματος χρήστη**, όπου και οι δύο διαδικασίες αντιμετωπίζουν τα ίδια επιστημονικά προβλήματα, τα οποία προκύπτουν από το πολύπλευρο ζήτημα της συνάφειας. Οι Baeza-Yates και Ribeiro-Neto (2011) αναλύουν περισσότερο τα προβληματικά σημεία της ΑΠ εστιάζοντας στο ζήτημα **«μετάφρασης» των πληροφοριακών αναγκών του χρήστη στο πλαίσιο μιας αναζήτησης.** Υποδεικνύουν τις επιμέρους επεξεργασίες νοητικά για τον χρήστη και αυτοματοποιημένα για τον υπολογιστή όπου αντίστοιχα **ο χρήστης πρέπει να περιγράψει την πληροφοριακή του ανάγκη και να την συμπιέσει σε ορισμένες λέξεις-κλειδιά** που θα αποτελέσουν το ερώτημα που θα εισάγει μέσω της διεπαφής ενώ από την άλλη το **ΣΑΠ θα πρέπει να ανατρέξει στο ευρετήριο των εγγράφων μιας συλλογής, να «κατανοήσει» το περιεχόμενο τους, να κρίνει τα συναφή έγγραφα και να τα επιστρέψει ως αποτελέσματα σε μια λίστα καταταγμένη ανάλογα με το βαθμό συνάφειας σε σχέση με το ερώτημα**. Για την παραπάνω διαδικασία είναι απαραίτητη η γλωσσολογική γνώση, η οποία αφορά την κατανόηση του κειμένου σε συντακτικό και σημασιολογικό επίπεδο.

Σε σχέση με το χρήστη (Woods et al. 2000), η αποτυχία εύρεσης κατάλληλης πληροφορίας εξαρτάται από τις **λέξεις-κλειδιά που εισάγει** ως ερώτημα και από το αν και **κατά πόσον αυτές χρησιμοποιούνται από το αντίστοιχο συναφές υλικό**, ως προς την πληροφοριακή του ανάγκη. Ακόμη (Woods et al. 2000) προβληματικό

σημείο είναι και **ο απαιτούμενος χρόνος που χρειάζεται να δαπανήσει ένας χρήστης διαβάζοντας τα ανακτηθέντα αποτελέσματα** της αναζήτησης ώστε να ανακαλύψει εάν η πληροφορία που αναζητά περιέχεται σε κάποιο από αυτά. Σύμφωνα με τον Lewis και την Sparck Jones (1996) **κατά τη διαδικασία της ανάκτησης εγγράφων, ο χρήστης έχει άγνοια ως προς την ακριβή πληροφορία που αναζητά** και αυτό επιφορτίζει το ΣΑΠ με το σημαντικό καθήκον της **σύνδεσης της πληροφοριακής ανάγκης του χρήστη με αντίστοιχα συναφή έγγραφα**. Για να πραγματοποιηθεί όμως αυτό θα πρέπει το υπολογιστικό σύστημα να αναζητήσει τις σχέσεις των δύο παραπάνω πλευρών. Η δυσκολία της σύνδεσης αυτών των δύο έχει να κάνει στο ότι δεν υπάρχει κάποια συγκεκριμένη συνθήκη με βάση την οποία να ικανοποιείται η πληροφοριακή ανάγκη του χρήστη, καθώς το ερώτημα του αποτελεί μια συντομευμένη περιγραφή, θεωρητικά αρκετά ασαφή.

Από την πλευρά των ΣΑΠ (Woods et al. 2000) υπάρχει ανάγκη βελτίωσης της αποτελεσματικότητας της αναζήτησης, μέσω **γλωσσικών συσχετίσεων στα κείμενα** όπως μορφολογικές σχέσεις μεταξύ των λέξεων και ταξινομικές σχέσεις μεταξύ εννοιών. Ο Van Rijsbergen (1979) αναφέρει πως το πρόβλημα εστιάζεται στην δυσκολία ερμηνείας της φυσικής γλώσσας με αυτόματο τρόπο από τα υπολογιστικά συστήματα και της κατανόησης του περιεχομένου των εγγράφων και πως η δυσκολία αυτή έχει δύο άξονες: την **διαδικασία ανάγνωσης** (με σκοπό την εύρεση συναφούς πληροφορίας), η οποία περιλαμβάνει συντακτική και σημασιολογική ανάλυση και την **απόφαση της συνάφειας ή μη** σε σχέση με μεμονωμένο έγγραφο.

Η **ευρετηρίαση** αφορά τον **καθορισμό του περιεχομένου των εγγράφων**, σύμφωνα με τον Lewis και την Sparck Jones (1996) και είναι υπεύθυνη για την αύξηση της ακρίβειας και ανάκλησης σε ένα ΣΑΠ. Τα **προβλήματα που εμφανίζονται** εσωτερικά **κατά την ευρετηρίαση** αφορούν διάφορους περιορισμού όπως: **γλωσσικοί περιορισμοί** με παράδειγμα το φαινόμενο της πολυσημίας, περιορισμοί **ως προς την έκφραση του ερωτήματος** (ασαφή ερωτήματα λόγω της άγνοιας χρήστη και ατελή ερωτήματα, τα οποία δεν περιλαμβάνουν αρκετές λεπτομέρειες) και περιορισμοί στην **αναπαράσταση εγγράφων**, με αποτέλεσμα την απώλεια πληροφορίας από τα πρωτότυπα έγγραφα. Όπως αναφέρει ο Van Rijsbergen (1979) μέσω των κατάλληλων **αναπαραστάσεων εγγράφων και ερωτημάτων θα κριθεί η** συνάφεια ή μη ενός εγγράφου και ειδικότερα για την ευρετηρίαση τονίζει ότι η πληροφορία που αποθηκεύεται στα ευρετήρια σε σχέση με τα έγγραφα αποτελεί σμικρυμένες **αναπαραστάσεις εγγράφων** και όχι αυτούσια όλες τις πληροφορίες που

περιέχουν αυτά με αποτέλεσμα να υπάρχει απώλεια στην πληροφορία. Ειδικότερα ως **προς τα μοντέλα ΑΠ** και τον καθορισμό συνάφειας ο Spoerri (1995) διακρίνει διάφορες προσεγγίσεις ανάκτησης, εκ των οποίων και **η στατιστική προσέγγιση** η οποία χρησιμοποιεί στατιστικές πληροφορίες. Στην προσέγγιση αυτή ανήκει και το Vector Space Model (VSM), το οποίο χρησιμοποιεί τις πληροφορίες στατιστικής εμφάνισης όρων στα έγγραφα, για την παραγωγή των συναφών αποτελεσμάτων σε μορφή λίστας κατάταξης. Στην στατιστική **προσέγγιση ένα από τα μειονεκτήματα των μοντέλων αυτών** που απαντώνται σύμφωνα με Spoerri (1995) αφορά την **έλλειψη σχετικής δομής ώστε να μπορούν να εκφραστούν γλωσσολογικά χαρακτηριστικά όπως φράσεις, καθώς και περιορισμοί εγγύτητας.**

Οι Galvez, de Moya-Anegon και Solana (2005) συμπληρώνουν ειδικότερα για το ζήτημα της **ανεπαρκούς αναπαράστασης εγγράφων και ερωτημάτων** πώς επηρεάζει την απόδοση των ΣΑΠ και τονίζουν τη σημαντικότατη **συμβολή των τεχνικών Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing/NLP)**, προκειμένου να αντιμετωπιστεί το ζήτημα της **γλωσσικής ανάλυσης των όρων** που περιέχουν τα έγγραφα, το οποίο αφορά σε ποικιλία λεξιλογικών, συντακτικών και μορφολογικών χαρακτηριστικών για κάθε μεμονωμένο όρο. Μέσω της ενσωμάτωσης τεχνικών Επεξεργασίας Φυσικής Γλώσσας στα ΣΑΠ **γίνεται μια απόπειρα εφαρμογής των πρακτικών της Υπολογιστικής Γλωσσολογίας (ΥΓ)**. Τα βασικά προβλήματα έχουν να κάνουν με την πολυσημία ενός όρου (ένας όρος/μια λέξη με πολλές σημασίες) και με τους όρους που αποτελούνται από πολλές φράσεις έχοντας έτσι δομή φράσεων, η οποία δομή επιδέχεται πολλές διαφορετικές παραλλαγές. Μάλιστα όπως συμπληρώνουν οι Galvez, de Moya-Anegon και Solana (2005), οι **γλωσσικές παραλλαγές** που παρατηρούνται στους όρους χωρίζονται σε τρεις κατηγορίες:

- Μορφολογικές παραλλαγές, αφορά τις διαφορετικές μορφές εμφάνισης ενός όρου (όπου εφαρμόζεται η διαδικασία περιστολής ώστε να επιλεγεί μια ενιαία μορφολογική παραλλαγή).
- Λεξιλογικές - σημασιολογικές παραλλαγές, αφορούν τη σημασιολογική εγγύτητα των λέξεων (μια έννοια - πολλοί όροι ή πολλές έννοιες – ένας όρος).
- Συντακτικές παραλλαγές, αφορούν τις διαφορετικές δομές όρων πολλών λέξεων (μείωση/ διαίρεση κανονικών/κανονικοποιημένων συντακτικών δομών (canonical syntactic structure).

1.2 Νευροεπιστήμες και Γλωσσολογία

Παραπάνω πραγματοποιήθηκε μια εισαγωγή στην ΑΠ, το αντικείμενο της, τα προβλήματα που αντιμετωπίζονται και τον τρόπο με τον οποίο συνδέεται με την επιστήμη της Γλωσσολογίας. Ειδικότερα η Γλωσσολογία μελετά την ανάπτυξη και παραγωγή της γλώσσας, η οποία είναι άμεσα συνδεδεμένη με τον ανθρώπινο εγκέφαλο, αντικείμενο που εμπίπτει στον επιστημονικό κλάδο των Νευροεπιστημών. Όπως αναφέρει και ο Miller (2003) η Γλωσσολογία συνδέεται άμεσα με τις Νευροεπιστήμες και την Επιστήμη των Υπολογιστών (ως βασικές επιστήμες), σε συνδυασμό ακόμη με τη φιλοσοφία, την ανθρωπολογία και τη ψυχολογία υπό τη σκέπη της γνωσιακής επιστήμης (cognitive science), η οποία αφορά ευρύτερα τη διεπιστημονική μελέτη του νου. Οι παραπάνω σχέσεις γίνονται εμφανής από το σχηματισμό νέων επιστημονικών κλάδων όπως η ΥΓ (Γλωσσολογίας και Επιστήμη Υπολογιστών) και η Νευρογλωσσολογία (Νευροεπιστήμες και Γλωσσολογία).

Μέσω της **Νευρογλωσσολογίας** οι δύο παραπάνω επιστήμες συμβάλλουν στην **κατανόηση της αντίληψης της γλώσσας, παραγωγής της και υλοποίησης των μηχανισμών της από τον ανθρώπινο εγκέφαλο**. Σύμφωνα με την βιβλιογραφία (Ahlsen 2006), οι Νευροεπιστήμες αποτελούν έναν **κλάδο εξ ορισμού διεπιστημονικό**, ο οποίος συνδυάζει εκτός της Γλωσσολογίας και των Νευροεπιστημών και πολλούς άλλους επιστημονικούς κλάδους όπως αυτόν της Επιστήμης των Υπολογιστών. Σύμφωνα με Ahlsen (2006) τη Νευρογλωσσολογία απασχολούν πολλά ερωτήματα όπως μοντέλα επεξεργασίας γλώσσας ή υπολογιστικές προσομοιώσεις για την επεξεργασία της γλώσσας. Η Νευρογλωσσολογία ξεκίνησε από τη μελέτη προβληματικών εγκεφάλων και πηγή δεδομένων της αποτελούν οι μετρήσεις της εγκεφαλικής δραστηριότητας κατά τη διάρκεια συνηθισμένων γλωσσικών λειτουργιών καθώς και οι απεικονίσεις του εγκεφάλου μέσω απεικονιστικών εξετάσεων όπως είναι παραδείγματος χάρη μια αξονική ή μαγνητική τομογραφία. Βέβαια αν και η Νευρογλωσσολογία συνδυάζει τη Γλωσσολογία και τις Νευροεπιστήμες όπως ήδη έχει αναφερθεί, παρατηρείται το φαινόμενο αδυναμίας αντιστοίχησης εννοιολογικών γενικεύσεων των δύο αυτών επιστημών, υπό την σκέπη μιας ενιαίας οντολογίας, γνωστό στη βιβλιογραφία (Poeppel και Embick 2005) και ως πρόβλημα Οντολογικής Ασσυμετρίας (Ontological Incommensurability Problem/OIP).

Ως Νευροεπιστήμες (Squire et al. 2008) ορίζονται οι «*διεπιστημονικές επιστήμες που αναλύουν το νευρικό σύστημα ώστε να κατανοήσουν την βιολογική βάση της συμπεριφοράς*». Έτσι σύμφωνα με Pulvermuller (2003) η γλώσσα αναφέρεται ως ένα σύστημα εγκεφαλικών κυκλωμάτων, το νευρικό σύστημα. Το βασικότερο στοιχείο του νευρικού συστήματος είναι ο νευρώνας, το νευρικό κύτταρο, το οποίο εκπέμπει και λαμβάνει σήματα (signals) προς και από άλλους νευρώνες μέσω των κομβίων επικοινωνίας, τις λεγόμενες συνάψεις. Όπως αναφέρεται στη βιβλιογραφία (Pulvermuller 2003) ο εγκεφαλικός φλοιός του ανθρώπου είναι ένα τεράστιο δίκτυο νευρώνων που επικοινωνούν μεταξύ τους ώστε να εκπέμψουν και να λάβουν κάποια πληροφορία την οποία και επεξεργάζονται. Μάλιστα οι νευρώνες είναι συνυφασμένοι μεταξύ τους μέσω των διαδικασιών που πρέπει να υλοποιηθούν και έτσι δημιουργούν σύνολα νευρώνων που αλληλεπιδρούν μεταξύ τους και απαρτίζουν τους ιστούς νευρώνων, των οποίων οι συνδέσεις είναι διαφορετικές για κάθε διαδικασία. Ουσιαστικά οι πληροφορίες που λαμβάνει ο εγκεφαλικός φλοιός αναπαρίστανται από έναν ιστό νευρώνων, οι οποίοι αποτελούν ένα σύνολο νευρώνων από διάφορες περιοχές του εγκεφάλου που συνδέονται μεταξύ τους άμεσα και αν και εργάζονται ως μια μονάδα τα λειτουργικά μέρη είναι ανεξάρτητα ώστε κάθε μέρος να συνεισφέρει στην λειτουργία του ιστού. **Μπορεί κανείς να αντιληφθεί τον εγκέφαλο ως μια μνήμη στην οποία αποθηκεύονται συσχετίσεις για τις πληροφορίες που λαμβάνει. Οι ιστοί νευρώνων είναι λειτουργικές μονάδες που αναπαριστούν γνωστικές οντότητες με αισθητήριες πτυχές και πτυχές ενέργειας όπως λέξεις και έννοιες.**

Σύμφωνα με τη βιβλιογραφία (Pulvermuller 2003) πατέρας της Γλωσσολογίας θεωρείται ο Ferdinand de Saussure, ο οποίος τόνισε την αναγκαιότητα συσχετισμού της Γλωσσολογίας με τις Νευροεπιστήμες, εστιάζοντας στη σύνδεση γλωσσικών περιγραφών με περιγραφές νευρώνων. Σύμφωνα με τον de Saussure (1966) αν και το αντικείμενο της Γλωσσολογίας αφορά στην παρατήρηση της ομιλίας του ανθρώπου, σε όλες τις διαφορετικές εκδηλώσεις, **ο γραπτός λόγος (κείμενο) αποτελεί τη σημαντικότερη πηγή μελέτης**. Η περιοχή του εγκεφάλου που έχει σχέση με την ομιλία και ό,τι την αφορά ονομάζεται περιοχή Broca και περιλαμβάνει και το γραπτό λόγο, τη γλώσσα. Με τον όρο «**γλώσσα**» γίνεται αναφορά σε συστατικό του εγκεφάλου, το οποίο μάλιστα παρουσιάζει μεγάλες ομοιότητες με την φύση του γενετικού κώδικα, καθώς είναι «***ιεραρχική, παραγωγική, επαναλαμβανόμενη και εικονικά απεριόριστη σε σχέση με το πεδίο έκφρασης της***», όπως αναφέρουν οι Hauser, Chomsky και Fitch (2002). Σύμφωνα με τον de Saussure (1966) η γλώσσα

αποτελεί ένα στάδιο της διαδικασίας ομιλίας. Χρησιμοποιείται ένας κώδικας για την έκφραση, η γλώσσα, η οποία στον ανθρώπινο εγκέφαλο απαρτίζεται από γλωσσικά σύμβολα (linguistic signs) που εκφράζουν τις έννοιες που αναπαριστούν την πραγματικότητα για το πώς αντιλαμβανόμαστε τον κόσμο και μεταφράζονται σε αλφάβητο για την έκφραση με γραπτό λόγο. **Ο ανθρώπινος εγκέφαλος ουσιαστικά αντιστοιχεί έννοιες με σύμβολα, τα οποία αναπαρίστανται από ήχο - εικόνα και αντίστοιχα από κάποιο συνδυασμό γραπτού κώδικα**. Συνεπώς **η γλώσσα αποτελεί μια αντιστοιχία συμβόλων που εκφράζουν ιδέες.**

1.3 Ερευνητικές τάσεις στην σύγχρονη Υπολογιστική και Ποσοτική Γλωσσολογία

Όπως αναφέρεται στη βιβλιογραφία (Gries 2013) την τελευταία δεκαετία παρατηρείται μια **αλματώδης στροφή της έρευνας της Γλωσσολογίας προς την ποσοτική μελέτη** και έρευνα των δεδομένων που την αφορούν καθώς μέσω της εφαρμογής **στατιστικών μεθόδων** είναι δυνατή η τεκμηρίωση της **περιγραφής δεδομένων σχετικά με φαινόμενα που παρατηρούνται, ο έλεγχος και η επαλήθευση υποθέσεων σχετικά με τις σχέσεις δεδομένων σε ένα σύνολο και οι προβλέψεις σχετικά με αυτά**. Έτσι έχει αυξηθεί κατακόρυφα η χρήση ποσοτικών μεθόδων σε όλους σχεδόν τους υπο-κλάδους της επιστήμης της Γλωσσολογίας. Όπως συμπληρώνεται σε άλλη μελέτη (Gries to appear in *International Encyclopedia of the Social and Behavioral Sciences*) οι τάσεις έρευνας στρέφονται (α) στην ποσοτική μελέτη των γλωσσολογικών φαινομένων χρησιμοποιώντας π.χ. τη θεωρία πιθανοτήτων ή τη θεωρία πληροφορίας κ.α. και (β) στην ενασχόληση με άλλους επιστημονικούς κλάδους που θεωρούνται συγγενικοί και εφαρμόζουν στατιστικές προσεγγίσεις, ξεπερνώντας τα διεπιστημονικά όρια.

Ειδικότερα για τη θεωρία της πληροφορίας, σύμφωνα με τη βιβλιογραφία (Nadel 2005), θεμελιωτής της ήταν ο μαθηματικός Shannon C. E., ο οποίος ανέπτυξε ένα θεωρητικό μοντέλο μετάδοσης πληροφορίας μέσω μηνύματος, που αναπαριστά τη διαδικασία της ανθρώπινης νόησης (cognition) στο πλαίσιο της διαδικασίας επικοινωνίας ως ένα σύστημα επεξεργασίας πληροφορίας (information processing system). Σε αυτό το μοντέλο όρισε την πληροφορία που περιέχει κάποιο μήνυμα για την επικοινωνία μεταξύ δέκτη και αποστολέα και το κανάλι μετάδοσης της, καθιστώντας και τους δύο ορισμούς μετρήσιμους.

Σύμφωνα με τη βιβλιογραφία (Gries to appear in *International Encyclopedia of the Social and Behavioral Sciences*) ένα τρανταχτό **παράδειγμα** χρήσης στατιστικών μεθόδων σε σχέση με τη Γλωσσολογία και άλλους επιστημονικούς κλάδους αποτελεί αυτό της ενασχόλησης με τις **τεχνολογικές εξελίξεις στον τομέα των υπολογιστών και του διαδικτύου**, όπου εμφανίζεται η ανάγκη επεξεργασίας **σωμάτων κειμένου (corpora)** σε σχέση με δεδομένα συχνότητας. Έτσι λοιπόν για όλους τους υπο-κλάδους της Γλωσσολογίας που ερευνούν την επεξεργασία σωμάτων κειμένου η χρήση της στατιστικής και ποσοτικής ανάλυσης είναι απαραίτητη.

Σύμφωνα με τη βιβλιογραφία (Gries to appear in *International Encyclopedia of the Social and Behavioral Sciences*) **η χρήση ποσοτικών μεθόδων έχει καθιερωθεί στη σύγχρονη Γλωσσολογία** και αποτελεί έναν ξεχωριστό **διαρκώς αναπτυσσόμενο** επιστημονικό κλάδο, την **Ποσοτική Γλωσσολογία (ΠΓ)**. Οι πιο **διαδεδομένες** ποσοτικές μέθοδοι (στατιστικά εργαλεία) της ΠΓ αφορούν στον **έλεγχο της ορθότητας υποθέσεων**, όπως π.χ. οι μηδενικές υποθέσεις (null hypothesis). Μάλιστα στη βιβλιογραφία (Gries to appear in *International Encyclopedia of the Social and Behavioral Sciences*) οι έλεγχοι αυτοί χωρίζονται σε 2 κατηγορίες:

1. **Έλεγχοι καλής προσαρμογής** (goodness of fit): ελέγχουν αν τα χαρακτηριστικά ενός συγκεκριμένου συνόλου δεδομένων διαφέρουν από τα καθιερωμένα χαρακτηριστικά γνωστών κατανομών των μελετών του τομέα.
2. **Έλεγχοι ανεξαρτησίας/διαφορών** (independence/differences): ανάλογα το πόσες μεταβλητές εμπεριέχουν διακρίνονται σε μονοπαραγοντικούς και πολυπαραγοντικούς.

1.4 Ανάκτηση Πληροφορίας και Γλωσσολογία

Η ΥΓ, όπως αναφέρθηκε και παραπάνω είναι ένα **διεπιστημονικό πεδίο**, το οποίο εμφανίστηκε περί τα τέλη του 1950 και το οποίο συνδέει τις επιστήμες της **Γλωσσολογίας, της Επιστήμης Υπολογιστών** και της Λογικής. Σύμφωνα με Espunya i Prat (1994) ο κύριος στόχος της ΥΓ είναι η «*κατασκευή υπολογιστικών συστημάτων που κατανοούν και ομιλούν κάθε ανθρώπινη γλώσσα (φυσική γλώσσα)*». Η ΥΓ **εστιάζει την έρευνα της** στην μηχανική μετάφραση μέσω ηλεκτρονικού υπολογιστή, στις διεπαφές μεταξύ ηλεκτρονικού υπολογιστή και χρηστών και στα **ΣΑΠ**.

Όπως έχει ηδη αναφερθεί τα ΣΑΠ προσπαθούν να αντιμετωπίσουν το πρόβλημα της **συνάφειας** στο πλαίσιο της αναζήτησης πληροφορίας. Ανατρέχουν σε συλλογές εγγράφων και ανακτούν έγγραφα **από πηγές που βρίσκονται σε φυσική γλώσσα** έχοντας ως κριτήριο τις πληροφοριακές ανάγκες του χρήστη, οι οποίες επίσης εισάγονται σε φυσική γλώσσα. Το ζήτημα που αφορά την ΥΓ, όπως αναφέρει η Espunya i Prat (1994) είναι **η κατασκευή προγραμμάτων ώστε να μπορούν να αναπαραστήσουν τη γλωσσική πληροφορία με κατανοητό τρόπο για τους Η/Υ**, προγράμματα που μπορούν να καταλάβουν ή να παράγουν υλικό σε φυσική γλώσσα. Βέβαια τα προγράμματα αυτά **αναπαριστούν τις πιο κοινές και προβλέψιμες δομές** των προτάσεων της φυσικής γλώσσας, λόγω της πολυπλοκότητας του εγχειρήματος, η οποία έγκειται σε πολλές παραμέτρους από τη μια γλωσσολογικής φύσεως όπως είναι οι υπο – κλάδοι της Γλωσσολογίας: Μορφολογία, Σύνταξη, Σημασιολογία και από την άλλη μη-γλωσσολογικές, που αφορούν τη μηχανική σχεδίαση των συστημάτων, η οποία αφορά την επιστήμη της μηχανικής (engineering).

Όπως αναφέρεται στη βιβλιογραφία (Brants 2004), **οι πιο σημαντικές βελτιώσεις στο πλαίσιο της ΑΠ προκύπτουν από τις πιο συνηθισμένες εργασίες επεξεργασίας κειμένου** (βλέπε ενότητα 3.1.4). **Τεχνικές που έχουν σχεδιαστεί με αποκλειστικό στόχο την ΑΠ θεωρούνται επιτυχημένες** όπως ο αλγόριθμος Porter που αφορά στη περιστολή και κανονικοποίηση.

Οι Manning, Raghavan και Schutze (2008) αναφέρουν τη **σημαντικότητα γλωσσικής επεξεργασίας στην κατασκευή του ευρετηρίου της συλλογής εγγράφων**, το οποίο όταν προϋπάρχει αυξάνει την αποδοτικότητα της μηχανής αναζήτησης και της ανάκτησης πληροφορίας. Ειδικότερα η γλωσσολογική προ-επεξεργασία παράγει μια λίστα με κανονικοποιημένους όρους για κάθε έγγραφο, που θα αποτελέσουν στη συνέχεια τους όρους ευρετηρίου, το λεξιλόγιο δηλαδή του ΣΑΠ που αποτελείται από το σύνολο ευρετηριασμένων όρων. Όπως αναφέρουν οι Lewis και Sparck Jones (1996), **η χρήση των όρων ευρετηρίου που αφορά σε τεχνικές Επεξεργασίας Φυσικής Γλώσσας αυξάνεται ολοένα και περισσότερο** αν και η ΑΠ χρησιμοποιεί παραδοσιακά κυρίως ελεγχόμενα λεξιλόγια. Μάλιστα τονίζουν ακόμη την μεγάλη ανάγκη επιστημονικής διερεύνησης στο πεδίο αυτό σχετικά με αποτελεσματικότερες προσεγγίσεις. Σύμφωνα με Strzalkowski et al. (1999) η γλωσσολογική επεξεργασία (Επεξεργασία Φυσικής Γλώσσας) μπορεί να εφαρμοστεί επιτυχώς και έχει βελτιωθεί με την πάροδο του χρόνου σε μεγάλο βαθμό ώστε να μπορεί να εφαρμοστεί στα προβλήματα της ΑΠ αν και υπάρχουν πάντοτε περιθώρια

βελτίωσης. Σύμφωνα με την Sparck και τον Kay (1977) ένα πρόβλημα σε σχέση με την ενσωμάτωση της Γλωσσολογίας στην ΑΠ είναι πως **αφορούν διαφορετικής κλίμακας μάζες**, καθώς η Γλωσσολογία αφορά σε μικρότερες μονάδες λόγου ενώ η ΑΠ ασχολείται με μάζες εγγράφων και με πιο μαζικά χαρακτηριστικά τους.

1.5 Ερευνητικά προβλήματα

Μέχρι στιγμής ο κλάδος της επιστήμης της ΑΠ σε συνδυασμό με την Γλωσσολογία (ΥΓ και ΠΓ) αποφέρει σημαντικές βελτιώσεις στον τομέα της αναζήτησης. Όμως υπάρχουν ανοικτά προβλήματα που από τη μια πλευρά έχουν σχέση με την κατανόηση της γραφής/φυσικής γλώσσας από τους υπολογιστές (Espunya i Prat 1994) και από την άλλη πλευρά αφορούν το συντακτικό τρόπο γραφής, ο οποίος παρουσιάζει πολλά προβλήματα όσον αφορά την κατανόηση (π.χ. πολυσημία).

Από τη σκοπιά του κλάδου των Νευροεπιστημών, όπως αναφέρθηκε παραπάνω, υποδεικνύεται ότι ο τρόπος σύνταξης και διάρθρωσης του γραπτού λόγου (Ahlsen 2006) έχει άμεση σχέση με τη δυνατότητα κατανόησης του εγκεφάλου, ο οποίος συσχετίζεται με τη φυσική μνήμη του ανθρώπου, καθώς όμως και με δεδομένα που άπτονται ποσοτικών παρατηρήσεων για να επεξηγήσουν την ανθρώπινη συμπεριφορά. Σύμφωνα με τον Altmann (2002) η «*αρχή της Ελάχιστης Προσπάθειας*» αφορά την τάση μείωσης της προσπάθειας στο πλαίσιο της επικοινωνίας, όταν μια λέξη/φράση επαναλαμβάνεται συχνά, να μειώνεται και η ίδια η λέξη/φράση, χωρίς βέβαια να εξαφανίζεται εντελώς. Για το λόγο αυτό, η ΠΓ προσπαθεί να ερμηνεύσει τον τρόπο σύνδεσης των λέξεων μεταξύ τους ως προς τη συχνότητα και την ποσότητα με στατιστικό και νομοτελειακό τρόπο (βλέπε ενότητα 3.3.1).

1.6 Σκοπός και στόχος

Σκοπός της παρούσας διατριβής είναι η δημιουργία νέων τεχνικών ΑΠ, οι οποίες όμως είναι συνδεδεμένες με ζητήματα της ΠΓ και ιδιαίτερα των μικρών κειμένων. Προς αυτήν την κατεύθυνση δημιουργήθηκε μια σειρά πειραματικών διαδικασιών στις οποίες χρησιμοποιήθηκαν μικρά κείμενα, που συνήθως απαρτίζονται από μερικές προτάσεις ή τίτλους επιστημονικών δημοσιεύσεων. Χρησιμοποιήθηκε ένας νέος αλγόριθμος βασισμένος στο διανυσματικό μοντέλο

ανάκτησης πληροφορίας VSM και στη συνέχεια αυτή η διαδικασία συνδέθηκε με φλέγοντα ζητήματα της ΠΓ.

Απώτερος στόχος της διατριβή αυτής είναι η διασύνδεση των Νευροεπιστημών με τους επιστημονικούς κλάδους της ΑΠ και της Γλωσσολογίας, ΥΓ και ΠΓ.

1.7 Διάρθρωση της εργασίας

Η διατριβή ξεκινά από τις ευχαριστίες και την περίληψη στην ελληνική και αγγλική γλώσσα. Στη συνέχεια αναπτύσσεται σε έξι βασικά κεφάλαια ως εξής:

Στο «*Κεφάλαιο 1ο – Εισαγωγή*» πραγματοποιείται μια αναλυτική παρουσίαση των επιμέρους επιστημονικών κλάδων, οι οποίοι μελετήθηκαν για τη διεξαγωγή της παρούσας διατριβής, δηλαδή στις ΑΠ, ΥΓ και ΠΓ. Ακόμη αναλύονται ζητήματα σχετικά με τη διεπιστημονική τους φύση και με άλλους κλάδους όπως αυτόν της Νευροεπιστήμης. Επίσης παρουσιάζονται ζητήματα που αφορούν το πως οι τρεις ανωτέρω κλάδοι διαλειτουργούν με στόχο την επίλυση προβλημάτων σχετικά με την κατανόηση δομής προτάσεων.

Στη συνέχεια στο «*Κεφάλαιο 2ο - Ανάκτηση Πληροφοριών*» πραγματοποιείται επισκόπηση της διαδικασίας ΑΠ και αναλύονται οι βασικές έννοιες της ΑΠ και πως αυτές αλληλεπιδρούν στο πλαίσιο του διαδικτύου. Δίνεται έμφαση στα μοντέλα, στην αξιολόγηση αποτελεσμάτων, στην αρχιτεκτονική συστημάτων και ειδικότερα στις μηχανές αναζήτησης που σχετίζονται με τη σύγχρονη ΑΠ.

Στο «*Κεφάλαιο 3ο - Υπολογιστική και Ποσοτική Γλωσσολογία*» παρουσιάζεται μια λεπτομερής ανάλυση ζητημάτων που αφορούν το κείμενο, καθώς αυτό αποτελεί το βασικό αντικείμενο της ΑΠ. Αναλύονται ζητήματα που αφορούν τις ιδιότητες, την επεξεργασία και τις παραμέτρους που καθιστούν ποιοτικά μετρήσιμο ένα κείμενο. Στη συνέχεια αναλύονται ζητήματα που εμπίπτουν από τη μια πλευρά στον επιστημονικό κλάδο της ΥΓ (Μορφολογία, Σύνταξη, Σημασιολογία) και από την άλλη σε αυτόν της ΠΓ (στατιστικοί νόμοι ΠΓ).

Στο «*Κεφάλαιο 4ο – Στατιστικά Θέματα*» γίνεται εκτενής ανάλυση στατιστικών ζητημάτων. Αρχικά πραγματοποιείται μια σύντομη εισαγωγή στη Στατιστική και στη συνέχεια δίνεται έμφαση σε επιμέρους ορολογία και ορισμούς ώστε να παρουσιαστούν τα στατιστικά εργαλεία που χρησιμοποιήθηκαν στο πλαίσιο της διατριβής.

Στο «*Κεφάλαιο 5ο - Προτεινόμενο Μοντέλο Ανάκτησης Σημαινόντων Όρων Εγγράφων*» αναλύεται το προτεινόμενο μοντέλο που αναπτύχθηκε στο πλαίσιο της υλοποίησης του σκοπού της διατριβής. Ακόμη παρατίθενται οι επιστημονικές δημοσιεύσεις που παρήχθησαν μέσω της επιστημονικής έρευνας κατά τη διάρκεια εκπόνησης της διατριβής.

Στο «*Κεφάλαιο 6ο – Συμπεράσματα*» παρουσιάζονται αναλυτικά: η καινοτομία που επιφέρει η διατριβή στον επιστημονικό χώρο, τα συμπεράσματα που προέκυψαν από την έρευνα και τα ανοικτά επιστημονικά ζητήματα που εγείρονται.

Τέλος, η διατριβή περιέχει ένα λεξικό (στην ελληνική και αγγλική γλώσσα) με την ορολογία που χρησιμοποιήθηκε για τη συγγραφή της διατριβής, δύο παραρτήματα που επεξηγούν το προτεινόμενο μοντέλο, καθώς και εκτενή βιβλιογραφία.

ΚΕΦΑΛΑΙΟ 2Ο

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

2.1 Εισαγωγή στην διαδικασία της ΑΠ

Το κεφάλαιο αυτό ασχολείται με το θεωρητικό υπόβαθρο της διδακτορικής διατριβής στο οποίο εντάσσονται οι διαδικασίες εξόρυξης γνώσης μέσω των πληροφοριών που διαμοιράζει το διαδίκτυο. Προκειμένου αυτή η διαδικασία να περιγραφεί με ένα φορμαλιστικό τρόπο επιλέχθηκε ως μοντέλο σχεδιασμού το σύστημα ταξινόμησης της ACM (ανακτήθηκε 20/02/2015 από http://www.acm.org/about/class/2012).

Σύμφωνα με την ταξινόμηση αυτή, δίνεται έμφαση σε έναν επιστημονικό υπο-κλάδο των πληροφοριακών συστημάτων αυτόν της ΑΠ, ο οποίος βέβαια είναι άμεσα συνυφασμένος με αυτόν του Παγκόσμιου Ιστού όπως φαίνεται και στην εικ. 1.

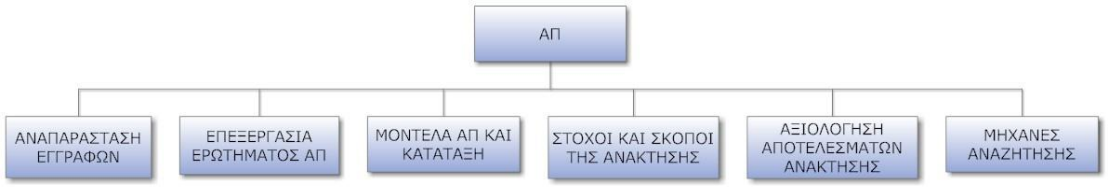
Εικ. 1 Υπο-κλάδοι πληροφοριακών συστημάτων σύμφωνα με ACM Classification

Ο παραπάνω κλάδος της ΑΠ επιλέχθηκε για τον εξής λόγο: διότι ασχολείται με όλες εκείνες τις διαδικασίες που αφορούν την αναπαράσταση εγγράφων ως επίσης και τις διαδικασίες ανάλυσης και αξιολόγησης. Ο δε κλάδος του Παγκόσμιου Ιστού αποτελεί ένα δυναμικό μέσο παραγωγής εγγράφων και έτσι στη σημερινή πραγματικότητα αυτοί οι δύο κλάδοι αποτελούν δύο αναπόσπαστα τμήματα διαλειτουργίας. Στην παρούσα διατριβή ο κλάδος του Παγκόσμιου Ιστού μελετάται κυρίως υπο το πρίσμα της διαδικασίας αναζήτησης και ανάκτησης πληροφορίας, χωρίς να περιλαμβάνει άλλες παραμέτρους του κλάδου αυτού. Η διδακτορική διατριβή χρησιμοποιεί ως μέσο τις πρακτικές ανάλυσης αυτών των εγγράφων με σκοπό να ευρεθεί ένα πρότυπο (pattern), το οποίο θα αποτελέσει δυνητικά ένα κυρίαρχο δομικό λίθο στον κλάδο της ΑΠ.

Σε αυτό το σημείο σκόπιμος είναι ο ορισμός της λέξης «**έγγραφο**», ο οποίος απαντάται πολύ συχνά στη διεθνή βιβλιογραφία της ΑΠ. Όπως αναφέρεται στην βιβλιογραφία (Baeza-Yates και Ribeiro-Neto 2011, Manning, Raghavan και Schutze 2008), οι τύποι πληροφοριακών αντικειμένων στην ΑΠ ποικίλουν, αν και η ΑΠ

εστιάζει κυρίως στην ανάκτηση εγγράφων (documents). Ο όρος **έγγραφο στην ΑΠ** χρησιμοποιείται υπό μια ευρεία έννοια και υποδηλώνει **την βασική μονάδα που απαρτίζει ένα ΣΑΠ**, είτε αυτή είναι κείμενο είτε οποιαδήποτε άλλη μορφή καθώς η πληροφορία που αναζητά ο χρήστης μπορεί να βρίσκεται σε οποιαδήποτε μορφή, όπως σε ένα αρχείο ήχου ή βίντεο. **Τα έγγραφα απαρτίζονται από δεδομένα** τα οποία και αυτά διακρίνονται σε 2 κατηγορίες: **τα δομημένα και αδόμητα** δεδομένα. Με βάση τους Manning, Raghavan και Schutze (2008), αδόμητα ονομάζονται τα δεδομένα εκείνα τα οποία για έναν υπολογιστή δεν θεωρούνται σαφή και σημασιολογικά εμφανή ως προς τη δομή τους. Δομημένα δεδομένα είναι το αντίθετο των αδόμητων δεδομένων. Παρόλα αυτά πάντοτε υπάρχει μια τυπική δομή στα εκάστοτε κείμενα, η οποία προέρχεται από τη χρήση της γλώσσας, καθώς απαιτεί συγκεκριμένη σύνταξη. Επιπλέον τα περισσότερα κείμενα διαθέτουν δομή, η οποία πηγάζει από το πως τα έχει διαμορφώσει ο δημιουργός τους τοποθετώντας επικεφαλίδες, παραγράφους, υποσημειώσεις. Π.χ. στο διαδίκτυο για τα έγγραφα η δομή αποδίδεται μέσω των γλωσσών σήμανσης (markup).

Στο παρακάτω διάγραμμα (βλέπε εικ. 2) φαίνονται οι κύριοι άξονες **της ΑΠ:**

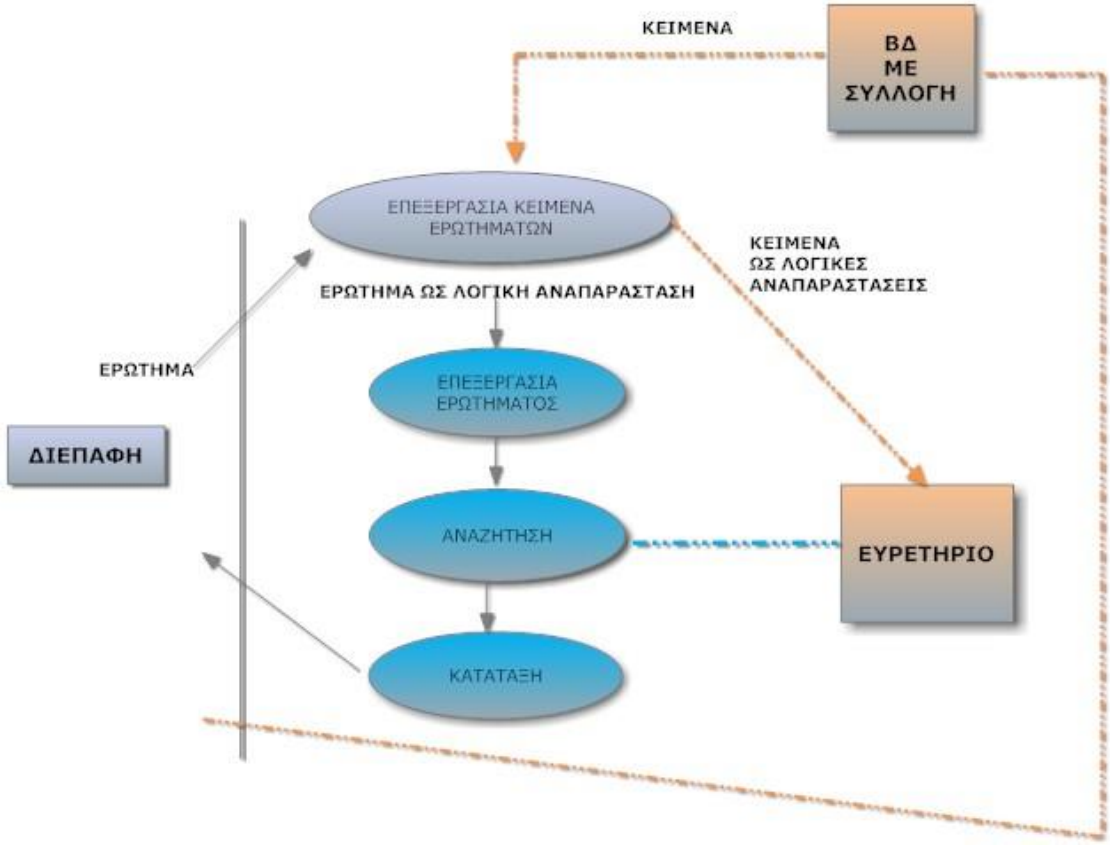


Εικ.2 Επισκόπηση ΑΠ σύμφωνα με ACM Classification

Στην παρούσα διατριβή θα αναπτυχθούν κυρίως οι ενότητες **«μοντέλα ΑΠ και κατάταξη», δίνοντας έμφαση στο μοντέλο VSM καθώς και «μηχανές αναζήτησης», εστιάζοντας στην αρχιτεκτονική τους, η οποία αφορά στα ζητήματα ανάκτησης πληροφορίας.** Προκειμένου να υπάρχει μια ολοκληρωμένη παρουσίαση είναι απαραίτητη η παρουσίαση της διαδικασίας ΑΠ συνοπτικά. Οι στόχοι της ΑΠ παρουσιάστηκαν στο προηγούμενο κεφάλαιο.

Συνοπτικά η διαδικασία ΑΠ έγκειται στο ερώτημα που εισάγει ένας χρήστης σε μια διεπαφή υπολογιστή προκειμένου να αναζητήσει - ανακτήσει την πληροφορία που χρειάζεται. Στην συνέχεια το ερώτημα υφίσταται ένα σύνολο επεξεργασιών (κατάλληλες μετατροπές ώστε να είναι επεξεργάσιμο από τον υπολογιστή) ώστε τέλος να πραγματοποιηθεί αναζήτηση σε μια πηγή πληροφοριών, όπως μια βάση δεδομένων και τέλος να επιστραφούν τα σχετικά αποτελέσματα στον χρήστη.

Μάλιστα σύμφωνα με τους Ceri et al. (2013) καθώς η ΑΠ ασχολείται κυρίως με δεδομένα κειμένου (έγγραφα και ερωτήματα σε μορφή κειμένου), ένα σύνολο κειμενικών διεργασιών είναι απαραίτητο. Το παρακάτω διάγραμμα παρουσιάζει αναλυτικά τη διαδικασία ΑΠ.



Εικ. 3 Διαδικασία ΑΠ

Σύμφωνα με τους Ceri et al. (2013) η διαδικασία ΑΠ συνοψίζεται στην εικ. 3 και αναλύεται ως ακολούθως: ο χρήστης εισάγει στην διεπαφή (μπορεί να είναι κάποιος φυλλομετρητής) κάποιες λέξεις κλειδιά που αποτελούν το ερώτημα του και εκφράζουν την πληροφοριακή ανάγκη. Εν συνεχεία το ερώτημα αυτό υφίσταται ένα σύνολο διεργασιών για την επεξεργασία του ως κείμενο καθώς βρίσκεται σε φυσική γλώσσα, διεργασίες όμοιες με αυτές που πραγματοποιούνται στα αποθηκευμένα έγγραφα του συστήματος κατά την ευρετηρίαση (βλέπε ενότητα 2.4.1.1). Ειδικότερα για την ΑΠ σε κειμενική πληροφορία είναι αναγκαία η μετατροπή τόσο του ερωτήματος χρήστη όσο και των εγγράφων σε λογικές αναπαραστάσεις. Στην συνέχεια το ερώτημα υποβάλλεται σε επιμέρους διεργασίες, πραγματοποιείται αναζήτηση στη βάση (η ταχύτητα της αναζήτησης εξαρτάται από το ευρετήριο) και επιστρέφονται συναφή αποτελέσματα στο χρήστη σε μια λίστα κατάταξης σύμφωνα

με τη συνάφεια ως προς την πληροφοριακή ανάγκη του χρήστη. Καθώς ο χρήστης περιηγείται στη λίστα μπορεί να εντοπίσει κάποιο υποσύνολο που θεωρεί πολύ συναφές, γνωστό ως ανάδραση (**feedback)** που καταχωρείται στο σύστημα.

2.2 Μοντέλα ΑΠ και κατάταξη

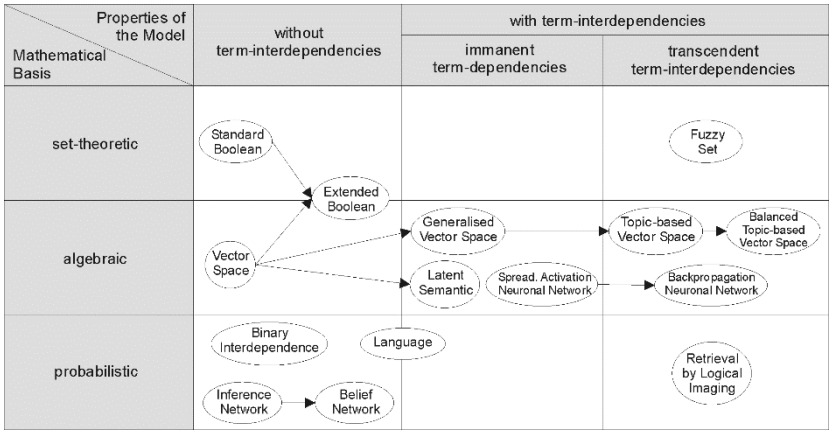
Η αναπαράσταση εγγράφων και η διαδικασία επεξεργασίας ερωτήματος χρήστη έχουν επεξηγηθεί συνοπτικά στην παραπάνω ενότητα. Έτσι σε αυτήν την ενότητα θα παρουσιαστούν τα μοντέλα ΑΠ (IR models) δίνοντας έμφαση στο VSM, το οποίο αποτελεί βάση της παρούσας διατριβής.

Τα μοντέλα ΑΠ αφορούν τη **διευκρίνηση της συνάφειας εγγράφων** σε σχέση με ένα ερώτημα χρήστη. Μάλιστα διαρκώς αξιολογούνται, ώστε να αναβαθμίζονται ικανοποιητικά προκειμένου να λειτουργούν όσο το δυνατόν αποτελεσματικότερα. Σύμφωνα με τους Baeza-Yates και Ribeiro-Neto (2011), η διάκριση των συναφών εγγράφων από τα υπόλοιπα της συλλογής, εξαρτάται κυρίως από τον αλγόριθμο κατάταξης (θεωρείται πυρήνας των ΣΑΠ) και από τις βασικές προϋποθέσεις που χρησιμοποιεί αυτός. Έτσι οι διαφορετικές παραλλαγές συνόλων προϋποθέσεων αποτελούν διαφορετικά μοντέλα ΑΠ, τα οποία χωρίζονται σε κάποιες επιμέρους κατηγορίες.

Οι Baeza-Yates και Ribeiro-Neto (2011) αναφέρουν τα **κλασσικά μοντέλα ΑΠ**:

1. **Λογικό (Boolean)**: βασίζεται στη θεωρία Συνόλων και τα έγγραφα και τα ερωτήματα αναπαρίστανται ως σύνολα όρων ευρετηρίου.
2. **Διανυσματικό (Vector)**: βασίζεται στη θεωρία Διανυσμάτων (ευρύτερα θεωρείται αλγεβρικό μοντέλο), στο οποίο τα αντίστοιχα έγγραφα και ερωτήματα αναπαρίστανται ως διανύσματα σε πολυδιάστατο διανυσματικό χώρο.
3. **Πιθανοτικό (Probabilistic)**: βασίζεται στη θεωρία Πιθανοτήτων.

Παρακάτω φαίνεται σχηματικά (βλέπε εικ. 4) μια λεπτομερέστατη **ταξινόμηση των μοντέλων ΑΠ** όπως παρουσιάζεται στη βιβλιογραφία (Kuropka 2004), όπου ο διαχωρισμός τους πραγματοποιείται από τη μια πλευρά σύμφωνα με την **μαθηματική βάση** των μοντέλων και από την άλλη με την ύπαρξη ή όχι **αλληλεξάρτησης των όρων**.



Εικ. 4 ταξινόμηση μοντέλων ΑΠ σύμφωνα με Kuropka (2004)

Στα κλασικά μοντέλα ΑΠ παραπάνω αναφέρθηκαν οι μαθηματικές βάσεις. Όσον αφορά τις εξαρτήσεις όρων οι Bendersky και Croft (2012) παρατηρούν μια αλλαγή ενδιαφέροντος προς μοντέλα ΑΠ που ενσωματώνουν εξαρτήσεις όρων την τελευταία δεκαετία, με παραδείγματα τα μοντέλα *Markov random fields* και *linear discriminant model*. Σύμφωνα με τους Metzler και Croft (2005) **οι εξαρτήσεις όρων** σε μια συλλογή εγγράφων υπάρχουν και εστιάζουν κυρίως σε **ζεύγη όρων**. Μάλιστα αναφέρουν ότι η εξάρτηση όρων μελετάται είτε **στο πλαίσιο των φράσεων, όσον αφορά την εγγύτητα** μεταξύ των όρων ή εστιάζει στις **εμφανίσεις των όρων που εξετάζονται από κοινού** σε έγγραφα. Όπως φαίνεται και από το παραπάνω διάγραμμα (εικ. 4) οι **παραλλαγές μοντέλων ΑΠ είναι πάρα πολλές** ώστε να επεξηγηθούν αναλυτικά. Παρακάτω **θα αναπτυχθεί το διανυσματικό μοντέλο VSM** το οποίο παρουσιάζει ιδιαίτερο ενδιαφέρον για την παρούσα διατριβή.

2.2.1 VSM

Το VSM απαντάται στη βιβλιογραφία (Dubin 2004) ως ένα μαθηματικό μοντέλο ιδιαιτέρως πολυχρηστικό καθώς μπορεί να εφαρμοστεί σε πολλούς επιστημονικούς τομείς αλλά και ειδικότερα στην ΑΠ αναπαριστώντας έγγραφα ως διανύσματα. Το VSM είναι αποτέλεσμα της έρευνας κατά την ανάπτυξη του ΣΑΠ *System for the Mechanical Analysis and Retrieval of Text (SMART)* στο Πανεπιστήμιο Cornell, με επικεφαλής της ερευνητικής ομάδας τον **Gerard Salton** (βλέπε ενότητα 1.1).

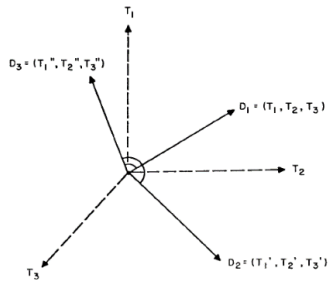
Ειδικότερα, το VSM ανήκει στα στατιστικά μοντέλα (Spoerri 1995), τα οποία για να καθορίσουν τη συνάφεια μεταξύ των εγγράφων μιας συλλογής και του ερωτήματος του χρήστη χρησιμοποιούν τη συχνότητα εμφάνισης των όρων στα

έγγραφα και επιστρέφουν ως αποτελέσματα λίστες κατάταξης εγγράφων ανάλογα με τη συνάφεια τους ως προς το ερώτημα του χρήστη.

Όπως αναφέρουν οι Raghavan και Wong (1986) η κεντρική ιδέα του VSM στηρίζεται στην **αναπαράσταση κάθε αντικειμένου ΑΠ** (λέξη, έγγραφο κλπ.) **ως διάνυσμα** ενώ οι Turney και Pantel (2010) συμπληρώνουν πως **κάθε έγγραφο μιας συλλογής αντικατοπτρίζεται ως διάνυσμα σε ένα διανυσματικό χώρο**. (όπως έχει ηδη αναφερθεί η έννοια του εγγράφου είναι ευρύτερη και αντικατοπτρίζει το αντικείμενο ανάκτησης ανάλογα με τη συλλογή). Μάλιστα όπως αναφέρεται στη βιβλιογραφία (Lan, Tan και Su 2009) κάθε όρος καθορίζεται από ένα **δείκτη σημασίας**, ο οποίος **αντικατοπτρίζει τη σημαντικότητα του** σε σχέση με το περιεχόμενο του εκάστοτε εγγράφου και αναπαριστά πόσο ο όρος αυτός συνεισφέρει στη σημασιολογία του εγγράφου. Για τους δείκτες αυτούς θα γίνει ανάλυση στην επόμενη ενότητα (βλέπε ενότητα 2.2.2).

Μια από τις βασικές παραδοχές του VSM σύμφωνα με τους Raghavan και Wong (1986) είναι πως το σύστημα στο οποίο υπάρχει ο διανυσματικός χώρος **είναι γραμμικό**. Σε ένα γραμμικό σύστημα ισχύουν βασικοί κανόνες και αξιώματα της άλγεβρας όπως η ιδιότητα πρόσθεσης δύο συστατικών του συστήματος από τα οποία προκύπτει ένα άλλο συστατικό του συστήματος και η δυνατότητα πολλαπλασιασμού κάθε συστατικού του συστήματος με πραγματικό αριθμό.

Χρησιμοποιώντας τη θεωρία διανυσμάτων **τα έγγραφα συγκρίνονται ως προς την ομοιότητα** (βλέπε ενότητα 3.2.3.2) τους. Όπως αναφέρεται στην βιβλιογραφία (Spoerri 1995) τα έγγραφα μετατρέπονται σε **διανύσματα** και τοποθετούνται **σε πολυδιάστατο χώρο**, τις **διαστάσεις του οποίου αποτελούν οι όροι που περιέχονται στα έγγραφα**. Στην εικ. 5 παρατίθεται ένα παράδειγμα αναπαράστασης διανυσμάτων εγγράφων σε τρισδιάστατο χώρο, με D να αντιστοιχεί στο έγγραφο και T στον όρο εγγράφου. Όπως μπορεί κανείς να διακρίνει οι διαστάσεις στο σχήμα αντιστοιχούν στον αριθμό των όρων εγγράφων.



Εικ. 5 Αναπαράσταση διανυσμάτων εγγράφων στον πολυδιάστατο χώρο από Salton, Wong and Yang (1975)

Οι όροι που περιέχονται στα έγγραφα χρησιμοποιούνται προκειμένου να δημιουργηθεί ένα ευρετήριο για την αναπαράσταση των εγγράφων της συλλογής και μέσω λεξιλογικής ανάλυσης πρέπει να καθοριστούν οι σημαντικοί όροι ως προς το σημασιολογικό περιεχόμενο του κάθε εγγράφου. **Όσο μεγαλύτερη είναι η απόσταση μεταξύ των διανυσμάτων τόσο λιγότερη ομοιότητα έχουν ενώ όσο πιο κοντά βρίσκονται στο διανυσματικό χώρο τόσο πιο κοντά είναι και ως προς το περιεχόμενο τους. Τα ερωτήματα του χρήστη και αυτά μετατρέπονται σε διανύσματα** ώστε να είναι εφικτή η σύγκριση μεταξύ των εγγράφων της συλλογής και του ερωτήματος και τελικά να επιστραφούν τα σχετικότερα αποτελέσματα στον χρήστη σε μορφή μιας λίστας κατάταξης.

Αναλυτικότερα σύμφωνα με τους Salton και Buckley (1988) η αναπαράσταση των εγγράφων από όρους - διανύσματα (term vectors) περιγράφεται από το μαθηματικό τύπο (1):

$$D = (t_1, t_j, \ldots, t_p) \tag{1}$$

Κάθε t_k αντιστοιχεί σε έναν όρο που ανατίθεται σε κάποιο έγγραφο D. Καθώς τα ερωτήματα των χρηστών αναπαρίστανται και αυτά είτε σε διανυσματική μορφή είτε ως λογικές δηλώσεις (boolean statements), ο παραπάνω τύπος μετατρέπεται ως ακολούθως όπως φαίνεται στους μαθηματικούς τύπους (2) και (3):

$$Q = (q_a, q_b, \ldots, q_r) \tag{2}$$

$$Q = (q_a \; and \; q_b) \text{ Or } (q_c \; and \; q_d \; and \; \ldots) \text{ or } \ldots \tag{3}$$

Όπου q_k αναπαριστά έναν όρο ο οποίος έχει ανατεθεί σε ένα ερώτημα Q.

Σύμφωνα με τους Salton και Buckley (1988) αν w_{dk} (ή w_{qk}) αναπαριστά το βάρος (<u>βλέπε ενότητα 2.2.2</u>) του όρου t_k στο έγγραφο D (ή το ερώτημα Q) και για την αναπαράσταση του περιεχομένου διατίθενται t όροι, τότε οι τύποι (1) και (2) μετατρέπονται ως ακολούθως στους μαθηματικούς τύπους (4) και (5) αντίστοιχα:

$$D = (t_0 w_{d_0}; t_1 w_{d_1}; \ldots; t_t w_{d_t}) \quad (4)$$

$$Q = (q_0 w_{q_0}; q_1 w_{q_1}; \ldots; q_t w_{q_t}) \quad (5)$$

Όταν ο όρος k δεν έχει ανατεθεί στο έγγραφο D ή στο ερώτημα Q τότε το βάρος του όρου w_{dk} ή w_{qk} είναι ίσο με 0 ενώ στην αντίστροφη περίπτωση είναι ίσο με το 1.

2.2.2 Όροι στάθμισης (βάρη)

Η απόδοση βάρους στους όρους αποτελεί ένα πολύ σημαντικό βήμα για την βελτίωση της αποτελεσματικότητας στην ΑΠ. Όπως αναφέρεται στη βιβλιογραφία (Lan, Tan και Su 2009) οι μέθοδοι απόδοσης βάρους που αναφέρονται είναι οι παρακάτω: (α) δυαδική (binary) αναπαράσταση, (β) πιο ευρέως χρησιμοποιούμενη *tf-idf* και παραλλαγές της. Όπως αναφέρουν και οι Salton και Buckley (1988) στόχος της απόδοσης βαρών σε όρους είναι η βελτίωση της **αποτελεσματικότητας (effectiveness) της ανάκτησης**, δηλαδή η βελτίωση της διαδικασίας ανάκτησης συναφών εγγράφων με βάση τις πληροφοριακές ανάγκες του χρήστη αλλά ταυτόχρονα και στην εκκαθάριση της λίστας κατάταξης από μη συναφή έγγραφα. Τα δύο βασικά μέτρα για την επίτευξη των παραπάνω επιμέρους στόχων είναι η **ανάκληση και η ακρίβεια,** μέσω των οποίων επιτυγχάνεται η αξιολόγηση των ΣΑΠ.

Επεξηγώντας αναλυτικότερα τις έννοιες της ανάκλησης και ακρίβειας οι Salton και Buckley (1988) τονίζουν ότι είναι **αναγκαίοι οι συμβιβασμοί στις απαιτήσεις** ως προς τις έννοιες αυτές σε σχέση με ένα ΣΑΠ. Ιδανικά ένα ΣΑΠ θα έπρεπε να έχει υψηλή ανάκληση και υψηλή ακρίβεια, μα καθώς η ανάκληση σχετίζεται με την υψηλή συχνότητα όρων στα έγγραφα της συλλογής ενώ η υψηλή ακρίβεια από τη χρήση σπανιότερων όρων, το ακριβώς αντίθετο δηλαδή είναι κατανοητή η ανάγκη για **συμβιβασμό μεταξύ των επιπέδων των δύο μέτρων. Προκειμένου να επιτευχθεί ο συμβιβασμός αυτός** στη βιβλιογραφία (Lan, Tan και Su 2009) αναφέρονται οι **παράγοντες απόδοσης βάρους** συνοπτικά και χωρίζονται σε τρεις κατηγορίες:

- *Παράγοντες συχνότητας εμφάνισης όρου (term frequency factors)*
    - Δυαδική αναπαράσταση (0 απουσία όρου, 1 παρουσία όρου), *tf* κ.α.
- *Παράγοντες συχνότητας εμφάνισης όρου στην συλλογή* (*collection Frequency Factor).*
- *Παράγοντες κανονικοποίησης (normalization Factor)*
    - Για την εξίσωση του διαφορετικού μήκους των εγγράφων ώστε να περιοριστεί το βάρος όρου στο διάστημα (0,1). Είναι προφανές ότι στη δυαδική αναπαράσταση η κανονικοποίηση δεν έχει χρησιμότητα αφού οι τιμές είναι ήδη 0 ή 1.

Αναλυτικότερα οι Salton και Buckley (1988) παρουσιάζουν τους πιο δημοφιλής παράγοντες για την απόδοση βάρους *tf-idf*. Ο παράγοντας συχνότητας εμφάνισης όρων *tf* (term frequency) μετρά τη συχνότητα εμφάνισης όρων στα έγγραφα ή τα ερωτήματα και αφορά την ανάκληση. Όμως η συχνότητα εμφάνισης των όρων δεν αρκεί διότι αν οι συχνότεροι όροι δεν βρίσκονται μονάχα σε κάποια μεμονωμένα έγγραφα αλλά είναι διάσπαρτοι σε όλη την συλλογή τότε είναι πιθανό να ανακτηθούν όλα τα έγγραφα με αποτέλεσμα τα επίπεδα της ακρίβειας να είναι ιδιαιτέρως χαμηλά.

Έτσι ο παράγοντας *idf* (inverse document frequency), ο οποίος εξαρτάται από τη συλλογή έρχεται για να εξισορροπήσει τα επίπεδα ανάκλησης - ακρίβειας. Ένας idf παράγοντας υπολογίζεται από $\log {}^{N}/_{n}$, όπου n τα έγγραφα στα οποία έχει ανατεθεί ένας όρος σε μια συλλογή και N το σύνολο των εγγράφων της συλλογής. Σύμφωνα πάντα με τους Salton και Buckley (1988) οι καλύτεροι όροι θα πρέπει από την μια πλευρά να έχουν υψηλή συχνότητα εμφάνισης όρων και από την άλλη χαμηλές συνολικά συχνότητες στη συλλογή, ώστε να απομονώσουν τα συναφή έγγραφα στο πλαίσιο μιας συλλογής. Έτσι ο υπολογισμός της σημαντικότητας των όρων πραγματοποιείται μέσω του γινόμενου: $tf \times idf$.

Τέλος, οι Salton και Buckley (1988) αναφέρονται στον παράγοντα κανονικοποίησης που καθιστά εφικτή την αντιμετώπιση όλων των συναφών εγγράφων ως ίσης σημασίας στην ανάκτηση. Καθώς όλα τα έγγραφα δεν έχουν το ίδιο μήκος, η σύγκριση τους χωρίς κανονικοποίηση δεν θεωρείται έγκυρη αφού ανάλογα με το μήκος των εγγράφων διαμορφώνεται και το μήκος των διανυσμάτων τους. Έτσι τα μικρού μήκους έγγραφα (short documents) τείνουν να αναπαρίστανται από μικρά διανύσματα όρων και αντίστοιχα τα μεγάλου μήκους έγγραφα

αναπαρίστανται από μεγάλα διανύσματα όρων. Συνεπώς τα μεγαλύτερου μήκους έγγραφα έχουν καλύτερες πιθανότητες ανάκτησης.

Σύμφωνα με τους Salton και Buckley (1988) από τις εξισώσεις (4) και (5) μπορεί να εξαχθεί μια τιμή ομοιότητας μεταξύ εγγράφων και ερωτήματος, συγκρίνοντας τα διανύσματα τους, όπως φαίνεται στον παρακάτω μαθηματικό τύπο (6):

$$similarity(Q, D) = \sum_{k=1}^{t} w_{qk} \cdot w_{dk} \tag{6}$$

Μάλιστα σύμφωνα με τους Salton και Buckley (1988) κάθε βάρος όρου, εξαρτάται σε κάποιο βαθμό από τα βάρη των υπόλοιπων όρων στο ίδιο διάνυσμα. Ένα τυπικό βάρος όρου που χρησιμοποιεί παράγοντα κανονικοποίησης μήκους διανύσματος φαίνεται στο μαθηματικό τύπο (7):

$$\frac{w_{dk}}{\sqrt{\sum_{vector} (w_{dk})^2}} \tag{7}$$

Τότε ο τύπος (6) σε συνδυασμό με τον παράγοντα κανονικοποίησης (7) δίνουν ως αποτέλεσμα τον τύπο *cosine vector similarity* (8) :

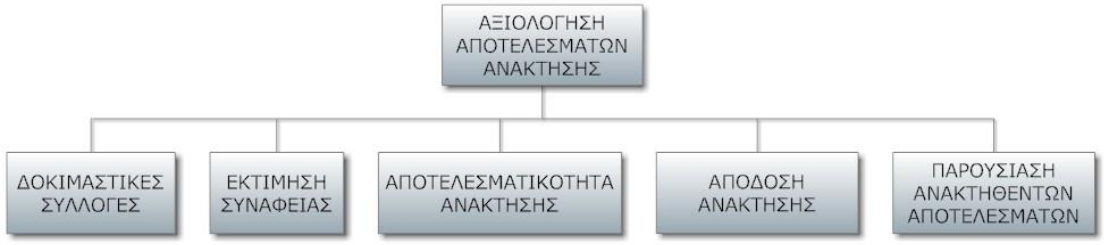
$$similarity(Q, D) = \frac{\sum_{k=1}^{t} w_{qk} \cdot w_{dk}}{\sqrt{\sum_{k=1}^{t} (w_{qk})^2 \cdot \sum_{k=1}^{t} (w_{dk})^2}} \tag{8}$$

2.3 Αξιολόγηση αποτελεσμάτων στην ΑΠ

Ένα βασικό κεφάλαιο στην ΑΠ αφορά στην αξιολόγηση των ΣΑΠ. Όπως έχει ήδη αναφερθεί η καρδιά των συστημάτων αυτών είναι τα μοντέλα ΑΠ και πιο συγκεκριμένα ο αλγόριθμος κατάταξης, με βάση τον οποίο παρουσιάζονται τα αποτελέσματα στο χρήστη. Εξ αρχής **ο απώτερος στόχος** της ΑΠ είναι η **ικανοποίηση των πληροφοριακών αναγκών του χρήστη.** Μιλάμε λοιπόν για **χρηστο-κεντρική αξιολόγηση** της αναζήτησης που εκτελεί το ΣΑΠ σύμφωνα με τους Croft et al. (2010), καθώς ο τελικός κριτής της ποιότητας αναζήτησης είναι ο χρήστης. Όμως τα **ερωτήματα** που εισάγει ο χρήστης αποτελούν **περιγραφές της πληροφοριακής του ανάγκης, μη συγκεκριμένες**. Έτσι αποτελεί ιδιαίτερη

πρόκληση το να καθοριστεί ένα έγγραφο συναφές ως προς αυτές καθώς η **υποκειμενική φύση** των ερωτημάτων και η **δυναμική τους ιδιότητα** αποτελούν τα μεγαλύτερα εμπόδια.

Έτσι η μέθοδος που ακολουθείται είναι η **διαρκής αξιολόγηση των αποτελεσμάτων της ανάκτησης, ώστε τα συστήματα να βελτιώνονται όλο και περισσότερο**. Παρακάτω φαίνονται σχηματικά (βλέπε εικ. 6) τα κύρια θέματα που αφορούν στην αξιολόγηση:



Εικ. 6 Αξιολόγηση ΑΠ σύμφωνα με ταξινόμηση ACM

Για να πραγματοποιηθεί αξιολόγηση ενός ΣΑΠ θα πρέπει να υπάρχουν κάποιες μετρήσιμες ιδιότητες του, ως προς τις οποίες θα εξεταστεί. Σύμφωνα με Ceri et al. (2013) πρέπει να είναι γνωστή η **ύπαρξη μιας συλλογή D και ενός συνόλου ερωτημάτων Q ως προς τα οποία να μπορεί να γίνει αναφορά και τέλος να υπάρχει μια πλειάδα $t_{jk} = (d_j, q_k, r^*)$ οπου για κάθε $q_k \in Q$ και $d_j \in D$ να υπάρχει μια αντίστοιχη δυαδική κρίση (binary judgment) r^* της μεταξύ τους συνάφειας (του d_j σε σχέση με το q_k** δηλαδή), όπως εκτιμήθηκε από μια αρχή αναφοράς.

2.3.1 Δοκιμαστικές συλλογές

Προκειμένου να είναι εφικτή η σύγκριση και αξιολόγηση ΣΑΠ αναπτύχθηκαν διάφορες τέτοιες συλλογές εγγράφων (standard test collections), όπως αναφέρθηκαν παραπάνω, οι οποίες χρησιμοποιούνται ήδη από τη δεκαετία του 1960. Ορισμένα ευρέως χρησιμοποιούμενα παραδείγματα αποτελούν η συλλογή Cranfield, TREC, Reuters και TRC2. Οι συλλογές αυτές περιλαμβάνουν, όπως αναφέρθηκε παραπάνω, σύνολα εγγράφων, ερωτημάτων και σύνολα κριτικών συνάφειας. Ο λόγος ύπαρξης των παραπάνω συλλογών είναι ο υπολογισμός της εκτίμησης συνάφειας με τις πληροφοριακές ανάγκες χρηστών, με σκοπό την αξιολόγηση.

2.3.2 Εκτίμηση συνάφειας

Είναι βασική προϋπόθεση φυσικά πως η πληροφοριακή ανάγκη προς εκτίμηση της συνάφειας της είναι σχετική με τη δοκιμαστική συλλογή (τα έγγραφα που περιέχει είναι σχετικά δηλαδή για αναζήτηση και ανάκτηση). Μάλιστα σύμφωνα με τους Manning, Raghavan και Schutze (2008) οι πληροφοριακές αυτές ανάγκες σχεδιάζονται από ειδικούς του τομέα. Η συγκέντρωση εκτιμήσεων συνάφειας έγκειται σε κάποιο ανθρώπινο δυναμικό το οποίο θα δαπανήσει ιδιαίτερο χρόνο ώστε να εφαρμόσει όλα τα ερωτήματα στη συλλογή, μια ιδιαίτερα δαπανηρή διαδικασία. Όπως αναφέρουν οι Manning, Raghavan και Schutze (2008) τέτοιες διαδικασίες είναι εφικτές για μικρές συλλογές ενω για μεγάλες συλλογές η εκτίμηση συνάφειας συνηθίζεται να εφαρμόζεται σε περιορισμένο υποσύνολο εγγράφων για κάθε ερώτημα και όχι για όλη τη συλλογή.

2.3.3 Αποτελεσματικότητα και απόδοση ανάκτησης

Στην ενότητα αυτή θα γίνει αναφορά σε δύο βασικές έννοιες που αφορούν την αξιολόγηση των ΣΑΠ: την αποτελεσματικότητα και την ανάκληση. Όπως λοιπόν αναφέρεται στη βιβλιογραφία (Croft et al. 2009) η **αποτελεσματικότητα ορίζεται** ως η **ανάκτηση των περισσότερο συναφών έγγραφων** σε σχέση με το ερώτημα χρήστη ενώ η **απόδοση** ως η όσο το δυνατόν **ταχύτερη επεξεργασία των ερωτημάτων** του χρήστη.

Σύμφωνα με τους Manning et al. (2008) για την αξιολόγηση της **αποτελεσματικότητας** ενός ΣΑΠ πρέπει να υπολογιστούν η ακρίβεια και η ανάκληση, μέτρα τα οποία χρησιμοποίησε πρώτος την δεκαετία του 1960 ο Cyril Cleverdon, πρωτοπόρος των ΣΑΠ. Στα μέτρα ανάκλησης και ακρίβειας έγινε σύντομη αναφορά σε παραπάνω ενότητα (βλέπε ενότητα 2.2.2). Όπως αναφέρεται στη βιβλιογραφία (Kowalski 2011) μέσω της ανάκλησης και της ακρίβειας αξιολογείται η ποιότητα της αποτελεσματικότητας της αναζήτησης. Παρακάτω αναλύονται οι μετρικές συνάφειας, **ακρίβεια** και **ανάκληση,** σύμφωνα με τον Ingwersen (1992) οι οποίες υπολογίζονται με τους παρακάτω τύπους και έχουν μια **αντιστρόφως ανάλογη σχέση μεταξύ τους**:

- Ακρίβεια: αριθμός συναφών ανακτηθέντων εγγράφων R προς τον αριθμό των ανακτηθέντων εγγράφων L.

- Ανάκληση: αριθμός συναφών εγγράφων που έχουν ανακτηθεί R προς τον συνολικό αριθμό συναφών εγγράφων στην συλλογή C.

Σύμφωνα με Ingwersen (1992) υπάρχει ακόμα ένα μέτρο, το οποίο μπορεί σε κάποιες περιπτώσεις να αντικαταστήσει την ακρίβεια και ονομάζεται αστοχία. Η αστοχία αντιπροσωπεύει την σχέση μεταξύ του αριθμού μη συναφών ανακτηθέντων εγγράφων και όλων των μη συναφών εγγράφων (όπου N ο συνολικός αριθμός συλλογής).

Σχετικά με την απόδοση των ΣΑΠ, όπως αναφέρεται στην βιβλιογραφία (Kowalski 2011), το αιτούμενο είναι ο καθορισμός **του χρόνος απόκρισης**, δηλαδή του **χρονικού διαστήματος στο οποίο εκτελείται η διαδικασία της αναζήτησης.** Η απόδοση του συστήματος είναι ιδιαιτέρως σημαντική καθώς όπως αναφέρεται στη βιβλιογραφία (Buettcher, Clarke και Cormack 2010) επηρεάζει το κόστος λειτουργίας του συστήματος αλλά και την ικανοποίηση των χρηστών. Σύμφωνα με Kowalski (2011) κάποιες βασικές έννοιες στον χρόνο απόκρισης είναι η αρχή και το τέλος του χρονικού διαστήματος αυτού, δηλαδή το πότε αρχίζει η αναζήτηση (με την εντολή έναρξης του χρήστη) και το πότε τελειώνει (ολοκλήρωση της αναζήτησης). Ακόμη ο Kowalski (2011) συμπληρώνει πως η ολοκλήρωση της αναζήτησης μπορεί να είναι ένα ζήτημα με διαφορετική προσέγγιση για τον χρήστη σε σχέση με το ΣΑΠ. Για το ΣΑΠ η αναζήτηση ολοκληρώνεται μονάχα όταν έχουν βρεθεί όλα τα αποτελέσματα, ενω κατά τον χρήστη μια αναζήτηση έχει ολοκληρωθεί ακόμη και αν έχει επιστραφεί έστω ένα αποτέλεσμα. Οι Croft et al. (2009) από την άλλη τονίζουν ότι η απόδοση εξαρτάται από τα ευρετήρια.

Ένα άλλο ζήτημα που επηρεάζει την απόδοση μιας μηχανής αναζήτησης είναι αυτό της **εξυπηρέτησης πληθώρας χρηστών από τη μηχανή ταυτόχρονα**. Όπως αναφέρεται στη βιβλιογραφία (Buettcher, Clarke και Cormack 2010) μια μηχανή αναζήτησης μπορεί να πρέπει να εξυπηρετήσει ταυτόχρονα χιλιάδες ερωτήματα για μεμονωμένους χρήστες και για να θεωρείται αποδοτική ώστε να μη χάσει τους χρήστες της, ο χρόνος απόκρισης που εξυπηρετεί τον καθένα ξεχωριστά θα πρέπει να μην επηρεάζεται σε τέτοιο βαθμό ώστε να τον αφήσει δυσαρεστημένο. Ο μέσος αριθμός ερωτημάτων που επεξεργάζεται η μηχανή σε συγκεκριμένο χρονικό διάστημα ονομάζεται **ρυθμός διαμεταγωγής (throughput).**

2.3.3.1 Ευαισθησία – Ειδικότητα

Οι όροι ευαισθησία και ειδικότητα εμφανίζονται κυρίως για την **αξιολόγηση διαγνωστικών τεστ** στην ιατρική αλλά και σε πιο εξειδικευμένες περιπτώσεις αξιολόγησης σε διάφορους επιστημονικούς κλάδους, όπως π.χ. στην **αξιολόγηση και εκπαίδευση νευρωνικών δικτύων** (Poulos et al. 2007), στην **αξιολόγηση απόδοσης ταξινομητών σε σχέση με δυαδική ταξινόμηση** (Powers 2011, Fawcett 2006) και γενικότερα στην **αξιολόγηση αλγορίθμων και μέτρηση απόδοσης μοντέλων**. Σύμφωνα με τους Manning, Raghavan και Schutze (2008) οι μετρικές τις ευαισθησίας και της ειδικότητας **στο πλαίσιο της ΑΠ συμπίπτουν με την ανάκληση και την ακρίβεια** αντίστοιχα και συχνά στο πλαίσιο της αξιολόγησης **αναπαρίστανται με γραφικό τρόπο μέσω μιας Receiver Operating Characteristics curve ή αλλιώς ROC curve**.

Ειδικότερα για τον έλεγχο υποθέσεων στα διαγνωστικά τεστ, οι Wong και Gek (2011) τονίζουν πως τα διαγνωστικά – κλινικά τεστ, τα οποία καθορίζουν την ύπαρξη ή μη μιας ασθένειας στα αντικείμενα της μελέτης, στον πληθυσμό δηλαδή που έχει υποβληθεί στο τεστ, **δεν είναι πάντοτε έγκυρα** και υπάρχει η πιθανότητα κάποιου ποσοστού του αποτελέσματος να είναι λάθος. Γι' αυτό τον λόγο, οι Wong και Gek (2011) τονίζουν πως πρέπει να γίνεται **αξιολόγηση των τεστ** αυτών (test validation) και αυτό επιτυγχάνεται μέσω της μέτρησης 4 παραμέτρων: **ευαισθησία, ειδικότητα, θετική προγνωστική αξία** (positive predictive value, PPV) και **αρνητική προγνωστική αξία** (NPV). Συνεπώς χρησιμοποιώντας τις παραπάνω παραμέτρους μπορεί να γίνει **έλεγχος της εγκυρότητας ή μη μιας υπόθεσης**, όχι μόνο στην ιατρική και τα διαγνωστικά τεστ αλλά και γενικότερα για στατιστικούς λόγους.

Οι Lalkhen και McCluskey (2008), συμπληρώνουν τους παρακάτω ορισμούς προκειμένου να κατανοηθεί στη συνέχεια η ανάλυση των όρων ευαισθησίας και ειδικότητας. Οι πιθανές τιμές που μπορεί να έχει ως αποτέλεσμα ένα διαγνωστικό τεστ είναι οι ακόλουθες:

- Αληθινό θετικό: όταν ο ασθενής έχει την ασθένεια και το τεστ βγαίνει θετικό.
- Ψεύτικο θετικό: όταν ο ασθενής δεν έχει την ασθένεια και το τεστ είναι θετικό
- Αληθινό αρνητικό: όταν ο ασθενής δεν έχει την ασθένεια και το τεστ είναι αρνητικό
- Ψεύτικο αρνητικό: όταν ο ασθενής έχει την ασθένεια αλλά το τεστ είναι αρνητικό.

Σύμφωνα με τον Akobeng (2007) αρκετά χρήσιμο στην κατανόηση των όρων ευαισθησία – ειδικότητα είναι η τοποθέτηση των παραπάνω πιθανών αποτελεσμάτων, όπως φαίνονται στον πίνακα 1:

Πίνακας 1. Πιθανά αποτελέσματα για υπολογισμό ευαισθησίας - ειδικότητας

	ασθενής	μη ασθενής
Θετικό	A	B
αρνητικό	C	D

Ο Akobeng (2007) στο άρθρο του αναφέρει τον υπολογισμό των όρων μέσω των (9), (10), (11) και (12):

$$\text{Ευαισθησία} = \frac{a}{a+c} \tag{9}$$

$$\text{Ειδικότητα} = \frac{b}{b+d} \tag{10}$$

$$PPV = \frac{a}{a+b} \tag{11}$$

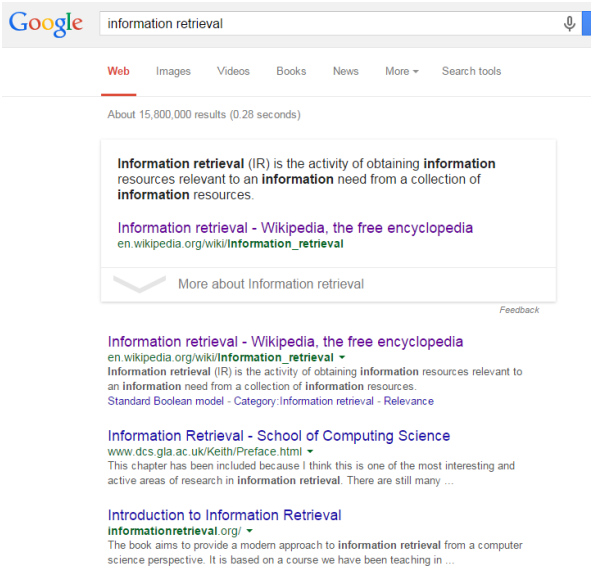
$$NPV = \frac{d}{c+d} \tag{12}$$

Οι Zhu, Zeng και Wang (2010) αναφέρουν ότι η ευαισθησία αφορά το πόσο καλό είναι το τεστ στην ανίχνευση των πραγματικών ασθενών (δηλαδή θετική απάντηση στο αποτέλεσμα του τεστ) ενώ η ειδικότητα αφορά την ανίχνευση των μη πασχόντων (δηλαδή αρνητική απάντηση στο αποτέλεσμα του τεστ).

2.3.4 Παρουσίαση ανακτηθέντων αποτελεσμάτων

Σύμφωνα με τη βιβλιογραφία (Wu, Bi και Zeng 2010, Kamps 2009), η παρουσίαση ανακτηθέντων αποτελεσμάτων συνδέεται άμεσα με την αξιολόγηση και **ο τρόπος παρουσίασης των ανακτηθέντων αποτελεσμάτων στους χρήστες μπορεί να καθορίσει το αν το ΣΑΠ θα θεωρηθεί επιτυχημένο** ή όχι. Ειδικότερα στην βιβλιογραφία αναφέρεται (Joho και Jose 2006) πως έχει αποδειχτεί ότι οι χρήστες όταν ψάχνουν για μια πληροφορία στο διαδίκτυο μέσω μηχανών αναζήτησης εξετάζουν όσο το δυνατόν μικρότερο αριθμό εγγράφων και συνήθως **περιορίζονται στην πρώτη σελίδα αποτελεσμάτων** από την λίστα. Έτσι **έγκειται στον τρόπο**

παρουσίασης των αποτελεσμάτων, στο σύνολο τους αλλά και σε κάθε μεμονωμένο αποτέλεσμα, να δελεάσουν και **να προτρέψουν τον χρήστη να επιλέξει κάποια από αυτά** τα αποτελέσματα **για περιήγηση**. Οι πληροφορίες που παρουσιάζονται συνήθως στα αποτελέσματα είναι **τίτλος, περίληψη, διεύθυνση url, αποσπάσματα εγγράφου.** Μέσω των στοιχείων που θα εμφανιστούν στο χρήστη μόλις δει τα αποτελέσματα για πρώτη φορά θα ληφθεί η απόφαση από τον ίδιο για τη συνάφεια τους και για τον αν θα ενδιαφερθεί να διαβάσει περαιτέρω κάποιο έγγραφο ώστε να βρει την πληροφορία που χρειάζεται. Τέλος από το **αν θα είναι ευχαριστημένος** από την όλη αυτή διαδικασία επιλογής και προβολής αποτελεσμάτων θα γίνει αξιολόγηση του συστήματος και θα προτιμάται ή όχι από τους χρήστες. Στην εικ. 7 φαίνεται ο τρόπος παρουσίασης αποτελεσμάτων για ένα τυχαίο παράδειγμα στο διαδίκτυο:



Εικ. 7 Παρουσίαση ανακτηθέντων αποτελεσμάτων από μηχανή αναζήτησης Google (ανακτήθηκε στις 25/02/2015 από https://www.google.gr/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=information+retrieval)

Κατά την κατάταξη των εγγράφων για την παρουσίαση των αποτελεσμάτων, σε κάθε έγγραφο ανατίθεται ένας βαθμός συνάφειας (Wu, Bi και Zeng 2010) μέσα στη συλλογή. Από τις βαθμολογίες αυτές γίνεται η κατάταξη τους με πιο συνηθισμένη μορφή παρουσίασης ανακτηθέντων αποτελεσμάτων αυτή της **λίστας καταταγμένων εγγράφων**. Όσα βρίσκονται ψηλότερα στην κατάταξη είναι πιο συναφή ενώ προχωρώντας προς τα κάτω στη λίστα ακολουθείται μια φθίνουσα πορεία σε σχέση με τη συνάφεια της πληροφοριακής ανάγκης του χρήστη. Ενναλακτική παρουσίαση είναι αυτή που χρησιμοποιεί **ιεραρχικές δομές** (βλέπε εικ.

8). Στη βιβλιογραφία αναφέρεται (Crestani και Wu 2006) πως σύμφωνα με πληθώρα ερευνών και πειραμάτων, η παρουσίαση ανακτηθέντων αποτελεσμάτων με συσταδοποίηση, δηλαδή ομαδοποίηση αποτελεσμάτων χωρίς να υπάρχουν προκαθορισμένες κλάσεις είναι αποτελεσματικότερη από μια απλή λίστα κατάταξης, καθώς βασίζεται στην υπόθεση ότι τα έγγραφα που ομαδοποιούνται μαζί έχουν παρόμοια συνάφεια για συγκεκριμένο ερώτημα.

Query: israel
Documents: 272, Clusters: 15, Average Cluster Size: 15.1 documents

Cluster	Size	Shared Phrases and Sample Document Titles
1 View Results Refine Query Based On This Cluster	16	Society and Culture (56%), Faiths and Practices (56%), Judaism (69%), Spirituality (56%); Religion (56%) , organizations (43%) Ahavat Israel - The Amazing Jewish Website! Israel and Judaism Judaica Collection
2 View Results Refine Query Based On This Cluster	15	Ministry of Foreign Affairs (33%), Ministry (87%) Publications and Data of the BANK OF ISRAEL Consulate General of Israel to the Mid-Atlantic Region The Friends of Israel Gospel Ministry
3 View Results Refine Query Based On This Cluster	11	Israel Tourism (36%), Comprehensive Israel (36%), Tourism (64%) Interactive Israel tourism guide - Jerusalem Ambassade d'Israel Travel to Israel Opportunites
4 View Results Refine Query Based On This Cluster	7	Middle East (57%), History (57%); WAR (42%) , Region (42%) , Complete (42%) , Listing (42%) , country (42%) Israel at Fifty: Our Introduction to The Six Day War Machal - Volunteers in the Israel's War of Independence HISTORY: The State of Israel
5 View Results Refine Query Based On This Cluster	22	Economy (68%), Companies (55%), Travel (55%) Israel Hotel Association Israel Association of Electronics Industries Focus Capital Group - Israel

Εικ. 8 Παρουσίαση αποτελεσμάτων βασισμένη σε συσταδοποίηση από Zamir O., Etzioni O. (1999)

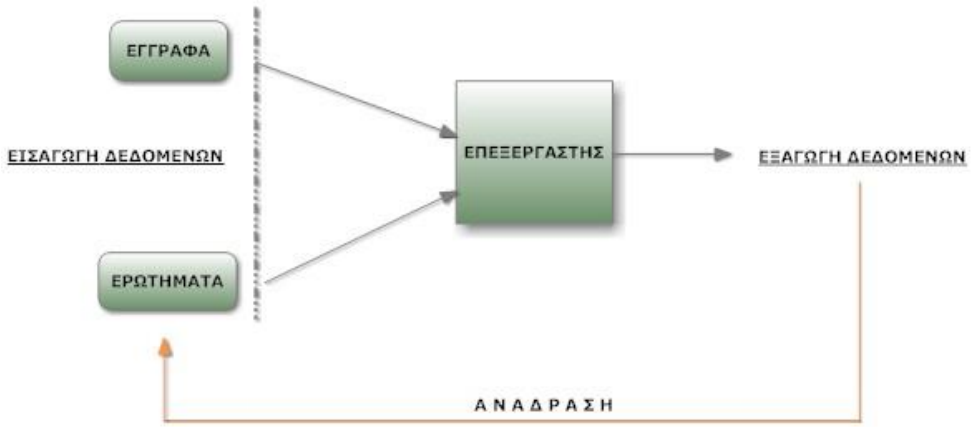
Όπως αναφέρουν οι Wu, Bi και Zeng (2010), η σχέση μεταξύ των βαθμολογιών εγγράφων για την κατάταξη αποτελεί τις πιθανότητες συνάφειας, οι οποίες μπορεί να είναι ιδιαίτερα χρήσιμες σε προχωρημένα ΣΑΠ.

2.4 Αρχιτεκτονική ΣΑΠ

Μέχρι αυτό το σημείο της διατριβής έχει γίνει αναφορά τόσο στη γενικευμένη έννοια των ΣΑΠ αλλά και στις μηχανές αναζήτησης που αποτελούν την πιο ευρέως χρησιμοποιούμενη εφαρμογή τους. Στη βιβλιογραφία (Canfora και Cerulo 2004), **οι μηχανές αναζήτησης θεωρούνται συνώνυμες με τα ΣΑΠ στο περιβάλλον του διαδικτύου**. Κατά συνέπεια, όπως είναι αναμενόμενο και σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) οι **βασικοί στόχοι** που πρέπει να επιτευχθούν για μια **μηχανή αναζήτησης** ως προς τον σχεδιασμό και την αρχιτεκτονική της, είναι ταυτόσημοι με αυτούς των ΣΑΠ (**αποτελεσματικότητα και απόδοση)**. Καθώς η αρχιτεκτονική μιας μηχανής αναζήτησης είναι ιδιαιτέρως πολύπλοκη, αρχικά θα γίνει παρουσίαση μιας απλουστευμένης μορφής ενός ΣΑΠ. Έτσι σύμφωνα με τον Van Rijsbergen (1979) ένα ΣΑΠ παρουσιάζεται στην απλούστερη και πιο γενικευμένη μορφή του ως ένα σύστημα αποτελούμενο από **τρία βασικά μέρη**: εισαγωγή

δεδομένων (input), επεξεργαστή και εξαγωγή δεδομένων (output), όπως φαίνεται σχηματικά στην εικ. 9:

- **Εισαγωγή δεδομένων:** το ΣΑΠ δέχεται ως δεδομένα τα έγγραφα από την μια πλευρά και από την άλλη τα ερωτήματα του χρήστη και τα δύο πρέπει να αναπαρίστανται με κατάλληλο τρόπο για τον υπολογιστή. Καθώς ο χρήστης αλλάζει τις απαιτήσεις του κατά την διάρκεια μιας ενότητας διαδικτυακής αναζήτησης, μπορεί να βελτιωθεί η αναζήτηση μέσω της διαδικασίας ανάδρασης.
- **Επεξεργαστής:** ασχολείται με τη διαδικασία ανάκτησης. Ορισμένες από τις διεργασίες που περιλαμβάνει είναι η δόμηση της πληροφορίας με κατάλληλο τρόπο και η διαδικασία αυτή καθαυτή της ανάκτησης του συνόλου εγγράφων.
- **Εξαγωγή δεδομένων:** περιλαμβάνει τα αποτελέσματα της διαδικασίας ανάκτησης.

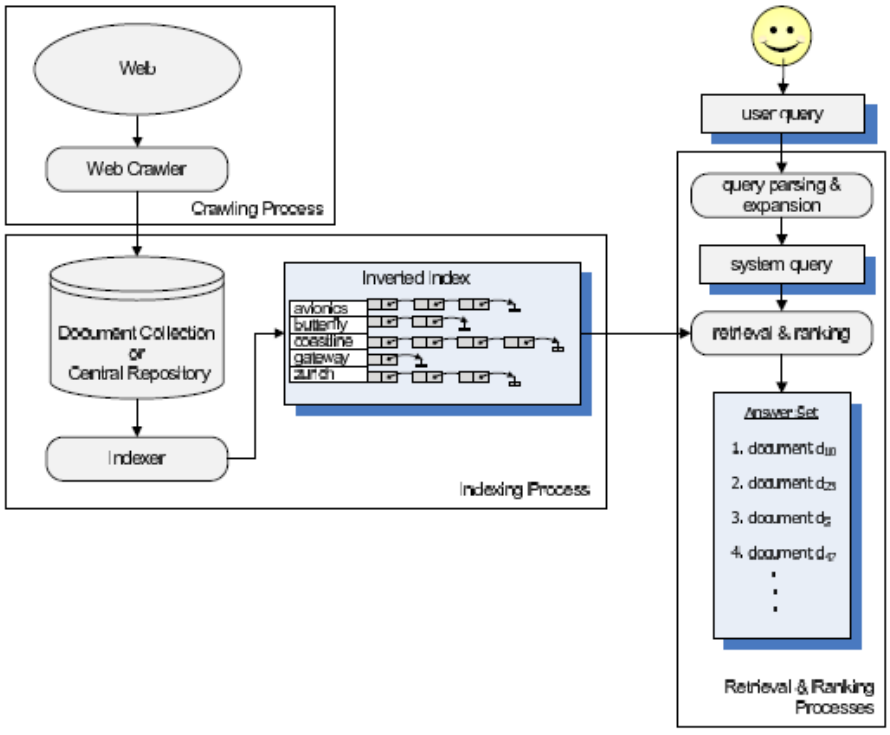


Εικ. 9 Αναπαράσταση αρχιτεκτονικής ΣΑΠ

Εμβαθύνοντας περαιτέρω στην **αρχιτεκτονική ενός ΣΑΠ**, μια αρκετά **πιο λεπτομερής** παρουσίαση (βλέπε εικ. 10) προέρχεται από τους Baeza-Yates και Ribeiro-Neto (2011), στην οποία ένα ΣΑΠ αποτελείται από δύο μεγάλες ενότητες διαδικασιών. Η πρώτη ενότητα αφορά τη **(α) συλλογή των εγγράφων** και όλες τις διαδικασίες που την αφορούν και η δεύτερη την **(β) ανάκτηση των εγγράφων** και όλων των διαδικασιών που περιλαμβάνονται για την υλοποίηση της. Αναλυτικότερα:

α. **Συλλογή εγγράφων:** είναι αποθηκευμένη σε ένα κεντρικό αποθετήριο και πρέπει να ευρετηριαστεί, δηλαδή να δημιουργηθεί ένα ευρετήριο από τα έγγραφα που την απαρτίζουν. Όλα τα βήματα για τη δημιουργία του ευρετηρίου αυτού αποτελούν τη διαδικασία **ευρετηρίασης**, η οποία πραγματοποιείται εκτός διαδικτύου.

β. **Ανάκτηση εγγράφων**: περιλαμβάνει όλα τα βήματα που ακολουθούνται ώστε να παραχθεί ένα σύνολο ανακτηθέντων εγγράφων, τα οποία θα είναι συναφή ως προς το ερώτημα χρήστη. Η διαδικασία ανάκτησης εγγράφων αφορά την αναζήτηση του χρήστη για πληροφορία με βάση την πληροφοριακή του ανάγκη ή σε περιήγηση πληροφορίας όταν ο χρήστης περιηγείται μέσω υπερσυνδέσμων. **Εστιάζοντας στην αναζήτηση πληροφοριών το ερώτημα του χρήστη αναλύεται, επεκτείνεται και επεξεργάζεται προκειμένου να επιστραφεί ένα υποσύνολο εγγράφων από το ευρετήριο**. Πιο συγκεκριμένα, το ερώτημα αναλύεται και τροποποιείται εκτελώντας κάποιες βασικές διεργασίες, όπως διόρθωση ορθογραφικών λαθών, αποκλεισμός όρων ως **διακόπτουσες λέξεις** (λέξεις που κατέχουν λειτουργικό ρόλο στο σχηματισμό δομής της πρότασης ενω ταυτόχρονα δεν συνεισφέρουν σημασιολογικά στο έγγραφο) και τέλος τροποποιείται ξανά μέσω προτάσεων από το σύστημα. Το τελικό τροποποιημένο ερώτημα επεξεργάζεται και σύμφωνα με αυτό επιστρέφεται ένα σύνολο ανακτηθέντων εγγράφων με έγγραφα που περιέχουν τους όρους του ερωτήματος χρήστη. Τα ανακτηθέντα έγγραφα κατατάσσονται με βάση τη συνάφεια σε σχέση με το ερώτημα του χρήστη και επιστρέφονται αυτά που κατά την αξιολόγηση κρίνονται πιο συναφή.



Εικ. 10 Αναπαράσταση αρχιτεκτονικής ΣΑΠ από Baeza-Yates και Ribeiro-Neto (2011)

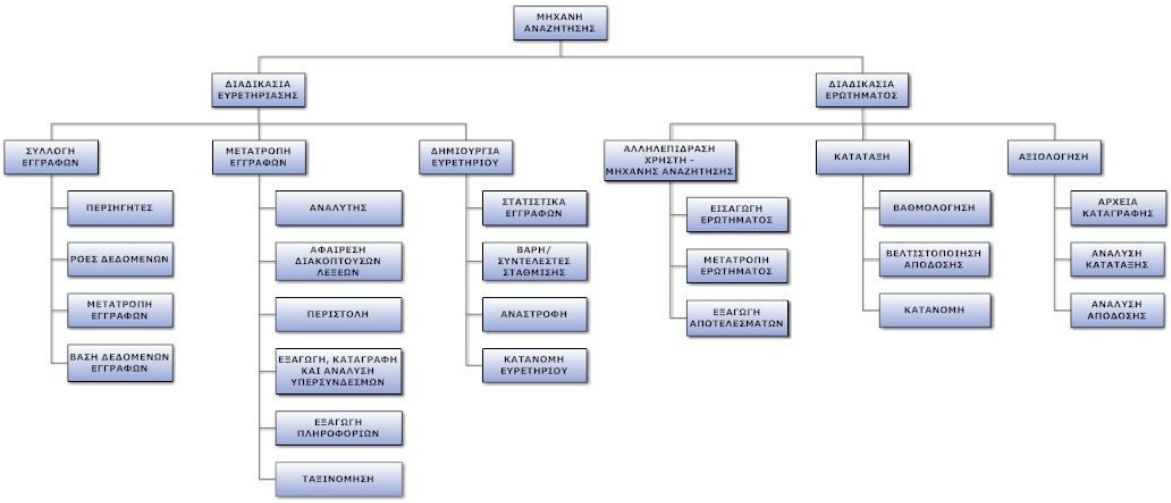
Όπως ήδη έχει αναφερθεί στην αρχή της ενότητας αυτής, οι μηχανές αναζήτησης αποτελούν την πιο διαδεδομένη εφαρμογή για την υλοποίηση τεχνικών ΑΠ και θεωρούνται σχεδόν ταυτόσημες με την έννοια της ΑΠ. Μάλιστα στη βιβλιογραφία (Croft et al. 2010) αναφέρεται ο παρακάτω ορισμός της μηχανής αναζήτησης: «***μια μηχανή αναζήτησης αποτελεί την πρακτική εφαρμογή τεχνικών ανάκτησης πληροφοριών σε μεγάλης κλίμακας συλλογές κειμένων***». Κατά συνέπεια τα βασικά ζητήματα ΑΠ αποτελούν βασικά ζητήματα και για τη σχεδίαση μηχανών αναζήτησης.

Φυσικά η ΑΠ εφαρμόζεται σε πληθώρα άλλων πεδίων εκτός των μηχανών αναζήτησης. Ορισμένα παραδείγματα χρήσης της στη βιβλιογραφία (Ceri et al. 2013) αφορούν συστήματα φιλτραρίσματος πληροφορίας, εξαγωγή περιλήψεων εγγράφων, συσταδοποίηση και ταξινόμηση εγγράφων, συστήματα απάντησης ερωτήσεων για την διαγλωσσική ανάκτηση.

Σε αυτό το σημείο κατόπιν ολοκλήρωσης της παρουσίασης της αρχιτεκτονικής ενός απλοποιημένου ΣΑΠ, στην ενότητα που ακολουθεί θα αναλυθεί η αρχιτεκτονική των μηχανών αναζήτησης ακόμη εκτενέστερα. Έτσι θα γίνει περισσότερο κατανοητός όχι μόνον ο τρόπος λειτουργίας τους αλλά και ο ρόλος των μοντέλων ΑΠ και ειδικότερα του αλγόριθμου κατάταξης.

2.4.1 Αρχιτεκτονική μηχανής αναζήτησης

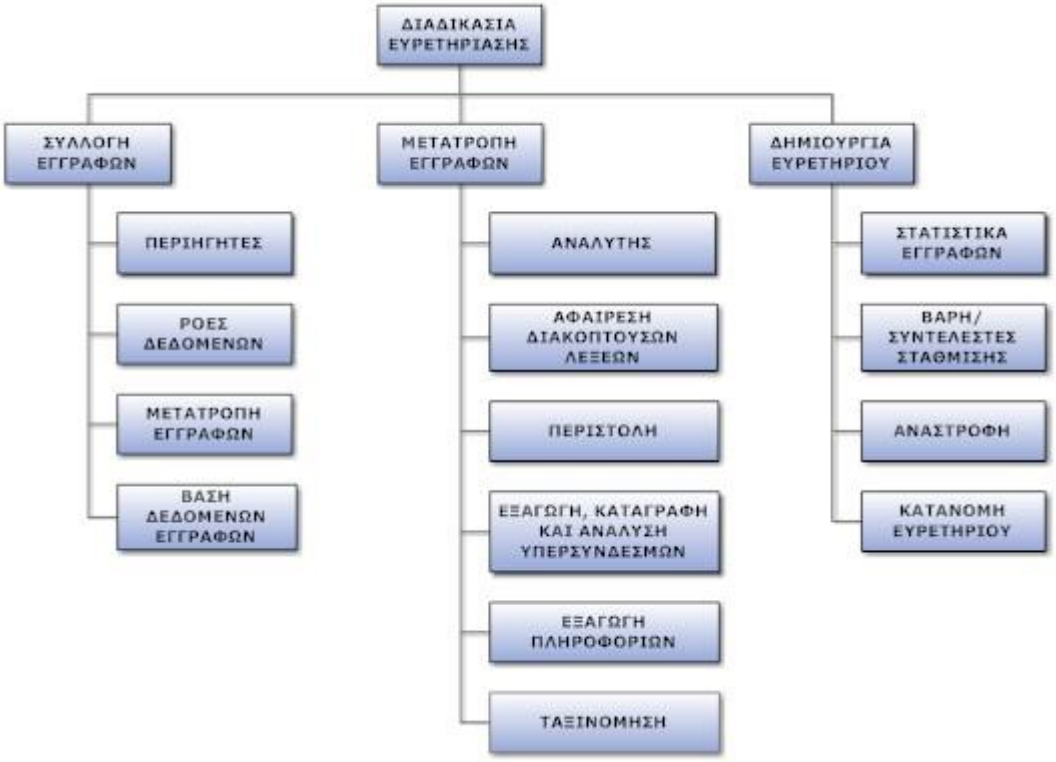
Η αρχιτεκτονική των μηχανών αναζήτησης αναλύεται σε βάθος από τους Croft B. et al. (2009), όπου παρουσιάζονται με ιδιαίτερη λεπτομέρεια όλα τα λειτουργικά τμήματα που αλληλεπιδρούν και κατηγοριοποιούνται σε **δύο κεντρικές λειτουργίες: τη διαδικασία ευρετηρίασης (indexing process) και τη διαδικασία ερωτήματος (query process)**. Σε αυτές έχει ήδη γίνει γενική αναφορά στην προηγούμενη ενότητα, εδώ θα παρουσιαστούν πολύ συγκεκριμένα. Σχηματικά η αρχιτεκτονική μιας μηχανής αναζήτησης (βλέπε εικ. 11) συνοψίζεται παρακάτω:



Εικ. 11 Αρχιτεκτονική μηχανής αναζήτησης

2.4.1.1 Αρχιτεκτονική μηχανών αναζήτησης - ευρετηρίαση

Ξεκινώντας με τη **διαδικασία ευρετηρίασης** στη βιβλιογραφία (Croft et al. 2009) εντοπίζονται τρεις επιμέρους βασικές διαδικασίες: **συλλογή εγγράφων** (text acquisition), **μετατροπή εγγράφων** (text transformation) και **δημιουργία ευρετηρίου** (index creation), όπως φαίνεται σχηματικά στην εικ. 12:



Εικ. 12 Διαδικασία ευρετηρίασης

Συλλογή εγγράφων: κατά τη διαδικασία αυτή αναγνωρίζονται και συλλέγονται τα έγγραφα της συλλογής (η οποία μπορεί να είναι ήδη αποθηκευμένη ή να δημιουργείται μέσω web crawling) και πραγματοποιούνται διαδικασίες ώστε τα έγγραφα αυτά να μπορούν να είναι προσβάσιμα κατά την αναζήτηση. Ειδικότερα περιλαμβάνει:

1. Τους **περιηγητές** (**crawlers)**, προγράμματα τα οποία αναγνωρίζουν και επιστρέφουν έγγραφα στη μηχανή αναζήτησης, όπως π.χ. η κατηγορία general web crawler. Το πρόγραμμα αυτό ακολουθεί τους υπερσυνδέσμους των ιστοσελίδων, εντοπίζει ιστοσελίδες και τις αποθηκεύει.
2. Τις **ροές δεδομένων** (**feeds),** μέσω των οποίων είναι δυνατή η πρόσβαση σε πραγματικό χρόνο ακολουθίας εγγράφων μέσω εγγραφής των χρηστών σε ιστοτόπους. Παράδειγμα αποτελεί το πρότυπο RSS.
3. Τη **μετατροπή (conversion)** των εγγράφων, σε κατάλληλο μορφότυπο αρχείου κειμένου, καθώς συνήθως ανακτώνται σε αρχεία xml, html, pdf κ.α.
4. Τη **βάση δεδομένων εγγράφων (document data store),** η οποία διαχειρίζεται τον όγκο των εγγράφων και των δομημένων δεδομένων τους όπως μεταδεδομένα ή άλλες πληροφορίες που προέρχονται από το έγγραφο (π.χ. υπερσύνδεσμοι).

Βασικό σε αυτό το σημείο είναι να οριστούν οι παραλλαγές ακολουθίας χαρακτήρων των εγγράφων και οι διαφορές τους. Μια **ακολουθία χαρακτήρων** σύμφωνα με τους Manning, Raghavan και Schutze (2008) μπορεί να διαιρεθεί σε **σύμβολα (tokens), τύπους (types) ή όρους (terms)**. **Σύμβολο** ονομάζεται μια ακολουθία χαρακτήρων που παρουσιάζεται σε συγκεκριμένο έγγραφο και αποτελεί μια σημασιολογική μονάδα. **Τύπος** ονομάζεται μια κλάση που αποτελείται από σύμβολα με ίδια ακολουθία χαρακτήρων. Τέλος, **όρος** ονομάζεται ένας τύπος ο οποίος εντάσσεται στο λεξικό του ΣΑΠ. Οι τύπου και οι όροι είναι συνήθως **κανονικοποιημένοι**, κάτι που θα που ήδη έχει παρουσιαστεί σε προηγούμενη ενότητα (βλέπε ενότητα 2.2.2). Θα παρουσιαστούν οι διαφορές των παραπάνω σε **ένα σύντομο παράδειγμα**:

- Έστω ότι υπάρχει μια πρόταση με 7 λέξεις από τις οποίες οι 2 θεωρούνται διακόπτουσες και πως οι 2 διακόπτουσες αυτές λέξεις είναι η ίδια λέξη "*of*" (δηλαδή η λέξη "*of*" επαναλαμβάνεται δύο φορές). Τότε:
    - Οι 7 λέξεις θεωρούνται σύμβολα, δηλαδή όσες είναι οι μονάδες της πρότασης τόσα και τα σύμβολα,

- Οι διαφορετικοί τύποι που εμφανίζονται είναι 6, καθώς η μια λέξη "*of*" εμφανίζεται 2 φορές
- Οι όροι που θα ενταχθούν στο ευρετήριο τελικά είναι 5, καθώς οι διακόπτουσες λέξεις αφαιρούνται.

Μετατροπή εγγράφων: συνεχίζοντας την παρουσίαση σύμφωνα με τη βιβλιογραφία (Croft et al. 2009), η διαδικασία αυτή περιλαμβάνει την **μετατροπή εγγράφων σε όρους ευρετηρίου** (index terms), με αποτέλεσμα να δημιουργείται ένα λεξιλόγιο όρων (index vocabulary). Ειδικότερα περιλαμβάνει:

1. Τον **αναλυτή (parser), ο οποίος διαιρεί την ακολουθία χαρακτήρων του εγγράφου σε σύμβολα (tokenization)**. Όταν υποβάλλεται το ερώτημα του χρήστη τότε αυτό διαιρείται σε σύμβολα ώστε να συγκριθεί με τα έγγραφα της συλλογής. Στη διαδικασία της διαίρεσης σε σύμβολα σημαντικά ζητήματα που πρέπει να αντιμετωπιστούν είναι οι ειδικοί χαρακτήρες, το ενωτικό και η απόστροφος. Ακόμη ο αναλυτής ασχολείται και με την λήψη της δομής τους εγγράφου από τη γλώσσα σήμανσης στην οποία έχει συνταχθεί (HTML ή XML), χρησιμοποιώντας το συντακτικό της γλώσσας και έτσι αντιλαμβάνεται τα μέρη από τα οποία αποτελείται ένα έγγραφο (τίτλος, επικεφαλίδα κλπ.).
2. Τις **διακόπτουσες λέξεις (stop words)**. Βασική εργασία είναι η **αφαίρεση** των λέξεων αυτών από την ακολουθία συμβόλων του εγγράφου, κατά συνέπεια και από το λεξιλόγιο όρων ώστε να μειωθεί το μέγεθος του ευρετηρίου. Οι διακόπτουσες λέξεις έχουν ήδη οριστεί σε προηγούμενη ενότητα (βλέπε ενότητα 2.4). Μάλιστα όλες οι διακόπτουσες λέξεις συγκεντρώνονται σε μια κοινή λίστα διακοπτουσών λέξεων (stop word list).
3. Την **περιστολή (stemming)** είναι η διαδικασία κατά την οποία ένας αλγόριθμος συγκεντρώνει λέξεις με κοινό τμήμα, τις ομαδοποιεί και τέλος τις αντικαθιστά με μια προκαθορισμένη λέξη προκειμένου να επιτύχει ομοιογένεια και κατά συνέπεια να βελτιωθεί η διαδικασία κατάταξης αποτελεσμάτων.
4. Την **εξαγωγή, καταγραφή και ανάλυση των υπερσυνδέσμων και του κειμένου αγκύρωσης (anchor text/** αποτελεί το κείμενο, το οποίο κάνει κλικ πάνω του ο χρήστης και μεταφέρεται στον υπερσύνδεσμο) των ιστοσελίδων που είναι αποθηκευμένες στη βάση δεδομένων εγγράφων, προκειμένου να χρησιμοποιηθούν από αλγόριθμους ανάλυσης υπερσυνδέσμων (όπως PageRank) με σκοπό να παρέχουν στην μηχανή αναζήτησης είτε μια κατάταξη

του πόσο δημοφιλείς είναι οι ιστοσελίδες αυτές στον ιστό είτε για το **authority** της κάθε ιστοσελίδας, το οποίο αφορά το κύρος των υπερσυνδέσμων που φέρει αυτή.

5. Την **Εξαγωγή Πληροφοριών**, ώστε να αναγνωριστούν όροι ευρετηρίου οι οποίοι έχουν πιο πολύπλοκη μορφή ως λέξεις. Ορισμένα παραδείγματα αποτελούν οι λέξεις που βρίσκονται σε πλάγια ή έντονη γραφή, λέξεις ως επικεφαλίδες κ.α. Ακόμη όσον αφορά την εξαγωγή πληροφοριών σχετικών με τη σύνταξη του κειμένου χρησιμοποιούνται **προγράμματα επισημείωσης μερών του λόγου (part-of-speech taggers)** (βλέπε ενότητα 3.2.2).
6. Την **ανάθεση εγγράφων σε ομάδες**. Αυτό μπορεί να γίνει **είτε μέσω τεχνικών ταξινόμησης με προκαθορισμένες κλάσεις** όπου ο ταξινομητής (classifier) πραγματοποιεί αναγνώριση σύμφωνα με τα μεταδεδομένα εγγράφων ή μερών τους και έτσι τα αναθέτει στις κατάλληλες προκαθορισμένες ετικέτες κλάσεων (θεματικές ή κατάταξη εγγράφων ως κακόβουλα (spam) ή αναγνώριση αντικειμένων που δεν αποτελούν μέρος του εγγράφου όπως διαφημίσεις), **είτε μέσω τεχνικών συσταδοποίησης** (clustering) όπου τα έγγραφα ομαδοποιούνται χωρίς να υπάρχουν προκαθορισμένες κλάσεις. Η ομαδοποίηση γίνεται με τέτοιο τρόπο ώστε να χρησιμεύει για την κατάταξη εγγράφων ή για την αλληλεπίδραση με το χρήστη.

Όπως αναφέρουν οι Baeza-Yates και Ribeiro-Neto (2011), η δημιουργία του ευρετηρίου αφορά κυρίως τη μείωση του χρόνου απόκρισης κατά τη διάρκεια της αναζήτησης. Όπως αναφέρουν η καλύτερη επιλογή ευρετηρίασης είναι τα ανεστραμμένα ευρετήρια αλλά υπάρχουν και άλλες τεχνικές ευρετηρίασης (πίνακες επιθεμάτων, αρχεία υπογραφών). Η διαδικασία ευρετηρίασης ολοκληρώνεται σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) με το τελευταία στάδιο της **δημιουργίας ευρετηρίου**, το οποίο αποτελείται από τα παρακάτω βήματα:

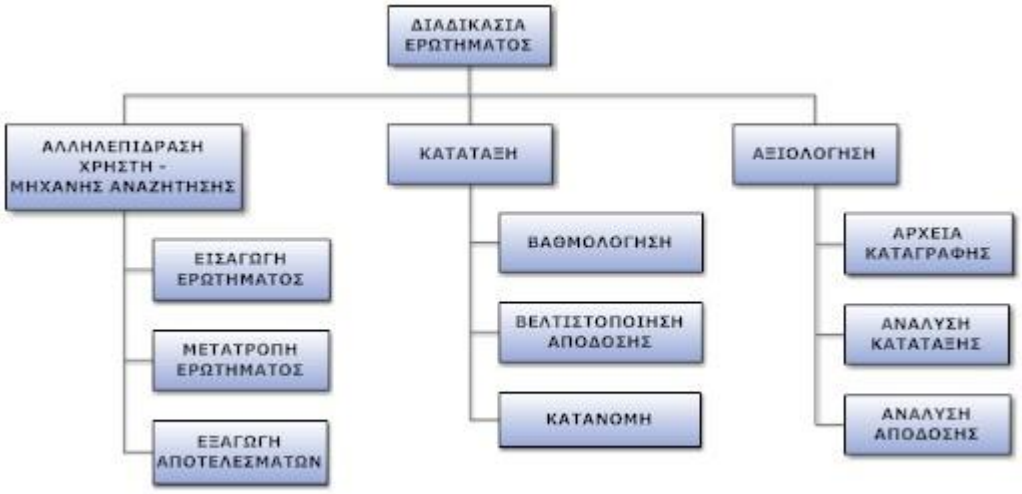
1. **Συγκέντρωση στατιστικών πληροφοριών σχετικές με τις λέξεις, τα έγγραφα και τα χαρακτηριστικά τους**. Παραδείγματα τέτοιων στατιστικών στοιχείων αποτελούν ο αριθμός εμφάνισης όρων ευρετηρίου σε κάθε έγγραφο της συλλογής ή η θέση κάθε όρου του ευρετηρίου στο έγγραφο, τα οποία χρησιμοποιούνται στην κατάταξη και στον υπολογισμό βαθμολογίας (score) εγγράφων, στα πλαίσια μιας αναζήτησης. Το ποιες στατιστικές πληροφορίες θα συλλεχθούν καθορίζεται από το μοντέλο ΑΠ που χρησιμοποιεί η μηχανή

αναζήτησης και ειδικότερα από τον αλγόριθμο κατάταξης που χρησιμοποιεί. Τέλος οι στατιστικές πληροφορίες αποθηκεύονται σε πίνακες αναζήτησης (look-up tables).

2. **Προσδιορισμός συντελεστών στάθμισης/βάρη (weighting)** όρων ευρετηρίου ώστε να διακρίνονται οι όροι μεταξύ τους με βάση τη διαφορετική σημασία που έχουν για κάθε έγγραφο, η οποία μεταβάλλεται από έγγραφο σε έγγραφο και βοηθά στον υπολογισμό βαθμολογίας εγγράφων στην κατάταξη. Οι συντελεστές στάθμισης/βάρη όρων αποθηκεύονται επίσης σε πίνακες αναζήτησης και η μορφή τους καθορίζεται από το μοντέλο ΑΠ. Ένα από τα πιο συνηθισμένα μοντέλα προσδιορισμού συντελεστών στάθμισης είναι το **term frequency - invert document frequency (tf-idf)**, το οποίο ήδη έχει επεξηγηθεί σε προηγούμενη ενότητα (βλέπε ενότητα 2.2).
3. **Αναστροφή (inversion)**, η οποία ουσιαστικά αφορά στην διαδικασία αντιστοίχισης όρων σε έγγραφα, ώστε όταν ζητείται ένας όρος κατά την αναζήτηση, να μπορεί να οδηγήσει στα έγγραφα της συλλογής που τον περιέχουν. Τα λεγόμενα ***αντεστραμμένα ευρετήρια*** βελτιώνουν την απόδοση της μηχανής αναζήτησης καθώς η αναζήτηση πραγματοποιείται πιο γρήγορα. Έτσι από την αποδόμηση και ανάλυση των εγγράφων της συλλογής με όλες τις παραπάνω διαδικασίες, πραγματοποιείται η αναστροφή, η οποία αποτελεί είδος συντομεύσεων από όρους προς έγγραφα της συλλογής.
4. **Κατανομή ευρετηρίων** (index distribution), πρόκειται για τη διανομή του ευρετηρίου σε πολλαπλούς υπολογιστές και σε πολλαπλές ιστοσελίδες στο διαδίκτυο, ώστε τόσο η ευρετηρίαση όσο και η διαδικασία υποβολής ερωτήματος του χρήστη να μπορούν να πραγματοποιηθούν παράλληλα, για τη συλλογή.

2.4.1.2 Αρχιτεκτονική μηχανών αναζήτησης – διαδικασία ερωτήματος

Η δεύτερη μεγάλη διαδικασία είναι αυτή της **διαδικασίας ερωτήματος**. Σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) περιγράφεται ως μια διαδικασία παραγωγής λίστας αποτελεσμάτων τα οποία θεωρούνται συναφή ως προς το ερώτημα του χρήστη (βλέπε εικ. 13). Η διαδικασία αυτή χωρίζεται ξανά **σε τρεις επιμέρους βασικές διαδικασίες**: την **αλληλεπίδραση χρήστη με την μηχανή αναζήτησης (user interaction), την κατάταξη (ranking) και την αξιολόγηση (evaluation)**.



Εικ. 13 Διαδικασία ερωτήματος

Αναλύονται λοιπόν οι τρεις υπό-διαδικασίες σύμφωνα με την βιβλιογραφία (Croft et al. 2009) ως εξής:

Αλληλεπίδραση χρήστη με τη μηχανή αναζήτησης: ουσιαστικά αφορά στο συστατικό της μηχανής που (α) παίρνει το ερώτημα που υποβάλλει ο χρήστης, το οποίο μετατρέπεται σε όρους ευρετηρίου, (β) οργανώνει τη λίστα κατάταξης εγγράφων και εμφανίζει τα αποτελέσματα στο χρήστη. Πιο συγκεκριμένα:

1. **Εισαγωγή δεδομένων ερωτήματος (query input)**: αποτελείται από την **διεπαφή για τον χρήστη και έναν αναλυτή γλώσσας ερωτημάτων (query language parser)**. Οι γλώσσες ερωτημάτων αποτελούνται από τους τελεστές (operators), δηλαδή σύνολο εντολών οι οποίες χρησιμοποιούνται για να υποδείξουν κείμενο το οποίο χρήζει ιδιαίτερης μεταχείρισης, ώστε να αποσαφηνιστεί το τι ψάχνει ο χρήστης μέσω του ερωτήματος που έχει εισάγει. Όπως ήδη έχει αναφερθεί σε προηγούμενες ενότητες το ερώτημα του χρήστη αποτελείται από κάποιες λέξεις κλειδιά σε φυσική γλώσσα, οι οποίες πρέπει να γίνουν κατανοητές από τη μηχανή.
2. **Μετατροπή ερωτήματος (query transformation)**: πρόκειται για ένα **σύνολο τεχνικών** που εφαρμόζονται στο ερώτημα του χρήστη, ώστε να βελτιωθεί πριν και μετά τη διαδικασία παραγωγής αποτελεσμάτων για το χρήστη σε μορφή καταταγμένων εγγράφων. Πολλές από αυτές τις τεχνικές είναι **ίδιες με αυτές που εφαρμόζονται κατά τη μετατροπή εγγράφων στην ευρετηρίαση** (διαίρεση σε σύμβολα, διακόπτουσες λέξεις κλπ.). **Άλλες πρόσθετες** αφορούν

στον ορθογραφικό έλεγχο (spell checking) και σε εναλλακτικές ερωτήματος (query suggestion), προκειμένου το ερώτημα να γίνει πιο συγκεκριμένο.

3. **Εξαγωγή αποτελεσμάτων (results output)**: ουσιαστικά αφορά στην παρουσίαση της λίστας καταταγμένων εγγράφων, όπου περιλαμβάνονται ποικίλες εργασίες όπως:
    - Παραγωγή περιλήψεων ανακτηθέντων εγγράφων γνωστά ως ψήγματα αποτελεσμάτων (result snippets).
    - Τονισμό σημαντικών λέξεων μέσα στο έγγραφο.
    - Ομαδοποίηση των αποτελεσμάτων ώστε να αναγνωριστούν οι επιμέρους κλάσεις (π.χ. θεματικές κλάσεις).
    - Αναζήτηση κατάλληλων διαφημίσεων για την προβολή αποτελεσμάτων.
    - Μετάφραση αποτελεσμάτων.

Κατάταξη εγγράφων: αποτελεί τη δεύτερη υπό-διαδικασία της αναζήτησης και σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) αφορά μόνο την κατάταξη των εγγράφων. Καθώς το ερώτημα του χρήστη έχει μετατραπεί ήδη σε όρους ευρετηρίου, βαθμολογούνται τα έγγραφα με βάση έναν αλγόριθμο κατάταξης (ανάλογα το μοντέλο ΑΠ) και παράγεται μια λίστα με τα καταταγμένα έγγραφα. Πιο συγκεκριμένα σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) τα συστατικά κατάταξης είναι τα ακόλουθα:

1. **Βαθμολόγηση (scoring)**, η διαδικασία κατά την οποία υπολογίζονται βαθμολογίες για τα έγγραφα χρησιμοποιώντας τον αλγόριθμο κατάταξης (ranking algorithm).
2. **Βελτιστοποίηση απόδοσης (performance optimization)**, αφορά στην σχεδίαση αλγορίθμων κατάταξης και ευρετηρίων ώστε ο χρόνος απόκρισης να ελαχιστοποιηθεί.
3. **Κατανομή (distribution)** της λίστας κατάταξης.

Αξιολόγηση: στο στάδιο αυτό πραγματοποιείται μέτρηση και έλεγχος της αποτελεσματικότητας και της απόδοσης της μηχανής. Σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) η αξιολόγηση αφορά στα:

1. **Αρχεία καταγραφής (log files ή logs) χρηστών**: πρόκειται για στοιχεία αλληλεπίδρασης μεταξύ των χρηστών με τις μηχανές αναζήτησης. Χρησιμοποιούνται σε πληθώρα εργασιών (π.χ. ορθογραφικός έλεγχος). Ιδιαίτερα χρήσιμα για την αξιολόγηση αλλά και τους αλγόριθμους κατάταξης θεωρούνται τα **αρχεία καταγραφής πλοήγησης (clickthrough logs)**, τα οποία περιλαμβάνουν τις επιλογές του χρήστη για περιήγηση για συγκεκριμένη λίστα κατάταξης.
2. **Ανάλυση κατάταξης (ranking analysis)**: συνίσταται στον **υπολογισμό της αποτελεσματικότητας** του αλγόριθμου κατάταξης. Επίσης περιλαμβάνει και τη σύγκρισή του με άλλους αλγόριθμους.
3. **Ανάλυση απόδοσης (performance analysis)**: **υπολογισμός της απόδοσης** της μηχανής αναζήτησης και τρόποι βελτίωσης.

ΚΕΦΑΛΑΙΟ 3ο

ΥΠΟΛΟΓΙΣΤΙΚΗ ΚΑΙ ΠΟΣΟΤΙΚΗ ΓΛΩΣΣΟΛΟΓΙΑ

3.1 Κείμενο και Γλωσσολογία

Όπως έχει ήδη αναφερθεί στο παραπάνω κεφάλαιο, το έγγραφο ως αντικείμενο πληροφορίας έχει κεντρική θέση στην ΑΠ. Η επεξεργασία των εγγράφων που απαρτίζουν τις συλλογές για την δημιουργία ενός ευρετηρίου προκειμένου να υπάρχει αποδοτικότερη αναζήτηση, είναι αναγκαία και επεξηγήθηκε αναλυτικά (βλέπε ενότητα 1.3.1.1).

Η επεξεργασία αυτή μπορεί να χωριστεί σε δύο κατηγορίες: την γλωσσολογική και τη μη-γλωσσολογική. Η γλωσσολογική επεξεργασία λοιπόν αφορά στον υπό-κλάδο της Γλωσσολογίας που ονομάζεται ΥΓ (βλέπε ενότητα 1.4) και όπως έχει αναφερθεί παραπάνω αφορά στην απόπειρα ανάπτυξης λογισμικού υπολογιστών προκειμένου να μπορέσει η μηχανή να κατανοήσει τη φυσική γλώσσα.

Για την κατανόηση της τομής των επιστημών της ΑΠ και της ΥΓ, είναι απαραίτητη η ανάλυση της έννοιας του κειμένου στη Γλωσσολογία η οποία αποτελεί το περιεχόμενο του εγγράφου στην ΑΠ.

3.1.1 Το κείμενο και τα χαρακτηριστικά του

Προκειμένου να οριστεί πλήρως η έννοια του κειμένου, είναι απαραίτητη η επεξήγηση της στο πλαίσιο μιας ευρύτερης διαδικασίας, αυτής της ανθρώπινης επικοινωνίας. Προκειμένου στην περιγραφή της διαδικασίας της **επικοινωνίας,** όπως περιγράφεται από τους Bolshakov και Gelbukh (2004), ορίζονται οι βασικές γλωσσολογικές έννοιες σημασία (meaning), κείμενο (text), γλώσσα (language) και οι μεταξύ τους σχέσεις. Ως **σημασία** ορίζεται ουσιαστικά **η πληροφορία που θέλει να μεταφέρει ένας άνθρωπος** μέσω της επικοινωνίας, η οποία αποτελεί και το βασικό στόχο της επικοινωνιακής διαδικασίας. Ως **κείμενο ορίζεται η φυσική αναπαράσταση** των σκέψεων μεταξύ δύο ατόμων που επικοινωνούν, το οποίο περιέχει λέξεις, κενά, σημεία στίξης και μέσω των συνδυασμών των παραπάνω σχηματίζονται προτάσεις και παράγραφοι. Τέλος, ως **φυσική γλώσσα ορίζεται ουσιαστικά ο μετατροπέας του νοήματος σε κείμενο** και αντιστρόφως**.** Από τους παραπάνω ορισμούς συμπεραίνεται ο ρόλος του κειμένου στο πλαίσιο της επικοινωνίας και η σχέση του με τη γλώσσα.

Σύμφωνα με Bolshakov και Gelbukh (2004) υπάρχουν τρία βασικά **χαρακτηριστικά** τα οποία διέπουν ένα κείμενο. Το πρώτο αφορά στον σκοπό ύπαρξης ενός κειμένου, δηλαδή στο ότι ένα κείμενο γεννιέται ώστε να κωδικοποιήσει

ένα σύνολο πληροφοριών, επομένως **έχει μια σημασία**, ένα νόημα το οποίο προορίζεται και αφορά κάποιους ανθρώπους. Για το λόγο αυτό ακριβώς πραγματοποιείται η επεξεργασία της φυσικής γλώσσας του κειμένου.

Το δεύτερο χαρακτηριστικό συνίσταται στο ότι όσες πληροφορίες και αν εμπεριέχονται σε ένα κείμενο, όσο πολύπλοκο και αν είναι η **δομή του είναι πάντοτε γραμμική**, αποτελεί δηλαδή ένα σύνολο από συμβολοσειρές, κενά, σημεία στίξης τα οποία σχηματίζουν μια πολύ μεγάλη γραμμή. Να σημειωθεί ότι οι πληροφορίες που αναπαρίστανται στο κείμενο με γραμμική δομή είναι μη γραμμικές.

Τέλος το τρίτο χαρακτηριστικό αφορά τη **δομή του κειμένου** από επιμέρους στοιχεία, τα οποία βρίσκονται «**εντεθειμένα σε ομοειδή δομή**» (nested structure) μέσα στο κείμενο. Για τα στοιχεία αυτά λοιπόν που απαρτίζουν το κείμενο στη βιβλιογραφία (Hoey 2003) γίνεται αναφορά στην ουσιαστική πραγματοποίηση **διαχωρισμού του κειμένου στα λεγόμενα κομμάτια (chunks) με αφορμή την αλληλεπίδραση μεταξύ συγγραφέα και αναγνώστη**, ώστε να μεταφερθεί η πληροφορία από τον μεν στον δε. Πιο συγκεκριμένα, ο διαχωρισμός αυτός πραγματοποιείται τόσο κατά την διάρκεια σύνταξης του κειμένου από τον συγγραφέα (απ' αρχής δημιουργίας του κειμένου δηλαδή, προκειμένου να υφίσταται μια λογική δομή στο κείμενο), όσο και κατά την διάρκεια της ανάγνωσης από τους αναγνώστες, οι οποίοι χρησιμοποιούν τη δομή ώστε να ερμηνεύσουν καλύτερα το κείμενο. Οι Bolshakov και Gelbukh (2004) αναφέρουν ακόμη **την σημασία των επιμέρους στοιχείων της δομής του κειμένου, τα οποία οργανώνονται σε λέξεις, προτάσεις, παραγράφους και όλα μαζί σχηματίζουν τον λόγο (discourse)**, ο οποίος έχει ως κύριο χαρακτηριστικό τη συνδετικότητα (connectivity) ή αλλιώς **συνεκτικότητα** (coherence).

Η συνεκτικότητα ενός κειμένου έγκειται στην κοινή σταθερότητα και συνέπεια όλων των στοιχείων του λόγου στο κείμενο ώστε να μεταφερθεί το νόημα που πρέπει μέσω αυτών. Έχοντας ένα κείμενο την οργάνωση αυτή, μέσω των παραπάνω χαρακτηριστικών τότε είναι εφικτή η ανάπτυξη μεθόδων ευφυούς επεξεργασίας κειμένου. Όπως τονίζεται στη βιβλιογραφία (Hoey 2003) **τα γραπτά κείμενα είναι συνεκτικά, έτσι ώστε να καθιστούν εφικτή την κατανόηση των σχέσεων των στοιχείων μέσα στο κείμενο**.

3.1.2 Αρχές κειμενικότητας

Οι De Beaugrande και Dressler (1981) στο πλαίσιο του υπο-κλάδου της **Κειμενογλωσσολογίας**, που μελετά τα κοινά χαρακτηριστικά που διέπουν τα κείμενα και τις μεταξύ τους διαφορές και διακρίσεις, χαρακτηρίζουν το **κείμενο ως μια επικοινωνιακή εμφάνιση** (communicative occurrence), η οποία πρέπει να πληροί **επτά αρχές κειμενικότητας**. Σε περίπτωση που κάποια από τις αρχές κειμενικότητας δεν ικανοποιείται τότε το κείμενο δε θεωρείται επικοινωνιακό και έτσι κατά συνέπεια δεν μπορεί να αντιμετωπιστεί καν ως κείμενο. Οι αρχές κειμενικότητας λοιπόν είναι άμεσα συνδεδεμένες με τη διαδικασία επικοινωνίας, όπως παρουσιάστηκε στην παραπάνω ενότητα (βλέπε ενότητα 3.1.1).

Οι αρχές κειμενικότητας χωρίζονται σε **κείμενο-κεντρικές** (απευθύνονται στο υλικό του κειμένου δηλαδή), που περιλαμβάνουν τις δύο πρώτες αρχές και σε **χρηστο-κεντρικές** (αφορούν παραγωγούς και αποδέκτες στην διαδικασία επικοινωνίας), που περιλαμβάνουν τις υπόλοιπες πέντε. Σύμφωνα με τους De Beaugrande και Dressler (1981), οι αρχές κειμενικότητας είναι οι ακόλουθες:

1. **Συνοχή** (cohesion): αφορά στους τρόπους σύνδεσης των συστατικών του επιφανειακού κειμένου (surface text), μέσα σε μια πρόταση, οι οποίοι έχουν να κάνουν κυρίως με γραμματικούς κανόνες. Οι De Beaugrande και Dressler (1981) τονίζουν πόσο σημαντική είναι η αλληλεπίδραση της συνοχής με τις υπόλοιπες αρχές κειμενικότητας προκειμένου να είναι αποδοτική η επικοινωνία.
2. **Συνεκτικότητα** (coherence): κάθε κείμενο για να μπορεί να είναι κατανοητό από το ανθρώπινο μυαλό πρέπει να αλληλεπιδρά η γνώση που περιέχεται στο κείμενο με τις γνώσεις που έχει ο άνθρωπος αποθηκευμένες στον εγκέφαλο του για τον κόσμο. Έτσι σε κάθε κείμενο υπάρχουν κάποιες έννοιες (concepts) και κάποιες σχέσεις (relations) οι οποίες συνδέουν τις έννοιες μεταξύ τους. Σύμφωνα με τους De Beaugrande και Dressler (1981) ο όρος **έννοια** ορίζεται ως «*το γνωσιακό περιεχόμενο το οποίο μπορεί να ανακτηθεί ή να ενεργοποιηθεί μέσα από την ενότητα και συνοχή του μυαλού*».
3. **Αποβλεπτικότητα** (intentionality): η πρόθεση του συγγραφέα-παραγωγού του κειμένου να δημιουργήσει ένα κείμενο που να πληροί τις πληροφοριακές ανάγκες του αποδέκτη.

4. **Αποδοχή** (acceptability): αφορά τη συμπεριφορά του δέκτη σε σχέση με το αν το κείμενο είναι επικοινωνιακό σε σχέση με τις πληροφοριακές του ανάγκες και αν το αποδέχεται.
5. **Πληροφοριακότητα** (informativity): σύμφωνα με τη βιβλιογραφία (Carstens 2001) η αρχή αυτή αφορά στην επικοινωνιακή αξία των μερών του κειμένου.
6. **Καταστασιακότητα** (situationality/contextuality): έχει να κάνει με τους διάφορους παράγοντες που καθιστούν ένα κείμενο σχετικό. Σύμφωνα με τη βιβλιογραφία (Carstens 2001) αφορά κυρίως το ρόλο του περικειμένου/συμφραζόμενα (context) στην ποιότητα της επικοινωνίας. Υποδηλώνει το κατά πόσο όσοι εμπλέκονται στην επικοινωνία έχουν γνώση των συμφραζομένων.
7. **Διακειμενικότητα** (intertextuality): σύμφωνα με τη βιβλιογραφία (Carstens 2001) η αρχή της διακειμενικότητας αφορά τη συμπεριφορά του παραγωγού ενός κειμένου στην πρόθεση του να παράγει πληροφορίες για έναν αποδέκτη. Μάλιστα σύμφωνα με τη βιβλιογραφία (Carstens 2001) οι αρχές κειμενικότητας αποδοχή και διακειμενικότητα θεωρούνται ζευγάρι καθώς για κάθε κείμενο υπάρχει απαραίτητα ο παραγωγός του και ο αποδέκτης του.

Τέλος, οι De Beaugrande και Dressler (1981) προσθέτουν πως εκτός των αρχών κειμενικότητας υπάρχουν ακόμη **κανονιστικές αρχές** (regulative principles), προκειμένου να υπάρχει έλεγχος της επικοινωνίας του κειμένου και αναφέρουν τις δύο βασικότερες: **αποδοτικότητα** (efficiency), όταν απαιτείται η ελάχιστη προσπάθεια (βλέπε ενότητα 3.1.5) από τους συμμετέχοντες ώστε η επικοινωνία να είναι ικανοποιητική και **αποτελεσματικότητα** (effectiveness), η αρχή η οποία αφορά τη δημιουργία συνθηκών για την επίτευξη ενός στόχου.

3.1.3 Περικείμενο

Στη βιβλιογραφία (Tanskanen 2006) παρουσιάζεται μια πολύ σημαντική έννοια στη Γλωσσολογία και αλληλένδετη με τη συνοχή του κειμένου, πρόκειται για την έννοια του περικείμενου. Το περικείμενο χωρίζεται στα εξής είδη: το **γλωσσολογικό περικείμενο**, το οποίο έχει να κάνει με το **γλωσσικό υλικό που βρίσκεται γύρω από το αντικείμενο προς εξέταση**, το **γνωσιακό περικείμενο** (cognitive) που αφορά τους γνωσιακούς παράγοντες επικοινωνίας (διανοητικές αναπαραστάσεις, γνωστική προσπάθεια που απαιτείται από τους επικοινωνούντες) και **κοινωνικό περικείμενο**, το οποίο αναφέρεται σε όλο το κανάλι επικοινωνίας, την

κατάσταση, τους επικοινωνούντες και τους ρόλους αλληλεπίδρασης. Σχετίζεται άμεσα με την καταστασιακότητα, μια από τις αρχές κειμενικότητας που αναλύθηκαν παραπάνω.

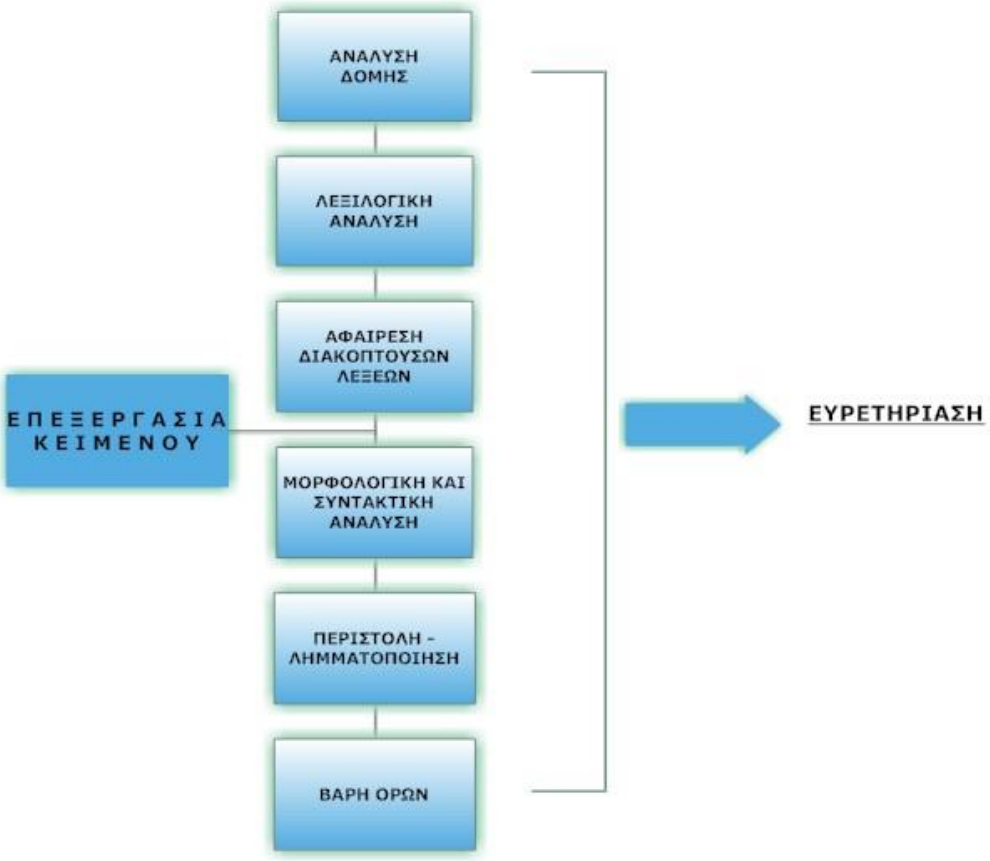
3.1.4 Επεξεργασία κειμένου για ανάκτηση και εξαγωγή πληροφορίας

Όπως έχει ήδη αναφερθεί (βλέπε ενότητα 1.1), η επεξεργασία της πληροφορίας είναι απαραίτητη προκειμένου να επιτευχθεί η ανάκτηση της. Πολλές φορές δε η ανάγκη ανάκτησης της πληροφορίας αφορά σε **ειδικευμένη ανάκτηση της πληροφορίας**. Σύμφωνα με τη βιβλιογραφία (Moens 2006) μέσω της Εξαγωγής Πληροφορίας (Information Extraction) είναι εφικτή η εξειδικευμένη ανάκτηση, καθώς η **Εξαγωγή Πληροφορίας** δεν εστιάζει μονάχα στην αντιστοίχιση του ερωτήματος του χρήστη (σε μορφή λέξεων κλειδιών φυσικής γλώσσας) με κάποια συναφή έγγραφα αλλά και στην αντιστοίχιση των σημασιολογικών τάξεων των οντοτήτων (και των μεταξύ τους σχέσεων) που φέρουν την πληροφορία στα έγγραφα.

Φυσικά η παραπάνω διαδικασία προϋποθέτει την κατανόηση της φυσικής γλώσσας στο πλαίσιο ενός ΣΑΠ. Όπως αναφέρεται στη βιβλιογραφία (Moens 2006) αυτό μπορεί να πραγματοποιηθεί μέσω της **Επεξεργασίας Φυσικής Γλώσσας,** στόχος της οποίας είναι η ανάλυση της **ανθρώπινης γλώσσας** ώστε να είναι εφικτή η **κατανόηση** της από τους υπολογιστές. Η Επεξεργασίας Φυσικής Γλώσσας εστιάζει στην επεξεργασία της δομής ενός κειμένου από γλωσσική άποψη και πιο συγκεκριμένα περιλαμβάνει **μορφολογική, συντακτική, σημασιολογική ανάλυση της γλώσσας**.

Με βάση λοιπόν τα παραπάνω, θα αναφερθούν επιγραμματικά οι συνηθέστερες διαδικασίες επεξεργασίας κειμένου, οι οποίες έχουν ενσωματωθεί στις μηχανές αναζήτησης για την ΑΠ σύμφωνα με τη βιβλιογραφία (Ceri et al. 2013), καθώς έχουν επεξηγηθεί αναλυτικά στο προηγούμενο κεφάλαιο (βλέπε ενότητα 2.4.1). Πρόκειται για τις διαδικασίες: ανάλυσης δομής εγγράφου, λεξιλογικής ανάλυσης, αφαίρεσης διακοπτουσών λέξεων, ανίχνευσης φράσεων (μέσω μορφολογικής και συντακτικής ανάλυσης), περιστολής - λημματοποίησης, απόδοσης βαρών όρων.

Στην εικ. 14 φαίνονται σχηματικά όλες οι παραπάνω διαδικασίες επεξεργασίας κειμένου:



Εικ. 14 Διαδικασίες επεξεργασίας κειμένου

3.1.5 Ποιότητα κειμένου (κατανόηση και αναγνωσιμότητα)

Η ποιότητα ενός κειμένου εξαρτάται από το κατά πόσο αυτό ικανοποιεί το βασικό σκοπό δημιουργίας του. Για το λόγο αυτό η μέτρηση της αποτελεί σημαντικό ζήτημα στη Γλωσσολογία. Όπως ήδη έχει αναφερθεί ο βασικός στόχος ενός κειμένου έγκειται στη μεταφορά ενός νοήματος στον αποδέκτη/αναγνώστη που απευθύνεται.

Για να γίνει αυτό, το κείμενο θα πρέπει να είναι κατανοητό από τον αναγνώστη. Συνεπώς οι διάφορες παράμετροι που επηρεάζουν την **κατανόησή** (comprehensiveness) του θεωρούνται πολύ σημαντικές. Στη βιβλιογραφία (Mikk 2005) η κατανόηση και η **αναγνωσιμότητα** (readability) θεωρούνται ταυτόσημες έννοιες. Η αναγνωσιμότητα (readability) σύμφωνα με τη βιβλιογραφία (Richards και Schmidt 2002) ορίζεται ως το «***πόσο εύκολα ένα υλικό μπορεί να αναγνωστεί και να κατανοηθεί***» και αναφέρονται παράγοντες που την επηρεάζουν όπως το μέσο μήκος πρότασης ενός κειμένου ή η πολύπλοκη γραμματική που το διέπει.

Η ανάγνωση και κατανόηση συνδέεται άμεσα με το κανάλι επικοινωνίας και την ευρύτερη εγκεφαλική αντίληψη του ανθρώπου. Μάλιστα στη βιβλιογραφία (Dubay 2004) αναφέρεται η αρχή της Ελάχιστης Δυνατής Προσπάθειας στην ανθρώπινη ομιλία, όπου ο γλωσσολόγος και φιλόλογος Zipf G. K. Χρησιμοποίησε τη στατιστική ανάλυση της γλώσσας για να δείξει το πώς αυτή λειτουργεί. Η ελάχιστη δυνατή προσπάθεια συνδέεται άμεσα με την εξοικονόμηση ενέργειας και τη συχνότητα εμφάνισης λέξεων. Η στατιστική μελέτη των λέξεων που χρησιμοποιεί ο άνθρωπος για την επικοινωνία (ανάλογα την δυσκολία - ευκολία της λέξης) οδήγησε σε διάφορους στατιστικούς νόμους που αφορούν την ΠΓ (βλέπε ενότητα 3.3).

Σύμφωνα με τους Dale και Chall (1949) η **επιτυχία ενός κειμένου** ως προς την αναγνωσιμότητα **εξαρτάται** από τους εξής αλληλένδετους παράγοντες, **εστιάζοντας στο κείμενο**:

1. Το περιεχόμενο του κειμένου και το πόσο ενδιαφέρει τον αναγνώστη.
2. Ο τρόπος έκφρασης.
3. Η δομή και οργάνωση του κειμένου.

Ως επιτυχία στην αναγνωσιμότητα σύμφωνα με τους Dale και Chall (1949) θεωρείται η κατανόηση και ανάγνωση του κειμένου από τους αναγνώστες στο ελάχιστο δυνατό χρόνο και με την καταβολή της ελάχιστης δυνατής προσπάθειας, η οποία βέβαια δεν **εξαρτάται** αποκλειστικά μόνο από το κείμενο όπως παρουσιάστηκε παραπάνω αλλά **και από τους ίδιους τους αναγνώστες** και πιο συγκεκριμένα από παράγοντες όπως:

1. Επιδεξιότητα στην ανάγνωση.
2. Ευφυία.
3. Εμπειρία.
4. Ωριμότητα.
5. Ενδιαφέροντα.
6. Σκοπός ανάγνωσης.

Στη βιβλιογραφία (Mikk 2005) οι μέθοδοι έρευνας σχετικά με την εξασφάλιση της κατανόησης και της αναγνωσιμότητας στρέφονται προς δύο διαφορετικές κατευθύνσεις: από τη μια πλευρά μελετώνται οι κανόνες προκειμένου ένα κείμενο να έχει υψηλή αναγνωσιμότητα και από την άλλη οι **readability formulae, προκειμένου να μετρηθεί και να αξιολογηθεί η αναγνωσιμότητα του κειμένου**. Ειδικότερα μέσω της μέτρησης αυτής στην βιβλιογραφία (Zamanian και Heydari 2012) αναφέρεται πως μπορεί να γίνει πρόβλεψη της δυσκολίας

αναγνωσιμότητας κάθε κειμένου και πως η πρόβλεψη αυτή είναι ιδιαιτέρως χρήσιμη σε διάφορους τομείς όπως η εκπαίδευση και η συγγραφή κειμένων. Ουσιαστικά μέσω των μετρήσεων αυτών μπορεί να διασφαλιστεί ότι το κατάλληλο ανάγνωσμα θα δοθεί στο κατάλληλο επίπεδο αναγνώστη και μπορεί να επιτευχθεί έτσι η αποβλεπτικότητα (βλέπε ενότητα 3.1.2).

Αναλυτικότερα σε σχέση με την εφαρμογή των readability formulae, σύμφωνα με τη βιβλιογραφία (Mikk 2005) προκειμένου να διερευνηθεί η κατανόηση κειμένου, αρχικά θα πρέπει να υπάρχει ένα σύνολο χαρακτηριστικών ως προς τα οποία θα εξεταστεί ένα αντιπροσωπευτικό δείγμα κειμένων. Τα χαρακτηριστικά αυτά συνήθως καθορίζονται από ειδικούς ή μέσα από έρευνα και ερωτηματολόγια. Εν συνεχεία, διαμορφώνονται οι readability formulae, οι οποίες υπολογίζουν το πόσο πολύπλοκο μπορεί να θεωρηθεί ένα κείμενο σε σχέση με την κατανόηση του, χρησιμοποιώντας έναν δείκτη αναγνωσιμότητας. Το δείγμα κειμένων υποβάλλεται σε πειράματα, εκτεταμένη ανάλυση, καθιέρωση τιμών για την κατανόηση και στατιστικές επεξεργασίες προκειμένου να υπολογιστούν οι σχέσεις μεταξύ των **μεταβλητών πρόβλεψης** (predictor variables). Δύο από τις βασικότερες μεταβλητές είναι η **πολυπλοκότητα περιεχομένου** που αφορά το **λεξιλόγιο** και η **πολυπλοκότητα δομής** που σχετίζεται με το **μήκος κειμένου**.

Στη βιβλιογραφία (Zamanian και Heydari 2012), αναφέρονται τα ακόλουθα **πλεονεκτήματα** χρήσης των **“readability formulae”**:

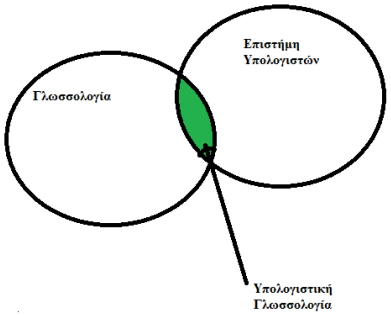
1. Μέσω των μετρήσεων ο συγγραφέας έχει στη διάθεσή του πληροφορίες ώστε να έρθει κοντά με το κοινό στο οποίο απευθύνεται και να μετατρέψει το κείμενο του σε ένα απλό κείμενο.
2. Η εφαρμογή τους πραγματοποιείται πριν το κείμενο φτάσει στον αναγνώστη μέσω υπολογιστών.

Ενώ παρουσιάζουν αντίστοιχα και τα ακόλουθα μειονεκτήματα:

1. Δεν προσδιορίζουν την κατανόηση από την πλευρά των αναγνωστών.
2. Παρουσιάζεται απόκλιση αποτελεσμάτων με χρήση διαφορετικών readability formulae.
3. Αδυναμία υπολογισμού διάφορων παραμέτρων όπως βαθμό ενδιαφέροντος ή συνεκτικότητας κειμένου.

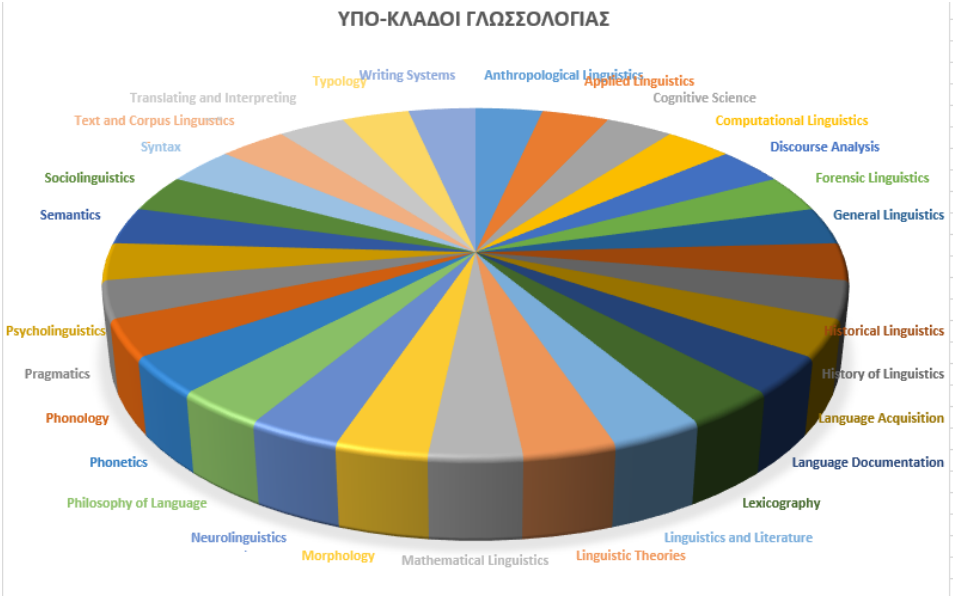
3.2 Εισαγωγή στην ΥΓ και τις βασικές της έννοιες

Σύμφωνα με τους Bolshakov και Gelbukh (2004) το κύριο μέλημα της ΥΓ είναι η δημιουργία υπολογιστικών προγραμμάτων για την αυτόματη Επεξεργασία Φυσικής Γλώσσας (π.χ. των λέξεων ή κειμένων). Η ΥΓ εφαρμόζεται σε πολλούς τομείς ορισμένοι από τους οποίους είναι ο διαχωρισμός λέξεων (hyphenation), ο γραμματικός έλεγχος (spell checking), **η ΑΠ**, η μηχανική μετάφραση, η εξόρυξη δεδομένων από το κείμενο. Όπως ήδη έχει αναφερθεί, η ΥΓ είναι ένας διεπιστημονικός κλάδος, ο οποίος αφορά στο συνδυασμό πολλών κλάδων επιστημών άλλων σε μικρότερο και άλλων σε μεγαλύτερο βαθμό, με κυρίαρχους την Επιστήμη των Υπολογιστών και τη Γλωσσολογία, όπως φαίνεται στην εικ. 15.



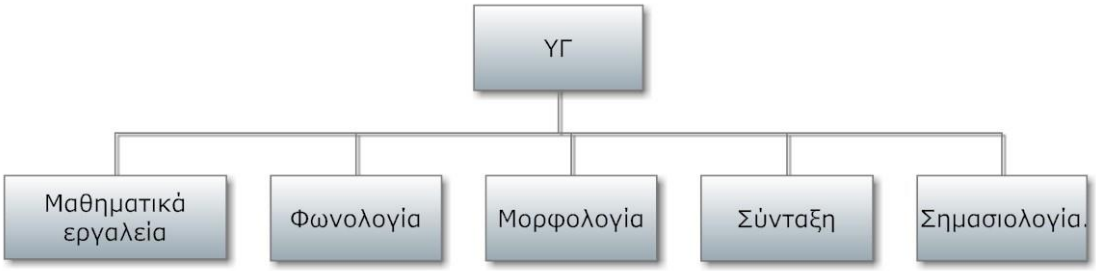
Εικ. 15 Ο υπο-κλάδος της ΥΓ

Γενικότερα, η ΥΓ σύμφωνα με την ***LINGUIST List*** (https://linguistlist.org/LL/LingSubfields.cfm#CogSci) αποτελεί έναν από τους πολλούς υπο-κλάδους της Γλωσσολογίας, όπως φαίνεται σχηματικά στην εικ. 16. Στο διάγραμμα παρατηρούνται και άλλοι επιστημονικοί και διεπιστημονικοί κλάδοι στους οποίους έχει ήδη γίνει αναφορά στη διατριβή όπως η Νευρογλωσσολογία και η Μορφολογία. Ο κλάδος της Μορφολογίας θα αναλυθεί παρακάτω (βλέπε ενότητα 3.2.1).



Εικ. 16 Υπο-κλάδοι Γλωσσολογίας

Οι Bolshakov και Gelbukh (2004) αναφέρουν τα πεδία που μελετά η ΥΓ, που αφορούν στον προφορικό και γραπτό λόγο. Στην εικ. 17 συνοψίζονται σε διαγραμματική μορφή:



Εικ. 17 Πεδία έρευνας ΥΓ

Αρχικά θα αναφερθούν κάποια βασικοί ορισμοί που χρησιμοποιούνται στην ΥΓ και αφορούν την παρούσα διατριβή.

Σκεπτόμενοι τα συστατικά δομής ενός κειμένου από το μικρότερο προς το μεγαλύτερο, η μικρότερη μονάδα φυσικής γλώσσας ονομάζεται **μόρφημα** (morph). Το **επόμενο επίπεδο γλωσσικής μονάδας αποτελεί η λέξη**. Σύμφωνα με τους Bolshakov και Gelbukh (2004) ως ***λέξη*** θεωρείται κάθε **υπο-συμβολοσειρά σε ένα κείμενο από τον πρώτο οριοθέτη (delimiter) ως τον επόμενο** οριοθέτη (δηλαδή ένα κενό ή κάποιο σημείο στίξης).

Στη συνέχεια θα επεξηγηθούν οι διάφοροι χαρακτηρισμοί που αφορούν τον όρο ***λέξη***. Με την έννοια ***εμφάνισης λέξης*** **(word occurrence)**, υποδηλώνεται η

επανάληψη λέξεων σε ένα κείμενο. Οι ομοιότητες των λέξεων, όπως το κοινό τους θέμα, μπορεί να αποτελέσει παράγοντα ομαδοποίησης τους μέσα στο συγκεκριμένο κείμενο, με βάση κάποιο κοινό νόημα, αν και μπορεί να έχουν διαφορετική μορφή. Συνεπώς το φαινόμενο αυτό πρέπει να χαρακτηριστεί με κάποιο τρόπο. Έτσι, **το σύνολο των συμβολοσειρών που έχουν το ίδιο νόημα αλλά εμφανίζονται σε διαφορετικές μορφές ονομάζεται *λέξημα* (lexeme), ενώ κάθε συμβολοσειρά του συνόλου αυτού ονομάζεται *μορφή λέξης* (word form)**. Έτσι λοιπόν ως μορφή λέξης θεωρείται κάθε εμφάνιση λέξης αλλά ταυτόχρονα οι μορφές λέξεων μπορούν να επαναλαμβάνονται μέσα στο κείμενο.

Για να είναι εφικτή η επεξεργασία της φυσικής γλώσσας **αποδίδονται σύμβολα για την ανάλυση των συστατικών (constituents) των κειμένων**, μέσω των οποίων αυτά κατατάσσονται σε διάφορες γραμματικές κατηγορίες. Παρατίθεται ένα σύνολο τέτοιων συμβόλων από τους Bolshakov και Gelbukh (2004) στον πίνακα 2 καθώς και ορισμένοι κανόνες παραγωγής:

Πίνακας 2. Γραμματικά σύμβολα και κανόνες παραγωγής

Γραμματικά σύμβολα	**Κανόνες παραγωγής**
S -- πρόταση	S → NP VP
NP – ονοματική φράση (με κέντρο το ουσιαστικό)	VP → V NP
VP – ρηματική φράση (με κέντρο το ρήμα)	NP → D N
N – ουσιαστικό	NP → N
V – ρήμα	
D -- προσδιοριστές (a, an, the)	

Εξετάζοντας τους κανόνες παραγωγής ισχύει η εξής σχέση μεταξύ των συμβόλων: από τη μια πλευρά παρουσιάζονται οι **κλάσεις μερών του λόγου με βάση τις οποίες κατηγοριοποιούνται οι λέξεις** και από την άλλη παρουσιάζονται **οι σχέσεις μεταξύ των κλάσεων**, ως συστατικά του κειμένου. Πιο συγκεκριμένα,

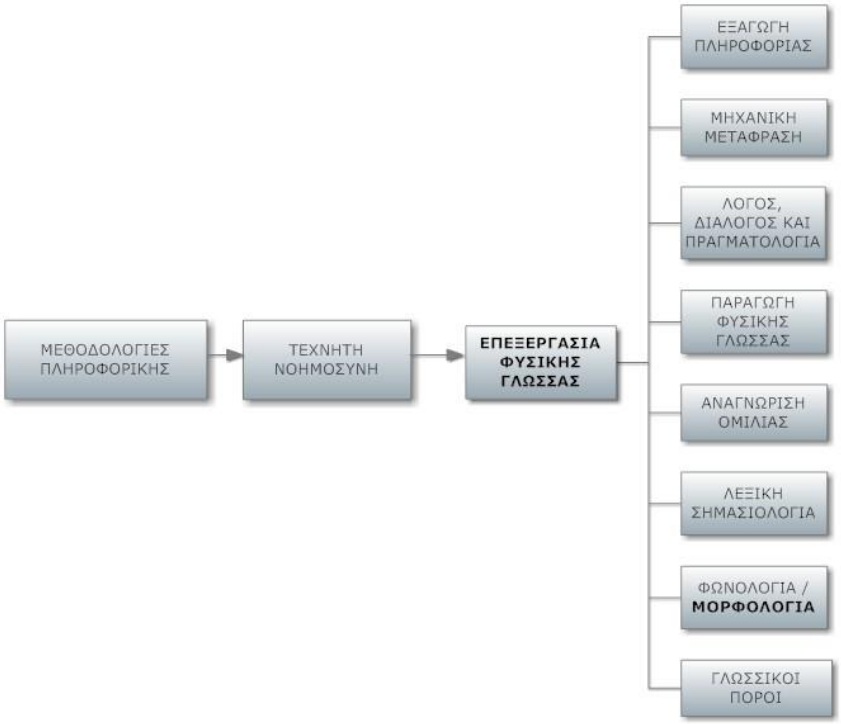
χρησιμοποιώντας σαν παράδειγμα τον πρώτο κανόνα παραγωγής (S → NP VP), είναι κατανοητό ότι η ονοματική και ρηματική φράση αποτελούν συστατικά της πρότασης.

Στις παρακάτω ενότητες θα αναλυθούν τα πεδία της ΥΓ που αφορούν το γραπτό λόγο και τα κείμενα (βλέπε εικ. 17), δηλαδή η Μορφολογία, η Σύνταξη και η Σημασιολογία.

3.2.1 Μορφολογία

Σύμφωνα με Jurafsky και Martin (2000) η Μορφολογία (Morphology) αφορά τη **μελέτη των βασικών μονάδων που σχηματίζουν τις λέξεις, τα μορφήματα**, τα οποία αποτελούν τη μικρότερη νοηματική μονάδα στη γλώσσα. Αποτελεί από μόνη της έναν ξεχωριστό υπο-κλάδο της Γλωσσολογίας όπως φαίνεται και στην εικ. 16, όμως αποτελεί αντικείμενο μελέτης και για την ΥΓ ειδικότερα (βλέπε εικ. 17).

Ακόμη στον κλάδο της Επιστήμης των Υπολογιστών, με βάση το σύστημα ταξινόμησης της ACM (ανακτήθηκε 25/02/2015 από http://www.acm.org/about/class/2012), η Μορφολογία, χρησιμοποιείται και μελετάται και από την Τεχνητή Νοημοσύνη και πιο συγκεκριμένα αποτελεί υπο-κλάδο της Επεξεργασίας Φυσικής Γλώσσας (βλέπε ενότητα 3.1.4) όπως φαίνεται και στην εικ. 18:



Εικ. 18 Επεξεργασία Φυσικής Γλώσσας και Μορφολογία

Στη Μορφολογία οι δύο βασικές ομάδες στις οποίες κατηγοριοποιούνται τα μορφήματα είναι **τα θέματα (stems)**, τα οποία αποτελούν το κύριο μόρφημα κάθε λέξης και δίνουν το βασικό νόημα στη λέξη και **τα προσφύματα (affixes)**, τα οποία δίνουν επιπλέον νοήματα σε μια λέξη. **Τα προσφύματα με τη σειρά τους διακρίνονται** και αυτά σε τέσσερις μεγάλες κατηγορίες: τα **προθήματα** (prefixes) που προηγούνται του θέματος, τα **επιθήματα** (suffixes) που έπονται του θέματος, τα ενθήματα (infixes) που εισάγονται μέσα στο θέμα και τα **περιθήματα** (circumfixes) τα οποία μπορεί να προηγούνται και να έπονται του θέματος. Μάλιστα μια λέξη μπορεί να έχει πάνω από ένα πρόσφυμα. Οι παραπάνω κατηγορίες χρησιμοποιούνται για τη συντακτική ανάλυση.

3.2.2 Σύνταξη

Ως **συντακτική ανάλυση (parsing)** οι Jurafsky και Martin (2000) ορίζουν την **εισαγωγή δεδομένων (input) και ως αποτέλεσμα την εξαγωγή κάποιας δομής για τα δεδομένα αυτά**. Έτσι στη συντακτική ανάλυση περιλαμβάνεται το θέμα της λέξης καθώς και άλλα μορφολογικά της χαρακτηριστικά, όπως το μέρος του λόγου που ανήκει. Προκειμένου να μπορεί να γίνεται μορφολογική ανάλυση αυτόματα, για να δημιουργηθεί δηλαδή ένας μορφολογικός συντακτικός αναλυτής (morphological parser) χρειάζονται τρία βασικά στοιχεία:

1. **Λεξικό** (lexicon), με τη λίστα θεμάτων και προσφυμάτων και συνοδευτικές βασικές πληροφορίες για αυτά. Το λεξικό συνεργάζεται με τους μορφοτακτικούς περιορισμούς ώστε να καλύψει όλους τους πιθανούς συνδυασμούς σχηματισμού λέξεων, καθώς όλες οι λέξεις μιας γλώσσας είναι αδύνατο να περιληφθούν σε ένα λεξικό.
2. **Μορφοτακτικοί περιορισμοί** (morphotactics), υποδεικνύουν την σειρά που ακολουθούν τα μορφήματα (χωρισμένα σε βασικές κλάσεις) προκειμένου να σχηματιστεί μια λέξη.
3. **Ορθογραφικοί κανόνες** (orthographic rules).

Σύμφωνα με τους Jurafsky και Martin (2000) προκειμένου να αναγνωριστεί μια εισαγωγή δεδομένων αν αποτελεί λέξη και να πραγματοποιηθεί η περαιτέρω ανάλυση της, τότε χρησιμοποιούνται τα finite state automata (FSAs), πρόκειται για λογισμικό, το οποίο χρησιμοποιεί το λεξικό σε συνδυασμό με τους μορφοτακτικούς κανόνες.

Η μορφολογική συντακτική ανάλυση μπορεί να πραγματοποιηθεί είτε χρησιμοποιώντας δύο είτε τρία επίπεδα, όταν περιλαμβάνει και τους κανόνες ορθογραφίας. Θα παρουσιαστεί η ανάλυση με τρία επίπεδα που θεωρείται και πιο πλήρης ώστε να γίνουν κατανοητά **τα στάδια που περνάει ένα αυτόματος μορφολογικός συντακτικός αναλυτής για την ανάλυση μιας λέξης**: μια λέξη μπορεί να διακριθεί σε τρία επίπεδα, το λεξικό (lexical), το οποίο αναπαριστά την λέξη ως απλή συνένωση μορφημάτων, το ενδιάμεσο (intermediate) το οποίο διορθώνει την ορθογραφία με βάση τους κανόνες και το επιφανειακό (surface), το οποίο δείχνει την πραγματική ορθογραφία της λέξης. Για να επιτευχθεί η μορφολογική συντακτική ανάλυση χρησιμοποιούνται κανόνες αντιστοίχισης (mapping rules), οι οποίοι αντιστοιχίζουν μια λέξη σε μια ακολουθία μορφημάτων συνοδευμένη με επιμέρους μορφολογικά χαρακτηριστικά. Έτσι ένας μετατροπέας αντιστοιχίζει μεταξύ του finite-state transducer (FST), μια υπολογιστική συσκευή για τη μοντελοποίηση της Μορφολογίας, από ένα σύνολο συμβόλων σε ένα άλλο.

Σύμφωνα με τους Jurafsky και Martin (2000) υπάρχουν και μορφολογικοί συντακτικοί αναλυτές – μετατροπείς, οι οποίοι δε χρησιμοποιούν λεξικά και αυτοί χρησιμοποιούνται κυρίως στην ΑΠ.

3.2.2.1 Μέρη του λόγου και επισημειωτές

Σύμφωνα με τους Jurafsky και Martin (2000) ως **μέρη του λόγου** (part of speech/POS) ορίζονται **οι κλάσεις στις οποίες ομαδοποιούνται οι λέξεις** και από τις οποίες αντλούνται αρκετές πληροφορίες σχετικά με τη λέξη αλλά και τις γειτονικές τις. Πολλές φορές στη βιβλιογραφία αναφέρονται και ως μορφολογικές κλάσεις ή λεξιλογικές ετικέτες (lexical tags). Οι πληροφορίες αυτές που αντλούνται από το μέρος του λόγου της λέξης χρησιμοποιούνται σε πολλούς επιστημονικούς κλάδους όπως και στην ΑΠ, με πολλές διαφορετικές εφαρμογές. Ειδικότερα αναφέρουν οι Jurafsky και Martin (2000), την διαδικασία της περιστολής (stemming) ώστε να μπορούν να αναγνωριστούν τα είδη των προσφυμάτων/καταλήξεων που επιδέχεται μια λέξη, την ανάκτηση προκειμένου να πραγματοποιηθεί ανάκτηση συγκεκριμένων μερών του λόγου, τη δημιουργία αλγόριθμων αποσαφήνισης σημασίας λέξεων (automatic word-sense disambiguating algorithms), τη συντακτική ανάλυση κειμένων και γενικότερες εφαρμογές εξαγωγής πληροφοριών.

Πιο συγκεκριμένα, σχετικά με τις κλάσεις των μερών του λόγου οι Jurafsky και Martin (2000) αναφέρουν πως **οι κλάσεις αυτές χωρίζονται σε δύο υπο-κλάσεις**: τις **κλειστές κλάσεις** (closed class), όπου τα μέλη τους είναι σταθερά σε αντίθεση με τις **ανοικτές κλάσεις** (open class), όπου υπάρχει πιθανότητα διαρκούς εμπλουτισμού των κλάσεων με νέες λέξεις που προκύπτουν είτε μέσω της επινόησης είτε δανεισμού από άλλες γλώσσες. Στις κλειστές κλάσεις ανήκουν: προθέσεις (prepositions), προσδιοριστές (determiners), αντωνυμίες (pronouns), σύνδεσμοι (conjunctions), βοηθητικά ρήματα (auxiliaries), μόρια (particles), αριθμοί (numerals). Στις ανοικτές ανήκουν: ουσιαστικά (nouns), ρήματα (verbs), επίθετα (adjectives), επιρρήματα (adverbs). Οι Jurafsky και Martin (2000) ορίζουν τη **διαδικασία επισημείωσης μερών του λόγου (part of speech tagging/tagging)** ως τη **διαδικασία με την οποία το λογισμικό αυτό αναγνωρίζει το μέρος του λόγου στο οποίο ανήκει μια λέξη**. Η διαδικασία της επισημείωσης **μπορεί να ταυτιστεί σε επίπεδο φυσικής γλώσσας με τη διαδικασία διαίρεσης σε σύμβολα (**βλέπε ενότητα 2.4.1.1**).** Οι επισημειωτές (taggers) είναι **πολύ σημαντικοί στην ΑΠ** αλλά το μεγαλύτερο τους πρόβλημα αποτελεί η ασάφεια των λέξεων και η σωστή ανάθεση μιας ετικέτας (tag), η οποία αντλείται από ένα σύνολο ετικετών (tagset).

Με βάση τους αλγόριθμους ανάθεσης ετικετών οι επισημειωτές χωρίζονται σε δύο κατηγορίες, τους **επισημειωτές βασισμένους σε κανόνες** (βάση δεδομένων με χειρόγραφους κανόνες αποσαφήνισης) και τους **στοχαστικούς επισημειωτές** (χρησιμοποιούν εκπαιδευμένο σώμα κειμένου για την αποσαφήνιση). Σύμφωνα με τους Jurafsky και Martin (2000) από το συνδυασμό των δύο παραπάνω κατηγοριών επισημειωτών έχουν προκύψει **οι επισημειωτές Transformation-Based Tagging ή Brill tagging**, οι οποίοι χρησιμοποιούν κανόνες προκειμένου να αναθέσουν ετικέτες σε λέξεις αλλά χρησιμοποιούν και την τεχνική εκμάθησης μηχανής, καθώς υποθέτουν ένα εκπαιδευμένο σώμα κειμένου (corpus) όπου ήδη έχουν ανατεθεί ετικέτες για αυτόματη πρόκληση των κανόνων. Ειδικότερα η λειτουργία τους είναι η εξής:

- Αρχικά πραγματοποιείται ανάθεση ετικετών στις λέξεις με βάση ένα σώμα κειμένου, που ήδη φέρει ετικέτες, όπως π.χ. το Brown corpus.
- Αφού ανατεθεί η πιθανότερη ετικέτα τότε εφαρμόζονται οι κανόνες μετατροπής και διορθώνονται οι αρχικές ετικέτες, αν χρειαστεί στην πορεία.

- Η εφαρμογή των κανόνων για ανάθεση ετικετών σε πρώτη φάση πραγματοποιείται σε μεγάλο μέρος του κειμένου με γενικούς κανόνες ενώ στη συνέχεια σταδιακά εφαρμόζονται κανόνες που εμπίπτουν σε όλο και μικρότερο μέρος κειμένου, φτάνοντας πια σε εντελώς εξειδικευμένους κανόνες για μεμονωμένες περιπτώσεις.

Όπως είναι φυσικό σε μια λέξη ανατίθενται αρκετές ετικέτες ώσπου να προκύψει η τελική απόφαση, η οποία διαμορφώνεται μέσα από την εφαρμογή των κανόνων. Απαραίτητο για την εφαρμογή όλων των προαναφερθέντων αλγορίθμων είναι ένα λεξικό με τα μέρη του λόγου για κάθε λέξη που υπάρχει (όσο αυτό είναι δυνατό), καθώς οι γλώσσες εμπλουτίζονται διαρκώς, συνεπώς είναι πρακτικά αδύνατο ένα λεξικό να τις περιέχει όλες.

Σύμφωνα με τους Jurafsky και Martin (2000), η σύνταξη αφορά το πως είναι οργανωμένες οι λέξεις και τα μέρη του λόγου μεταξύ τους, με σημαντικότατη την έννοια της **συστατικότητας (constituency),** η οποία αναφέρεται σε κλάσεις λέξεων/φράσεων οι οποίες συμπεριφέρονται ως μια μονάδα, δηλαδή ως ένα συστατικό. Όλες αυτές οι δομές μεταξύ των λέξεων και των συστατικών μοντελοποιούνται με μαθηματικά συστήματα όπως το context-free grammar (CFG). Αυτά τα συστήματα αποτελούνται από ένα σύνολο από κανόνες βάση των οποίων ομαδοποιούνται σύμβολα της γλώσσας καθώς και από ένα λεξικό που περιέχει λέξεις και τα αντίστοιχα σύμβολα.

Η συντακτική ανάλυση στην ΥΓ ταυτίζεται με την διαδικασία της περιστολής στην ΑΠ.

3.2.3 Σημασιολογία

Όπως φαίνεται και στην εικ. 17, η ΥΓ εκτός από την Μορφολογία και την Σύνταξη αφορά και την Σημασιολογία.

Όπως αναφέρει ο Kracht (2007) ως Σημασιολογία (Semantics) ορίζεται ο **τομέας της Γλωσσολογίας που ασχολείται με την σημασία των λέξεων**. Πιο συγκεκριμένα, ο Hayes (2010) συμπληρώνει πως η Σημασιολογία μελετά το **πώς η σημασία των λέξεων μεταφέρεται μέσω της γλώσσας**. Ως σημασία ορίζονται τα νοήματα τα οποία μεταφέρονται μέσα από τις λέξεις, τις προτάσεις. Ουσιαστικά η γλώσσα μπορεί να θεωρηθεί ως ένα από τα πιο περίπλοκα συστήματα συμβόλων. Έτσι και οι προτάσεις με τη σειρά τους αποτελούν σύμβολα τα οποία εκφράζουν τη σκέψη του ατόμου που τις δομεί. Συνεπώς μέσω της γλώσσας επιτυγχάνεται η

μεταφορά των σκέψεων του. Όπως αναφέρει ο Kracht (2007), στην Σημασιολογία οι προτάσεις οι οποίες μπορούν να χαρακτηριστούν αληθής ή ψευδής ονομάζονται δηλώσεις (statements), οι οποίες εκφράζουν λογικές προτάσεις (propositions). Τέλος, ο Hayes (2010) αναφέρεται στον στόχο της Σημασιολογίας, ο οποίος είναι «*η μελέτη του πως η γλώσσα εμπεριέχει τη σκέψη, χωρίς ακόμα να υπάρχει μια καλώς ανεπτυγμένη θεωρία περί της σκέψης*».

3.2.3.1 Προβλήματα στη Σημασιολογία

Το μεγάλο όμως **εμπόδιο για την επικοινωνία** και συνάμα το αδύνατο σημείο της γλώσσας αποτελεί, όπως αναφέρεται στη βιβλιογραφία (Karaman 2003), η **αμφισημία** (ambiguity), η οποία αν και αποτελεί μια αναπόσπαστη ιδιότητα της φυσικής γλώσσας, **παρακωλύει τη διαδικασία της επικοινωνίας**. Η έλλειψη αυτή της σαφήνειας του νοήματος αντιμετωπίζεται από τον άνθρωπο υποσυνείδητα καθώς εκτελούνται πραγματικές και σημασιολογικές διεργασίες, στις οποίες μεγάλο ρόλο διαδραματίζει το περικείμενο (βλέπε ενότητα 3.1.3), το οποίο περιβάλλει την λέξη που εμφανίζει την ασάφεια.

Σύμφωνα με τον Kennedy (2009) **η αμφισημία (ambiguity) αφορά τη συσχέτιση μιας ακολουθίας χαρακτήρων με πολλές διαφορετικές σημασίες** και χωρίζεται σε φωνολογική, λεξιλογική, δομική και πεδίου. Ειδικότερα όσον αφορά την σημασιολογική αμφισημία όπως αναφέρουν οι Jurafsky και Martin (2000) η αμφισημία είναι το φαινόμενο όπου το μέρος του λόγου στο οποίο ανήκει μια λέξη μέσα σε μια πρόταση δεν είναι σαφές. Έτσι μέσω της διαδικασίας της αποσαφήνισης (disambiguation) διευκρινίζεται ο ρόλος της λέξης.

Σύμφωνα με τους Jurafsky και Martin (2000) ως **πολυσημία** ορίζεται το φαινόμενο κατά το οποίο **ένα λέξημα μπορεί να αντιστοιχεί σε πολλαπλές σχετικές έννοιες.** Μάλιστα στη βιβλιογραφία (Kovacs 2011) αναφέρεται πως το φαινόμενο της πολυσημίας έγκειται στο ότι δεν είναι εφικτή η διάκριση μεταξύ των διαφορετικών εννοιών μιας λέξης, καθώς και στο ότι δεν υπάρχει σαφής εικόνα ως προς το πόσες διαφορετικές έννοιες μπορεί να έχει μια λέξη. Ακόμη το νόημα μιας λέξης μπορεί να διακρίνεται σε κυριολεκτικό και μεταφορικό. Τέλος, ένα από τα μεγαλύτερα προβλήματα που αντιμετωπίζονται σε σχέση με την **πολυσημία είναι η διάκριση της από την ομωνυμία** (homonymy). Οι Jurafsky και Martin (2000) αναφέρουν την ομωνυμία ως τη σχέση μεταξύ λέξεων που έχουν την ίδια μορφή αλλά τα νοήματα

τους δεν συσχετίζονται. Ακόμη στη βιβλιογραφία αναφέρεται (Kovacs 2011) πως στην ομωνυμία οι λέξεις δε σχετίζονται ετυμολογικά ενώ προφέρονται με τον ίδιο τρόπο ή έχουν την ίδια ορθογραφία. Από την άλλη πλευρά στην πολυσημία η ετυμολογία είναι η ίδια, άρα υπάρχει σημασιολογική σχέση και τα διαφορετικά νοήματα πηγάζουν από τη μεταφορική χρήση της λέξης. Καθώς με την ομωνυμία και την πολυσημία η αμφισημία ενδυναμώνεται έγκειται στο περικείμενο η αποσαφήνιση του νοήματος. Τέλος, σύμφωνα με τους Jurafsky και Martin (2000) και η **συνωνυμία** αποτελεί ένα φαινόμενο της γλώσσας που παρακωλύει την κατανόηση της γλώσσας με αυτόματο τρόπο, καθώς πρόκειται για το φαινόμενο κατά το οποίο **διαφορετικά λεξήματα αντιστοιχούν σε ένα κοινό νόημα.**

Γενικότερα **τόσο η πολυσημία όσο και η συνωνυμία αποτελούν σημαντικά ζητήματα και για την ΑΠ καθώς επηρεάζουν την ακρίβεια και την ανάκληση** (βλέπε ενότητα 2.3.3). Πιο συγκεκριμένα, όπως αναφέρουν οι Jurafsky και Martin (2000), χωρίς να είναι κανείς απόλυτος, η πολυσημία τείνει να μειώνει την ακρίβεια, καθώς επιστρέφει αποτελέσματα μη σχετικά ως προς τις πληροφοριακές ανάγκες του χρήστη ενω η συνωνυμία τείνει να μειώνει την ανάκληση, εφόσον έγγραφα που είναι σχετικά με τις πληροφοριακές ανάγκες του χρήστη παραλείπονται.

3.2.3.2 Μέτρα σημασιολογικής εγγύτητας (ομοιότητας)

Τα μέτρα σημασιολογικής εγγύτητας, όπως αναφέρουν οι Harispe et al. (2013), αποτελούν **μέτρα, με την έννοια μαθηματικού εργαλείου, μέσω των οποίων είναι δυνατός ο υπολογισμός της σημασιολογικής συνάφειας στοιχείων, τα οποία μπορεί να είναι γλωσσικές μονάδες, έννοιες ή οντότητες, οι οποίες αντλούνται από κείμενα** (αδόμητα ή ημι-δομημένα). Για τον υπολογισμό της εγγύτητας αναφέρουν πως υπάρχει πληθώρα μέτρων που υπολογίζουν την ομοιότητα ή διαφορά (γενικότερα την απόσταση) ανάμεσα σε συγκεκριμένες δομές δεδομένων (π.χ. διανύσματα) και σε τύπους δεδομένων (π.χ. αριθμητικά δεδομένα ή συμβολοσειρές). Η εφαρμογή των σημασιολογικών μέτρων έχει διεπιστημονικό χαρακτήρα και αφορά τις Γνωσιακές Επιστήμες, τη Γλωσσολογία, την Επεξεργασία Φυσικής Γλώσσας, το Σημασιολογικό Ιστό και άλλους επιστημονικούς κλάδους.

Όπως αναφέρεται στη βιβλιογραφία (Harispe et al. 2013) μέσω της σημασιολογικής εγγύτητας **ένα στοιχείο προς σύγκριση προέρχεται από έναν σημασιολογικό χώρο,** όπου π.χ. το στοιχείο θα μπορούσε να αντιστοιχεί σε μια

πρόταση και ο σημασιολογικός χώρος σε ένα κείμενο. Μάλιστα η **έννοια του σημασιολογικού χώρου μπορεί να αντιστοιχηθεί με έναν δειγματικό χώρο με σημασιολογική υπόσταση** και τα αντίστοιχα στοιχεία που περιλαμβάνει δύναται να αναπαρασταθούν μέσω μιας συγκεκριμένης δομής δεδομένων όπως π.χ. ένα διάνυσμα. Έτσι και η αναπαράσταση αυτή παίρνει πλέον **σημασιολογικές διαστάσεις.**

Τα **σημασιολογικά μέτρα ομοιότητας** κατηγοριοποιούνται όπως αναφέρεται στη βιβλιογραφία (Harispe et al. 2013) σύμφωνα με το είδος των στοιχείων προς σύγκριση, τη σημασιολογική πηγή (semantic proxy) από όπου προέρχονται οι πληροφορίες για τα στοιχεία και τέλος την κανονική μορφή (canonical form) που χρησιμοποιείται για την αναπαράσταση των στοιχείων. Για τη σχεδίαση των σημασιολογικών μέτρων ορίζεται μια συνάρτηση υπολογισμού της ομοιότητας.

Ειδικότερα για την περίπτωση αναπαράστασης λέξεων, όπως αναφέρεται στη βιβλιογραφία (Harispe et al. 2013) η κλασική κανονική μορφή υλοποιείται μέσω της αναπαράστασης με διανύσματα, η οποία για την ΑΠ αντιστοιχεί στο γνωστό μοντέλο VSM (βλέπε ενότητα 2.2.1).

Έτσι σύμφωνα με τη βιβλιογραφία (Harispe et al. 2013) τα σημασιολογικά μέτρα ομοιότητας χωρίζονται σε **τρεις μεγάλες κατηγορίες: τα κατανεμημένα (distributional), τα βασισμένα σε γνώση (knowledge-based) και το συνδυασμό αυτών**. Πιο συγκεκριμένα τα κατανεμημένα αφορούν σύγκριση γλωσσικών μονάδων, οι οποίες προέρχονται από σημασιολογική πηγή που αποτελεί κείμενο, ενώ τα βασισμένα σε γνώση αφορούν μονάδες όπως έννοιες, ομάδες εννοιών, οι οποίες προέρχονται από σημασιολογική πηγή δομημένης γνώσης, όπως θησαυροί ή οντολογίες.

Ειδικότερη σημασία για τη διατριβή έχουν τα **κατανεμημένα** καθώς αφορούν σε κείμενα και χωρίζονται στις εξής προσεγγίσεις:

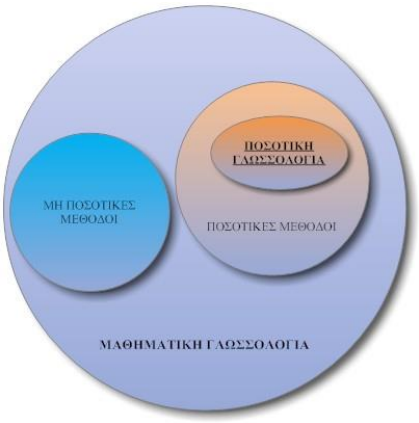
- **Γεωμετρικές** (geometric), τα στοιχεία που συγκρίνονται αναπαρίστανται ως διανύσματα και το μέτρο ομοιότητας που χρησιμοποιείται περισσότερο είναι η μέτρηση του συνημίτονου της γωνίας που σχηματίζουν μεταξύ τους τα διανύσματα (βλέπε ενότητα 2.2.1).
- **Βασισμένες σε ασαφή σύνολα** (fuzzy – set based), συγκρίνεται συνήθως ο αριθμός κοινής εμφάνισης συγκρινόμενων στοιχείων και συχνά χρησιμοποιούνται σχήματα απόδοσης βάρους.

- **Πιθανοτικές** (probabilistic), αφορά τη δύναμη της σχέσης μεταξύ των συγκρινόμενων στοιχείων και την κοινή πληροφορία που πιθανώς φέρουν, η οποία μεταφράζεται στην πιθανότητα τα δύο στοιχεία να συνυπάρχουν στην ίδια συλλογή.

3.3 Ποσοτική Γλωσσολογία

Η **συγκέντρωση δεδομένων, η επεξεργασία τους και η έκφραση τους με ποσοτικό τρόπο με σκοπό τη** διεξαγωγή συμπερασμάτων από αυτά, αποτελεί μια σημαντική εργασία στο πλαίσιο πολλών επιστημονικών κλάδων. Σύμφωνα με τον Johnson (2008) **για την επιστήμη της Γλωσσολογίας ειδικότερα αποτελεί πρακτική δεκαετιών** για πολλούς υπο-κλάδους της. Σε αυτούς συγκαταλέγεται και **η ΥΓ** όπου τα παραπάνω αποτελούν βασικό ένα κομμάτι εκπαίδευσης της.

Εκτός από την ΥΓ που έχει αναλυθεί έως τώρα, όπως ήδη έχει αναφερθεί (βλέπε ενότητα 1.3), από τη στροφή της Γλωσσολογίας προς την ποσοτική μελέτη έχει δημιουργηθεί ένας ξεχωριστός υπο-κλάδος, αυτός της ΠΓ. Μάλιστα οι Bolshakov και Gelbukh (2004) κατατάσσουν την ΠΓ (Quantitative Linguistics) ως **υπο-κλάδο της Μαθηματικής Γλωσσολογίας** (Mathematical Linguistics). Σύμφωνα με το Oxford dictionaries ως ΠΓ ορίζεται: «***Η συγκριτική μελέτη της συχνότητας και της κατανομής των λέξεων και των συντακτικών δομών σε διαφορετικά κείμενα***». Ακόμη στη βιβλιογραφία αναφέρεται πως (Liu και Huang 2012) η ΠΓ αφορά στα διάφορα γλωσσικά φαινόμενα, γλωσσικές δομές, δομικές ιδιότητες και τους μεταξύ τους συσχετισμούς στις δραστηριότητες επικοινωνίας στην πραγματική ζωή. Μέσω διαφόρων ποσοτικών τεχνικών η ΠΓ διεξάγει ακριβείς μετρήσεις, παρατηρήσεις, προσομοιώσεις, μοντελοποιήσεις και επεξήγηση αυτών των φαινομένων, με σκοπό την ανακάλυψη μαθηματικών νόμων που αφορούν τα γλωσσικά φαινόμενα. Ακόμη οι Bolshakov και Gelbukh (2004) συμπληρώνουν πως μέσω της ΠΓ **παρέχονται οι μέθοδοι για τη λήψη αποφάσεων στην επεξεργασία κειμένου** (text processing), **με βάση τα στατιστικά δεδομένα**. Τέλος, η Tesitelova (1992) συμπληρώνει πως η ΠΓ αφορά σε ποσοτικές μεθόδους που εφαρμόζονται στο πλαίσιο της Μαθηματικής Γλωσσολογίας. Σχηματικά η ΠΓ αναπαρίσταται στην εικ. 19:



Εικ. 19 ΠΓ ως υπο-κλάδος της Μαθηματικής Γλωσσολογίας

Σύμφωνα με τη βιβλιογραφία (Johnson 2008) οι **βασικοί στόχοι της ΠΓ** είναι οι εξής: **(α)** μετατροπή δεδομένων (μείωση των ποσοτικών δεδομένων στα ουσιαστικά τους τμήματα ώστε να περιγραφούν τάσεις και κοινά στοιχεία), **(β)** διεξαγωγή συμπερασμάτων (μέσω ελέγχου υποθέσεων), **(γ)** έρευνα και περιγραφή σχέσεων μεταξύ των δεδομένων και τέλος, **(δ)** εξερεύνηση και περιγραφή διαδικασιών που έχουν ως βάση τη θεωρία Πιθανοτήτων, όπως η θεωρία της Πληροφορίας.

3.3.1 Στατιστικοί νόμοι στην ΠΓ

Όπως αναφέρθηκε ήδη στην παραπάνω ενότητα, η ΠΓ περιγράφει τα διάφορα υπό μελέτη γλωσσικά φαινόμενα μέσω της μαθηματικής έκφρασης και πιο συγκεκριμένα μαθηματικών ή αλλιώς στατιστικών νόμων. Μάλιστα στην βιβλιογραφία (Hogan 2014) οι **νόμοι** αυτοί κατηγοριοποιούνται σε **δύο ομάδες στην ΠΓ**: στους νόμους **κατανομής (distributional)** που αφορούν στις κατανομές πιθανοτήτων με χαρακτηριστικό παράδειγμα το νόμο του **Zipf** και στους **λειτουργικοί (functional)** νόμους, οι οποίοι αφορούν στη σύνδεση μεταξύ γλωσσικών ιδιοτήτων, με χαρακτηριστικό παράδειγμα το νόμο **Menzerath – Altmann**.

Σε αυτό το σημείο κρίνεται απαραίτητη η επεξήγηση του ορισμού της έννοιας του «*νόμου*». Όπως αναφέρεται στη βιβλιογραφία (Hogan 2014) ο επιστημονικός **νόμος ουσιαστικά αποτελεί μια υπόθεση** η οποία ισχύει πάντοτε για τα αντικείμενα που αφορά, η οποία από την μια πλευρά **συνδέεται με άλλες υποθέσεις ενός επιστημονικού κλάδου και από την άλλη επιβεβαιώνεται μέσω εμπειρικών δεδομένων**. Μάλιστα στη βιβλιογραφία (Hogan 2014) ο επιστημονικός νόμος

αποτελεί τη **βάση για την κατασκευή μιας θεωρίας**, καθώς και αυτή με τη σειρά της αποτελείται από ένα σύνολο νόμων που συσχετίζονται μεταξύ τους, δηλαδή ένα **σύστημα νόμων**. Οι νόμοι χρησιμοποιούνται προκειμένου να υποστηρίξουν και να τεκμηριώσουν επιστημονικές εξηγήσεις. Τέλος, να επεξηγηθεί η διάκριση των νόμων και των θεωριών από τα **εμπειρικά δεδομένα.** Όπως αναφέρει ο Hempel (1964) οι εμπειρικές επιστήμες αφορούν στην **περιγραφή φαινομένων του κόσμου σύμφωνα με την εμπειρία και στην καθιέρωση γενικών αρχών ώστε τα φαινόμενα αυτά να μπορούν να επεξηγηθούν και να προβλεφθούν.**

Στη Γλωσσολογία **υπάρχει πληθώρα νόμων** σε σχέση με στατιστικές παρατηρήσεις και φαινόμενα που αφορούν τη γλώσσα. Παρακάτω θα αναλυθούν οι νόμοι των Menzerath – Altmann (MA) και Zipf καθώς αφορούν την παρούσα διατριβή. Ακόμη θα αναλυθεί η μεταξύ τους σχέση.

3.3.1.1 Νόμος των Menzerath-Altmann

Ένας λοιπόν από τους σημαντικότερους νόμους της ΠΓ είναι ο νόμος MA. Σύμφωνα με τη βιβλιογραφία (Buk και Rovenchak 2008), ο παραπάνω νόμος αφορά στην **περιγραφή της σχέσης μεταξύ δομήματος (construct) και συστατικού (constituent), όπου όσο μεγαλώνει το πρώτο, τόσο μικραίνει το δεύτερο**. Το 1928 ο Paul Menzerath, πειραματικός ψυχολόγος και φωνολόγος, παρατήρησε ότι όσο ο μέσος όρος του μήκους μιας συλλαβής μειώνονταν τόσο ο αριθμός των συλλαβών που απαρτίζουν τη λέξη αυξανόταν. Σύμφωνα με τον Eroglu (2013) ο Menzerath, ο οποίος θεωρείται ένας από τους πρωτοπόρους ερευνητές στην ΠΓ πρώτος ανέδειξε την **αρνητική συσχέτιση στη σχέση μεταξύ του μήκους ενός γλωσσικού δομήματος με τα συστατικά του.** Σύμφωνα με τη βιβλιογραφία (Buk και Rovenchak 2008) αργότερα ο Gabriel Altmann εξέφρασε με μαθηματική μορφή την παραπάνω σχέση μεταξύ δομήματος και συστατικού του Menzerath έτσι ώστε σήμερα ο παραπάνω νόμος είναι γνωστός ως νόμος των MA. Παρότι ο Menzerath αρχικά αναφέρθηκε στη σχέση μεταξύ συλλαβής - λέξης είναι ξεκάθαρο ότι ο παραπάνω νόμος μπορεί να εφαρμοστεί σε διάφορα επίπεδα γλωσσικών μονάδων αρκεί να περιγράφει μια σχέση δομήματος - συστατικού, σύμφωνα με τη βιβλιογραφία (Mikros και Milicka 2014).

Σύμφωνα με τον Eroglu (2013), παρότι ο νόμος MA αποτελεί έναν από τους βασικότερους νόμους της ΠΓ, η χρήση του δεν αφορά μόνο στην επιστήμη της

Γλωσσολογίας αλλά **βρίσκει εφαρμογή σχεδόν σε οποιαδήποτε περιγραφή οργανωτικής δομής** σε πληθώρα επιστημονικών πεδίων όπως για την περιγραφή οργανωτικής δομής μουσικών κειμένων. Μάλιστα οι Baixeries et al. (2013) αναφέρουν σε άρθρο τους τη χρήση του νόμου ΜΑ για την περιγραφή της σχέσης μεταξύ γονιδιωμάτων και χρωμοσωμάτων, όπου παρατήρησαν την ίδια αρνητική συσχέτιση, δηλαδή πως όσο μεγαλύτερο είναι το γονιδίωμα τόσο μικρότερα είναι τα χρωμοσώματα (σε ζεύγη βάσεως).

Όπως αναφέρει ο Altmann (1980), ένα βασικό ζήτημα στο νόμο ΜΑ αποτελεί το τι θεωρείται συστατικό. Δηλαδή **ποιες μονάδες θεωρούνται συστατικά**, οι αμέσως επόμενες στην ιεραρχία ή και άλλες, προερχόμενες από μετέπειτα επίπεδα ιεραρχίας. Με βάση αυτό το ζήτημα, προβλήματα στην εφαρμογή της μαθηματικής έκφρασης του νόμου των ΜΑ μπορούν να παρατηρηθούν για γλώσσες που περιέχουν μη συλλαβικές λέξεις.

Πέραν του θεωρητικού μέρους του νόμου ΜΑ που αναπτύχθηκε παραπάνω, θα παρατεθούν και οι μαθηματικές εκφράσεις του νόμου. Σύμφωνα με τον Altmann (1980), στην παρακάτω εξίσωση εκφράζεται ένας «***σταθερός ρυθμός μείωσης του μήκους του συστατικού y***» μέσω του τύπου (13):

$$\frac{y'}{y} = -c \tag{13}$$

Μέσω της μαθηματικής **ολοκλήρωσης** (integration), ο παραπάνω μαθηματικός τύπος παίρνει τη μορφή του (14):

$$y = Ae^{-cx} \tag{14}$$

Όπου το **x αντιστοιχεί στο δόμημα** και το **c σε μια μεταβλητή.** Έτσι μέσω της εξίσωσης (14), παρουσιάζεται, όπως αναφέρεται στη βιβλιογραφία (Mikros και Milick 2014), η σχέση του μήκους δομήματος x και μήκους συστατικού y, η οποία περιγράφεται από μια **μονότονη φθίνουσα συνάρτηση** (monotonic decreasing function).

Στην παρακάτω διαφορική εξίσωση από τη βιβλιογραφία (Kulacka και Macutek 2007), ορίζεται ξανά η σχέση μεταξύ συστατικού και δομήματος. Όπως φαίνεται από το μαθηματικό τύπο (15) το συστατικό y'είναι ανάλογο του μέσου μήκους y και αντιστρόφως ανάλογο του μήκους x.

$$y' = \frac{cy}{x} \tag{15}$$

Σύμφωνα με τον Eroglu (2013), *«ο νόμος του ΜΑ αποτελεί ένα συνεχόμενο μοντέλο κατανομής πιθανοτήτων (probability distribution model), το οποίο χρησιμοποιείται για να περιγράψει την πιθανοτική σχέση μεταξύ των διακριτών αποτελεσμάτων των ποσοτήτων»* όπως φαίνεται στο μαθηματικό τύπο (16):

$$y(x \mid A,b,c) = Ax^{b}e^{-cx} \tag{16}$$

Ξανά, το y αντιστοιχεί σε συστατικό, το x σε δόμημα και οι A, b, c αναπαριστούν ελεύθερες παραμέτρους.

3.3.1.2 Νόμος του Zipf

Σύμφωνα με τον Wyllys (1981) ο νόμος του Zipf περιγράφει τη σχέση μεταξύ της συχνότητας εμφάνισης λέξεων σε ένα σώμα κειμένου και της κατάταξης τους, η οποία περιγράφηκε από τον George Kingsley Zipf, καθηγητή φιλολογίας και γλωσσολόγος στο Harvard University. Το 1935 στο βιβλίο του "*The Psycho-Biology of Language*", ο Zipf περιγράφει την αλγεβρική μαθηματική έκφραση, γνωστή αργότερα ως νόμο του Zipf. Στην επιστήμη των μαθηματικών ο νόμος του Zipf κατατάσσεται στους νόμους των δυνάμεων (power laws). Με βάση το Oxford Dictionary (2015) ως power law ορίζεται «*μια σχέση μεταξύ δύο ποσοτήτων έτσι ώστε η μια είναι ανάλογη σε μια σταθερή δύναμη της άλλης*». Σύμφωνα με τη βιβλιογραφία (Powers 1998) ο νόμος του Zipf βρίσκει εφαρμογή σε πληθώρα επιστημονικών κλάδων που ασχολούνται με τη φυσική γλώσσα όπως αυτών της Γλωσσολογίας και ειδικότερα της ΠΓ και της Υπολογιστικής Ψυχογλωσσολογίας. Παρατηρείται στη διεθνή βιβλιογραφία (Piantadosi 2014) ότι ο νόμος του Zipf βρίσκει εφαρμογή εκτός από τη φυσική γλώσσα και σε πληθώρα άλλων επιστημών όπως στη μουσική, στα υπολογιστικά συστήματα, στο διαδίκτυο και στα φυσικά και βιολογικά συστήματα.

Σύμφωνα με τους Manning και Schutze (1999) ο νόμος του Zipf μπορεί να χρησιμοποιηθεί ως μια γενικευμένη περιγραφή της κατανομής συχνότητας των λέξεων στις ανθρώπινες γλώσσες και η περιγραφή αυτή διαχωρίζει τη συχνότητα εμφάνισης των λέξεων σε χαμηλή, μεσαία και υψηλή συχνότητα εμφάνισης λέξεων. Οι Manning και Schutze (1999) αναφέρουν πως **με το νόμο του Zipf μπορεί να ερευνηθεί η σχέση μεταξύ της συχνότητας εμφάνισης μιας λέξης f σε ένα σώμα κειμένου και της θέσης κατάταξης r της λέξης αυτής**, αφού έχει ήδη μετρηθεί η συχνότητα εμφάνισης για όλες τις λέξεις (ως τύπος, βλέπε ενότητα 2.4.1.1) του

σώματος κειμένου και έχει συνταχθεί μια λίστα σε φθίνουσα σειρά ξεκινώντας με τη λέξη με τη συχνότερη εμφάνιση. Μάλιστα τονίζουν ότι χρησιμοποιώντας το νόμο του Zipf τα δεδομένα στη γραφική παράσταση σχετικά με τη χρήση των περισσότερων λέξεων θα είναι συνήθως εξαιρετικά αραιά (sparse).

Στην εργασία του Wyllys (1981) περιγράφεται ο νόμος του Zipf αναλυτικότερα. Για να επεξηγηθεί, υποτίθεται η ύπαρξη ενός σώματος κειμένου σε φυσική γλώσσα και όπως και παραπάνω μετριέται η συχνότητα εμφάνισης των λέξεων σε αυτό ώστε να σχηματιστεί η λίστα κατάταξης των λέξεων σε φθίνουσα σειρά με την πιο συχνή λέξη να έχει την πρώτη θέση. Ως νόμος του Zipf ορίζεται η παρακάτω μαθηματική έκφραση (17):

$$r \cdot f = c \tag{17}$$

όπου το *r* αντιστοιχεί στην κατάταξη μιας λέξης, το *f* στη συχνότητα εμφάνισης της λέξης και *c* αντιστοιχεί σε μια σταθερά, η οποία εξαρτάται από το σώμα κειμένου.

Σύμφωνα με τη βιβλιογραφία (Sorell 2012), τοποθετώντας τις λέξεις ενός σώματος κειμένου σε φθίνουσα σειρά ξεκινώντας από αυτήν με τη μεγαλύτερη συχνότητα, τότε η δεύτερη συχνότερη λέξη θα εμφανίζεται περίπου τις μισές φορές από ότι η πρώτη και η τρίτη συχνότερη λέξη περίπου 1/3 φορές από ότι η πρώτη κ.ο.κ. Έτσι πολλαπλασιάζοντας την κατάταξη *r* με τη συχνότητα *f* της κάθε λέξης, η σταθερά *c* θα πρέπει να παραμένει περίπου ίδια για κάθε λέξη. Υποδεικνύει λοιπόν μια αντιστρόφως ανάλογη σχέση μεταξύ κατάταξης και συχνότητας των λέξεων ενός κειμένου.

Στην εργασία του Wyllys (1981) ο νόμος του Zipf εκτός από την παραπάνω αλγεβρική του έκφραση περιγράφεται επίσης ως ισοδύναμος με τη γραφική αναπαράσταση:

$$\log r \cdot \log f = \log c \tag{18}$$

Όπου στη σχεδίαση των ζευγών σημείων που προκύπτουν, ο λογάριθμος της κατάταξης *r* τοποθετείται στον οριζόντιο άξονα και ο λογάριθμος της συχνότητας *f* στον κάθετο άξονα και έτσι τα σημεία σχηματίζουν μια ελαφρά καμπύλη γραμμή, γνωστή ως καμπύλη Zipf (Zipf's curves).

Ο Wyllys (1981) αναφέρει ότι ο υπολογισμός με βάση το νόμο του Zipf έχει πιο έγκυρα αποτελέσματα κυρίως όσον αφορά την κατάταξη λέξεων με μεσαία τάξη εμφάνισης παρά για τις λέξεις με πολύ υψηλή ή χαμηλή συχνότητα εμφάνισης. Ακόμη αναφέρει ότι η εργασία του Zipf δείχνει πως το μέγεθος του δείγματος θα

πρέπει να αποτελείται από τουλάχιστον 5000 λέξεις ώστε το $r \times f$ να είναι σταθερό, ακόμη και για τις μεσαίες κατατάξεις.

Σύμφωνα με τον Altmann (2002) ο νόμος του Zipf εστιάζει στις σχέσεις μεταξύ των διαφορετικών οντοτήτων της γλώσσας και αναφέρει πολλά διαφορετικά είδη σχέσεων που έχουν παρατηρηθεί μεταξύ των οντοτήτων αυτών. Στη βιβλιογραφία (Hrebicek 2002) οι οντότητες αυτές διακρίνονται για το νόμο του Zipf, σε λεξιλογικές μονάδες και σώματα κειμένου, τα οποία στο πλαίσιο μιας συγκεκριμένης δομής κειμένου έχουν αμοιβαίες σχέσεις μεταξύ τους. Οι **σχέσεις αυτές που αναπτύσσονται μεταξύ των οντοτήτων έχουν διάφορες ιδιότητες** εκ των οποίων οι δύο παρακάτω, όπως αναφέρονται στη βιβλιογραφία (Hrebicek 2002):

1. «*Κάθε παρατηρούμενη συνεχής ακολουθία ήχου μιας γλώσσας λειτουργεί ως φορέας της μη συνεχούς ακολουθίας κωδικών συμβόλων διαφορετικών γλωσσικών επιπέδων*».
2. «*Οι μονάδες διαφορετικών επιπέδων μπορούν να περιγραφούν ως σύνολα που διακρίνονται από αυτοομοιότητα*» (δηλαδή την ιδιότητα ενός σχήματος να είναι όμοιο με ένα ή περισσότερα τμήματά του).

Οι παραπάνω ιδιότητες προκύπτουν από την ιεραρχική σχέση μεταξύ των οντοτήτων, όπως αναφέρει και ο Altmann (2002), δηλαδή τις σχέσεις ανάμεσα σε διαφορετικά επίπεδα, **οι οποίες εκφράζονται κυρίως μέσω του νόμου ΜΑ**. Με βάση το νόμο ΜΑ λοιπόν μπορούμε να δούμε το κείμενο ως μια δομούμενη γλωσσική μονάδα, όπου η βασική δομική της μονάδα είναι ο κώδικας της γλώσσας. Συνεπώς διαφαίνεται ήδη **η σχέση μεταξύ των δύο νόμων, ΜΑ και Zipf**, καθώς οι δύο παραπάνω ιδιότητες προκύπτουν μέσω του νόμου των ΜΑ και αποτελούν βασικές ιδιότητες που διέπουν τις σχέσεις μεταξύ των γλωσσικών μονάδων, οι οποίες αφορούν κυρίως το νόμο του Zipf. Η σχέση μεταξύ των δύο νόμων αποδεικνύεται μαθηματικά μέσω μιας υποπερίπτωσης της **εξίσωσης (16)** (βλέπε ενότητα 3.3.1.1), όπου **περιγράφεται ο νόμος ΜΑ.** Ειδικότερα μέσω της **συγκεκριμένης υποπερίπτωσης** του, όπως αναφέρεται σε Eroglu (2013), όταν b≠0 και c=0 τότε κανείς οδηγείται **στην εξίσωση του νόμου Zipf**, όπως φαίνεται παρακάτω στον τύπο (19):

$$y(x \mid A,b) = Ax^{-b} \tag{19}$$

Συνεπώς κάτω από συγκεκριμένες συνθήκες ο νόμος του Zipf, αποτελεί υποπερίπτωση του νόμου ΜΑ.

3.3.2 Θεωρία της Πληροφορίας και ΠΓ

Όπως ήδη έχει αναφερθεί παραπάνω (βλέπε ενότητα 1.3) παρατηρείται η **εστίαση της ενασχόλησης των μεθόδων ΠΓ** και των τάσεων στην έρευνα της ΠΓ σχετικά με την εφαρμογή διαφόρων **μαθηματικών και υπολογιστικών θεωριών**, όπως η θεωρία Πληροφορίας και η θεωρία Πιθανοτήτων. Για το λόγο αυτό σε αυτή την ενότητα θα παρουσιαστεί **η θεωρία της Πληροφορίας του Shannon** C. E. και η σύνδεσή της με το σχήμα απόδοσης βάρους tf – idf (βλέπε ενότητα 2.2.2) καθώς και σύνδεση της με τη θεωρία Πιθανοτήτων (η οποία θα αναλυθεί στην συνέχεια στην ενότητα 4.1.3).

Όπως αναφέρεται στη βιβλιογραφία (Nadel 2005) ο Shannon θεωρείται ο θεμελιωτής της θεωρίας της Πληροφορίας, η οποία μπορεί να χαρακτηριστεί ως μια **μαθηματική θεωρία για την επικοινωνία**. Όρισε στο πλαίσιο της διαδικασίας επικοινωνίας ένα σύστημα αποτελούμενο από έναν δέκτη, ένα κανάλι μεταφοράς και έναν αποδέκτη. Ο σκοπός του συστήματος αυτού είναι η μεταφορά κάποιας πληροφορίας σε μορφή μηνύματος. Πιο συγκεκριμένα **κάθε μήνυμα περιέχει μια ποσότητα πληροφορίας**, με την οποία ασχολήθηκε ο Shannon, ο οποίος εισήγαγε τη **μέτρηση τόσο της ποσότητας της πληροφορίας αυτής, όσο και της χωρητικότητας του καναλιού μετάδοσης**.

Όπως αναφέρεται στη βιβλιογραφία (Nadel 2005), ο ορισμός της πληροφορίας του Shannon είναι αρκετά τεχνικός και προσεγγίζει την ποσότητα της πληροφορίας συνδυάζοντας τη θεωρία Πιθανοτήτων. Πιο συγκεκριμένα, θεωρεί την πληροφορία αυτή ως ένα ενδεχόμενο δειγματικού χώρου για το οποίο υπάρχει αβεβαιότητα εάν θα συμβεί ή όχι και η αβεβαιότητα μειώνεται μονάχα με την παρατήρηση πραγματοποίησης του. Τότε μπορεί κανείς να είναι σίγουρος για το αποτέλεσμα της πιθανότητας του. Έτσι η ποσότητα κατά την οποία μειώνεται η αβεβαιότητα για το πιθανό αποτέλεσμα αποτελεί την περιεχόμενη πληροφορία για ένα ενδεχόμενο. Αντιστοίχησε δηλαδή την πραγματοποίηση ενός ενδεχομένου με μια πιθανότητα πραγματοποίησης του. Όπως π.χ. κατά την ρίψη ενός ζαριού, όπου θα αντιστοιχεί η πιθανότητα να ισχύει μια από έξι πιθανές τιμές και τότε η πιθανότητα που αντιστοιχεί στην πληροφορία είναι ίση με 1/6.

Στη βιβλιογραφία (Nadel 2005) αναφέρεται ακόμη πως **η πληροφορία μεταφέρεται από σύμβολα**, άρα στα σύμβολα αναλογεί η πιθανότητα εμφάνισης. Ο Shannon ασχολήθηκε αρχικά με την αντιστοίχιση ενός συμβόλου, η οποία όμως μπορεί να επεκταθεί και σε μια ροή συμβόλων, όπως π.χ. είναι μια λέξη. Υπολόγισε

το μέσο όρο πληροφορίας που δύναται να μεταφερθεί ανά σύμβολο σε μια ροή συμβόλων, ποσότητα που ονόμασε ως **εντροπία**.

Τα δύο βασικά προβλήματα της θεωρίας Πληροφορίας, σύμφωνα με τη βιβλιογραφία (Nadel 2005) είναι η **αποδοτικότητα μετάδοσης** (συμπίεση δεδομένων) και η **αξιοπιστία μετάδοσης όταν το κανάλι έχει θόρυβο**, ο οποίος συνδέεται με τον υπολογισμό της εντροπίας.

Στη βιβλιογραφία (Robertson 2004) αναφέρεται ακόμη πως **η θεωρία του Shannon μπορεί να συνδεθεί με το συστατικό *idf***, το οποίο χρησιμοποιείται κυρίως στο σχήμα απόδοσης βαρών *tf – idf* (βλέπε ενότητα 2.2.2). Το *idf* συστατικό ορίζεται από τον παρακάτω μαθηματικό τύπο (20):

$$idf(t_i) = \log \frac{N}{n_i} \tag{20}$$

Όπου το *idf* υποδηλώνει τη σπανιότητα ενός όρου σε μια συλλογή. Παρατηρώντας το λόγο του λογαρίθμου μπορεί κανείς να παρατηρήσει το λόγο για τον υπολογισμό μιας πιθανότητας (βλέπε ενότητα 4.1.3), αλλά αντεστραμμένο. Έτσι θα μπορούσε να υπολογιστεί η πιθανότητα για ένα τυχαίο έγγραφο να περιέχει έναν όρο t_i, από τον μαθηματικό τύπο (21):

$$P(t_i) = \frac{n_i}{N} \tag{21}$$

Έτσι μέσω των τύπων (20) και (21), είναι εφικτή η σύνδεση του *idf* συστατικού με τη θεωρία της Πληροφορίας του Shannon, η οποία είναι εμφανής στον παρακάτω μαθηματικό τύπο (22):

$$idf(t_i) = -\log P(t_i) \tag{22}$$

Να σημειωθεί πως σύμφωνα με τον Nadel (2005) για τη θεωρία της Πληροφορίας η ποσότητα που αντιστοιχεί στο $-\log P(t_i)$, αποτελεί την ποσότητα πληροφορίας στη μετάδοση μηνύματος, η οποία παίρνει μέρος στη συνάρτηση υπολογισμού της εντροπίας. Τα δε μηνύματα μετάδοσης πληροφορίας του Shannon μπορούν κάλλιστα να αντιστοιχηθούν σε μεταβλητές, που αντιστοιχούν σε ένα δειγματικό χώρο (με όλα τα πιθανά μηνύματα), σε μια συνάρτηση υπολογισμού για την πιθανότητα και σε κάποιο μέτρο πιθανότητας.

ΚΕΦΑΛΑΙΟ 4ο

ΣΤΑΤΙΣΤΙΚΑ ΘΕΜΑΤΑ

4.1 Εισαγωγή σε στατιστικά θέματα

Η Στατιστική επιστήμη θεωρείται **συναφής για οποιαδήποτε επιστημονική διερεύνηση** και χρησιμοποιείται εκτεταμένα από πολλούς επιστημονικούς κλάδους πέραν των μαθηματικών για την παρατήρηση φαινομένων και την διεξαγωγή συμπερασμάτων. Μάλιστα στη βιβλιογραφία (Diggle και Chetwynd 2011) αναφέρεται ότι η **επιστημονική μέθοδος συνίσταται από τη θεωρία**, η οποία προβλέπει τις διάφορες διαδικασίες και έτσι παρέχει αντίστοιχα μοντέλα για αυτές **και την παρατήρηση**, η οποία επιβεβαιώνει ή όχι την ορθότητα μιας θεωρίας μέσω της διεξαγωγής πειραμάτων είτε μέσω της άμεσης παρατήρησης. Έτσι η επιστημονική μέθοδος είναι άμεσα συνυφασμένη με τη συμπερασματολογία που διέπει τη **Στατιστική** και ειδικότερα **μπορεί να θεωρηθεί ως ένα εργαλείο σύνδεσης μεταξύ των πραγματικών δεδομένων και των μοντέλων της θεωρίας.**

Στο κεφάλαιο αυτό θα πραγματοποιηθεί μια **εισαγωγή** πάνω στον κλάδο της **Στατιστικής**, ώστε να επεξηγηθούν στη συνέχεια τα απαραίτητα **στατιστικά εργαλεία που χρησιμοποιήθηκαν** στην παρούσα διατριβή για την ανάπτυξη του προτεινόμενου μοντέλου (βλέπε 5ο κεφάλαιο) και ειδικότερα για την επιβεβαίωση της επιστημονικής θεωρίας και παρατηρήσεων μέσω των πειραμάτων που διεξήχθησαν.

Σύμφωνα με τον Fisher (1950) η Στατιστική αποτελεί κλάδο των Εφαρμοσμένων Μαθηματικών, στον οποίο εφαρμόζεται η γνώση των μαθηματικών σε δεδομένα που αποτελούν παρατηρήσεις σχετικές με έναν *πληθυσμό* (επεξηγείται παρακάτω). Ο κλάδος της Στατιστικής περιλαμβάνει την θεωρία και τις τεχνικές ώστε να συλλεχθεί, οργανωθεί, παρουσιαστεί, αναλυθεί και ερμηνευθεί ένα σύνολο δεδομένων με σκοπό τον καθορισμό των βασικών χαρακτηριστικών του.

Σύμφωνα με τη βιβλιογραφία (Fisher 1950, Αδαμόπουλος, Δαμιανός και Σβέρκος 2014) η **Στατιστική** μπορεί να χωριστεί **σε τρείς τομείς**: (α) **στον Σχεδιασμό Πειραμάτων (Experimental Design)**, που αφορά στην συλλογή των δεδομένων, (β) στην **Περιγραφική Στατιστική (Descriptive Statistics)** μέσω της οποίας επεξεργάζονται και παρουσιάζονται τα δεδομένα και στην (γ) **Επαγωγική Στατιστική ή Στατιστική Συμπερασματολογία (Inferential Statistics)** όπου μέσω της διερεύνησης των δεδομένων ο ερευνητής οδηγείται σε συμπεράσματα. Ο Panik (2012) συμπληρώνει ειδικότερα για την Περιγραφική Στατιστική πως αφορά στις τεχνικές συμπύκνωσης δεδομένων και παρουσίασης των κατανομών συχνοτήτων

τους μέσω πινάκων και διαγραμμάτων, ενώ για την Επαγωγική Στατιστική αναφέρει πως εκτός της διεξαγωγής συμπερασμάτων αφορά και τη λήψη αποφάσεων.

Οι παραπάνω τομείς της Στατιστικής αντιστοιχούν στα απαραίτητα βήματα για τη **διεξαγωγή μιας ολοκληρωμένης στατιστικής μελέτης** όπως την παρουσιάζουν και οι Τσίμπος και Γεωργιακώδης (1999): **(α)** η λήψη του δείγματος, **(β)** η επεξεργασία δεδομένων για το δείγμα, μέσω στατιστικών μέτρων για τον καθορισμό παραμέτρων του πληθυσμού και τέλος, **(γ)** η διεξαγωγή των συμπερασμάτων.

4.1.1 Σχεδιασμός Πειραμάτων

Η Στατιστική αφορά το σχεδιασμό κάποιου πειράματος που διεξάγεται από έναν ερευνητή. Χρησιμοποιείται η **θεωρία Συνόλων** ώστε να καθοριστεί το πεδίο εφαρμογής του πειράματος. Ακόμη χρησιμοποιείται η θεωρία Πιθανοτήτων μέσω της οποίας μπορούν να ερμηνευθούν τα διάφορα πειράματα και ειδικότερα τα στοχαστικά πειράματα όπως αναφέρουν και οι Δαμιανού, Παπαδάτος και Χαραλαμπίδης (2003).

Σύμφωνα με Μπούτσικα (2003) τα πειράματα χωρίζονται σε δύο κατηγορίες:

- **Αιτιοκρατικά** (deterministic), όπου οι γνωστές *μεταβλητές* (η έννοια επεξηγείται στην περιγραφική στατιστική) επαρκούν για την πρόβλεψη των αποτελεσμάτων του πειράματος.
- **Στοχαστικά** (stochastic / probabilistic), όπου οι γνωστές μεταβλητές δεν επαρκούν για την πρόβλεψη των αποτελεσμάτων του πειράματος.

Περισσότερη ανάλυση σε σχέση με τη θεωρία Συνόλων και Πιθανοτήτων θα δοθεί στην ενότητα 4.1.3 καθώς απαιτείται εξοικείωση με ορισμούς που θα αναπτυχθούν στο πλαίσιο της Περιγραφικής Στατιστικής στην αμέσως επόμενη ενότητα.

4.1.2 Περιγραφική Στατιστική

4.1.2.1 Βασικοί ορισμοί – δεδομένα - μεταβλητές

Προτού παρουσιαστούν οι επιμέρους ενότητες που αφορούν την Περιγραφική Στατιστική θα γίνει μια σύντομη εισαγωγή σε βασικούς ορισμούς: σύμφωνα με τους Τσίμπος και Γεωργιακώδης (1999) ο υπό μελέτη **πληθυσμός** αναφέρεται στο σύνολο των δεδομένων μιας μελέτης ενώ το **δείγμα** αφορά μέρος του πληθυσμού, ώστε να μη χρειαστεί η εξονυχιστική ανάλυση ολόκληρου του πληθυσμού.

Τα διαφορετικά είδη δεδομένων υπό μελέτη είναι πολύ σημαντικό ζήτημα ώστε να επιλεγεί η κατάλληλη στατιστική προσέγγιση. Έτσι, τα είδη των δεδομένων με βάση τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012) χωρίζονται σε δύο κατηγορίες ανάλογα με την τιμή που μπορούν να πάρουν:

- **Αριθμητικά/Ποσοτικά**: Όταν η πληροφορία που προκύπτει από τα δεδομένα έχει αριθμητική τιμή. Αυτά διακρίνονται στις εξής υπο-κατηγορίες:
    - **Συνεχή**: οι αριθμητικές τιμές εκφράζουν κάποιο διάστημα τιμών, όπως π.χ. έναν δεκαδικό αριθμό.
    - **Ασυνεχή/Διακριτά**: οι αριθμητικές τιμές αφορούν ακέραιους αριθμούς.
- **Ονομαστικά/Ποιοτικά:** όταν η πληροφορία που προκύπτει από τα δεδομένα έχει ως τιμή κάποια λέξη που την περιγράφει. Όπως είναι κατανοητό στα ποιοτικά δεδομένα δεν μπορούν να εφαρμοστούν μαθηματικές πράξεις, αλλά μπορούν οι ποιοτικές τους τιμές αυθαίρετα να αντικατασταθούν από αριθμητικές για τη διευκόλυνση της διεξαγωγής της στατιστικής μελέτης.

Γενικότερα σύμφωνα με τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012) οι τιμές των δεδομένων και των μεταβλητών (που αναλύονται στη συνέχεια) μπορούν να υπαχθούν σε τέσσερις κλίμακες μέτρησης ξεκινώντας από τη χαμηλότερη προς την υψηλότερη βαθμίδα. Η χαμηλότερη είναι η (α) ονομαστική (nominal), έπειτα ακολουθεί η (β) τακτική (ordinal), στη συνέχεια η (γ) διαστημική (interval) και τέλος, η (δ) αναλογική (ratio). Ειδικότερα ο Panik (2012) αναλύει τις παραπάνω κλίμακες με κάθε κλίμακα να φέρει ένα νέο χαρακτηριστικό το οποίο προστίθεται στην αμέσως επόμενη κλίμακα μαζί με το νέο. Έτσι μέσω της (α) **ονομαστικής κλίμακας** είναι εφικτός μονάχα ο προσδιορισμός κατηγοριών. Μέσω της (β) **τακτικής** αναφέρεται η δυνατότητα διάταξης και κατάταξης των στοιχείων. Μέσω της (γ) **διαστημικής** κλίμακας εντάσσεται η ιδιότητα καθορισμού της απόστασης μεταξύ των μετρήσεων και τέλος μέσω της (δ) **αναλογικής** προστίθεται μια μηδενική κατάσταση, μέσω της οποίας είναι εφικτές οι συγκρίσεις μεταξύ των μετρήσεων. Ανάλογα λοιπόν το είδος των δεδομένων και των μεταβλητών καθορίζονται και οι στατιστικές μέθοδοι που ακολουθούνται.

Έως τώρα έχει γίνει αναφορά στα δεδομένα ενός πληθυσμού, τα οποία συλλέγονται και επεξεργάζονται με κατάλληλο τρόπο. Η εξέταση όμως αυτή πραγματοποιείται ως προς κάποιο **χαρακτηριστικό ή κάποια ιδιότητα για το δείγμα**. Σύμφωνα με τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012)

το σύνολο των χαρακτηριστικών αυτών ονομάζονται **πληθυσμιακά χαρακτηριστικά**. Οι δε τιμές που παίρνουν τα πληθυσμιακά χαρακτηριστικά αποτελούν τιμές των δεδομένων ενώ το **κάθε πληθυσμιακό χαρακτηριστικό προς εξέταση ονομάζεται μεταβλητή**, η οποία με βάση τις μετρήσεις που πραγματοποιούνται **έχει και αυτή τιμές**, που μπορούν πάλι να υπάγονται σε μια από τις **κατηγορίες των ειδών δεδομένων** που αναπτύχθηκαν παραπάνω.

Σύμφωνα με τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012), μια μεταβλητή γράφεται με κεφαλαίους λατινικούς χαρακτήρες όπως π.χ. X. Αυτή μετράται για κάθε στοιχείο i, το οποίο ξεκινά από το πρώτο στοιχείο μέχρι το τελευταίο του πληθυσμού, με μέγεθος πληθυσμού $N = (i = 1, 2, \dots, N)$ ενώ όταν αφορά εξέταση δείγματος πληθυσμού (δηλαδή εξέταση τμήματος του πληθυσμού) αντιστοιχεί σε μέγεθος δείγματος $n = (i = 1, 2, \dots, n)$. Έτσι η εξεταζόμενη μεταβλητή για κάποιο στοιχείο i του πληθυσμού παίρνει τη μορφή x_i. Σύμφωνα με τον Panik (2012), στον παρακάτω τύπο (23) εκφράζεται το άθροισμα όλων των τιμών της μεταβλητής για το δείγμα του πληθυσμού:

$$X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i \tag{23}$$

Ακόμη όπως αναφέρεται στη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012), οι μεταβλητές διακρίνονται σε:

- **Εξαρτημένες και ανεξάρτητες** ανάλογα με το αν εξαρτώνται ή όχι από τις τιμές άλλων μεταβλητών.
- **Τυχαίες/Στοχαστικές και μη τυχαίες/προσδιορίσιμες** ανάλογα με το αν μπορούν να προσδιοριστούν εκ των προτέρων. Έτσι οι τιμές των πρώτων διέπονται από την θεωρία των Πιθανοτήτων προκειμένου να προσδιοριστούν, ενώ οι δεύτερες προσδιορίζονται εκ των προτέρων.

4.1.2.2 Κατανομή συχνότητας και παρουσίαση δεδομένων

Η καταμέτρηση των συχνοτήτων εμφάνισης των **μεταβλητών** είναι πολύ σημαντική για την ερμηνεία πειραμάτων και φαινομένων. Έτσι όπως αναφέρουν και οι Τσίμπος και Γεωργιακώδης (1999), οι συχνότητες (frequencies) είναι αριθμητικές τιμές που εκφράζουν την καταμέτρηση της συχνότητας εμφάνισης τιμών που λαμβάνουν οι μεταβλητές και συμβολίζονται ως f_i, όπου το $i = 1, 2, \dots, k$ και αποτελεί κάποια κατηγορία της μεταβλητής υπό εξέταση. Οι συχνότητες μπορούν να

εκφραστούν σε ποσοστιαία ή αναλογική κατανομή και τότε ονομάζονται σχετικές. Μάλιστα σύμφωνα με τους Τσίμπος και Γεωργιακώδης (1999) «*επειδή οι σχετικές συχνότητες δηλώνουν την κατ' αναλογία συχνότητα εμφάνισης κάποιου πληθυσμιακού χαρακτηριστικού στο σύνολο των εκ παρατηρήσεως μετρήσεων, μπορούμε να θεωρήσουμε ότι επέχουν* ***θέση πιθανοτήτων****»*. Όπως συμπληρώνουν οι Αγγελής και Δημάκη (2011) η κατανομή πιθανότητας (probability distribution) μιας τυχαίας μεταβλητής δίνει πληροφορίες για την πιθανότητα εμφάνισης των δυνατών τιμών της. Ειδικότερα σύμφωνα με τη βιβλιογραφία (Everitt 2006) η κατανομή πιθανότητας μιας διακριτής τυχαίας μεταβλητής είναι ουσιαστικά ένας μαθηματικός τύπος που έχει ως αποτέλεσμα την πιθανότητα για κάθε τιμή της μεταβλητής. Ειδικότερα, το μέτρο πιθανότητας όπως αναφέρεται τοποθετείται μεταξύ του διαστήματος των τιμών 0 έως 1, με την τιμή 1 να αντιστοιχεί σε ένα βέβαιο να συμβεί ενδεχόμενο. Αντίστοιχα όταν η τυχαία μεταβλητή είναι συνεχής ο μαθηματικός τύπος καθορίζει ένα διάστημα τιμών και όσον αφορά τη γραφική απεικόνιση του αποτελεί μια καμπύλη. Ένα παράδειγμα κατανομής πιθανότητας αποτελεί η κανονική κατανομή. Άλλος όρος που μπορεί να χρησιμοποιηθεί είναι η πυκνότητα πιθανότητας (probability density). Για την περιγραφή της τυχαίας μεταβλητής απαιτείται η χρήση μέτρων κεντρικής τάσης τα οποία αναφέρονται παρακάτω.

Ο όρος αθροιστική συχνότητα είναι επίσης πολύ σημαντικός στη Στατιστική, για την ανάλυση του όμως πρώτα θα χρειαστεί η επεξήγηση του όρου της κλάσης. Τα ποσοτικά δεδομένα, σύμφωνα με τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012), εντάσσονται είτε σε διαστημική είτε σε αναλογική κλίμακα και η κατάταξη τους πραγματοποιείται μέσω **δημιουργίας κλάσεων**. Για να οριστούν οι κλάσεις, πρέπει να υπολογιστεί το εύρος (range) της μεταβλητής, το οποίο αποτελεί τη διαφορά μεταξύ της μέγιστης και ελάχιστης τιμής που παίρνει η μεταβλητή. Στη συνέχεια από τον παρακάτω τύπο (24), μπορεί να υπολογιστεί το πλάτος των διαστημάτων κλάσεων δ όπως ονομάζεται, όπου R αντιστοιχεί στο εύρος της μεταβλητής και k στον επιθυμητό αριθμό κλάσεων:

$$\delta = \frac{R}{k} \tag{24}$$

Η παραπάνω αναφορά στον αριθμό κλάσεων, το εύρος και το πλάτος των διαστημάτων κλάσεων κρίθηκε απαραίτητη ώστε να αναλυθεί παρακάτω έννοια των αθροιστικών συχνοτήτων. Οι συχνότητες έχουν ήδη αναλυθεί. Σύμφωνα λοιπόν με

τους Τσίμπος και Γεωργιακώδης (1999) με τον όρο αθροιστικές συχνότητες F_i (cumulative frequencies) εννοείται το άθροισμα των συχνοτήτων, το οποίο για όλες τις κλάσεις πρέπει να ισούται με όλα τα στοιχεία του πληθυσμού. Έτσι εκφράζεται από τον παρακάτω τύπο (25):

$$\sum_{i=1}^{k} f_i = n \tag{25}$$

Ειδικότερα σύμφωνα με τον Everitt (2006) η συνάρτηση αθροιστικής κατανομής (cumulative distribution function) είναι μια κατανομή που εμφανίζει το σύνολο των τιμών μιας τυχαίας μεταβλητής, οι οποίες είναι λιγότερες/περισσότερες από τις δεδομένες τιμές. Η κατανομή αθροιστικών συχνοτήτων (cumulative frequency distribution) κατατάσσει τα παραπάνω σε πίνακα.

Στην περιγραφική στατιστική για την **παρουσίαση των δεδομένων** χρησιμοποιούνται οι **πίνακες** και οι **γραφικές παραστάσεις**. Στους μεν πίνακες περιέχονται τα συγκεντρωμένα στοιχεία, στις δε γραφικές παραστάσεις οι ίδιες πληροφορίες αναπαρίστανται με γραφικό τρόπο.

Ειδικότερα για τις γραφικές παραστάσεις όταν τα δεδομένα είναι **ποιοτικά** τότε χρησιμοποιούνται τα **κυκλικά διαγράμματα ή ραβδογράμματα**. Σύμφωνα με τους Τσίμπος και Γεωργιακώδης (1999), στα **ποσοτικά δεδομένα** αντιστοιχούν τα **ιστογράμματα συχνοτήτων** (αφορούν τον βαθμό συγκέντρωσης των δεδομένων σε σχέση με τις κλάσεις), τα **πολύγωνα και οι καμπύλες συχνοτήτων** (για σύγκριση πολλών κατανομών ταυτόχρονα όταν αφορούν συνεχή δεδομένα, ώστε να εξαχθούν συναρτήσεις που περιγράφουν την σχέση τιμών μεταβλητών και συχνότητας εμφάνισης τους στον πληθυσμό), οι **αψίδες ή πολύγωνα αθροιστικών συχνοτήτων** και τα φυλλογραφήματα.

Ειδικότερα, σύμφωνα τους με Τσίμπος και Γεωργιακώδης (1999) ανάλογα το **σχήμα των καμπυλών των γραφικών παραστάσεων**, υποδεικνύονται τέσσερις κατηγορίες κατανομών: **μονοκόρυφες, σχήματος U, σχήματος J, σύνθετες κατανομές**. Τα **άκρα των καμπυλών** ονομάζονται **ουρές** (tails). Η μονοκόρυφη κατανομή όταν έχει το χαρακτηριστικό σχήμα καμπάνας τότε είναι **συμμετρική** και το εμβαδόν του σχήματος της είναι ίσο με το σύνολο των συχνοτήτων. Η χρήση αυτής της κατανομής είναι πολύ συνηθισμένη στην Επαγωγική Στατιστική και **όταν είναι συμμετρική ονομάζεται κανονική ή γκαουσιανή κατανομή (normal/**

gaussian distribution) ενώ σε άλλη περίπτωση χαρακτηρίζεται ασύμμετρη θετικά ή αρνητικά, το οποίο αφορά τις ουρές της κατανομής.

4.1.2.3 Βασικά στατιστικά μέτρα

Για την επιτυχημένη μελέτη των κατανομών προκειμένου να περιγραφούν τα δεδομένα είναι απαραίτητη η χρήση των παρακάτω στατιστικών περιγραφικών μέτρων τα οποία με βάση τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012) χωρίζονται σε πέντε κατηγορίες:

1. **Μέτρα κεντρικής τάσης** (measures of central tendency): όπου υπολογίζονται ποικίλοι μέσοι όροι. Π.χ. αριθμητικός μέσος ή μέση τιμή.
2. **Μέτρα θέσης** (measures of location): τα οποία αφορούν τη θέση στην κατανομή για τις διάφορες τιμές της μεταβλητής. Π.χ. Διάμεσος
3. **Μέτρα διασποράς ή διασκόρπισης** (measures of variability, dispersion): τα οποία αφορούν τη συγκέντρωση ή μη γύρω από την κεντρική τιμή της μεταβλητής και τη μεταβλητότητα των δεδομένων. Π.χ. εύρος, διακύμανση, τυπική απόκλιση και συντελεστής μεταβλητότητας.
4. **Μέτρα ασυμμετρίας** (measures of skewness): υπολογίζεται ο βαθμός ασυμμετρίας.
5. **Μέτρα κύρτωσης** (kyrtosis measures): αφορά τον υπολογισμό αιχμηρότητας της καμπύλης.

Τα παραπάνω ονομάζονται **στατιστικές** όταν αφορούν την περιγραφή του δείγματος. Στην περίπτωση που αφορούν το σύνολο του πληθυσμού τότε ονομάζονται **παράμετροι.** Από τα παραπάνω μέτρα θα παρουσιαστούν ο αριθμητικός μέσος (κεντρικής τάσης), η διακύμανση και τυπική απόκλιση (κύμανσης), τα οποία αποτελούν κάποια από τα βασικότερα και ευρέως χρησιμοποιούμενα μέτρα.

Όπως αναφέρεται στη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999) ο **αριθμητικός μέσος ή και απλά μέσος** υπολογίζεται από το λόγο του αθροίσματος των τιμών των δεδομένων προς είτε το μέγεθος του **πληθυσμού (*N*)** είτε του **δείγματος (*n*) ανάλογα.** Ο **μέσος πληθυσμού** συμβολίζεται με το γράμμα ***μ*** και υπολογίζεται από τον παρακάτω μαθηματικό τύπο (26):

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \ldots + X_N}{N} \qquad (26)$$

Αν αντικατασταθεί το N με n υπολογίζεται ο μέσος δείγματος, ο οποίος αντί για το γράμμα μ συμβολίζεται με $\bar{x}$.

Όπως αναφέρεται στη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999), **η διακύμανση (variance) και η τυπική απόκλιση (standard deviation)** είναι μέτρα που υπολογίζουν τη **διασπορά για τις τιμές των δεδομένων γύρω από το μέσο τους**. Ειδικότερα η διακύμανση συμβολίζεται σ^2 ή V(X), όταν αντιστοιχεί στη διακύμανση πληθυσμού, ενώ όταν αντιστοιχεί στο δείγμα συμβολίζεται με s^2. Στην τυπική απόκλιση αντιστοιχεί το σύμβολο σ όταν αφορά τον πληθυσμό και s όταν αφορά δείγμα. Η διακύμανση και η τυπική απόκλιση ορίζονται από τους παρακάτω μαθηματικούς τύπους (27) και (28) αντίστοιχα:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{\mathrm{N}} \tag{27}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{\mathrm{N}}} \tag{28}$$

Μάλιστα μεγάλες τιμές για τη διακύμανση και τυπική απόκλιση υποδηλώνουν μεγάλη μεταβλητότητα, δηλαδή μεγάλη απόσταση μεταξύ μιας τιμής δεδομένων από το μέσο της. Τέλος, θα πρέπει τόσο η διακύμανση όσο και η τυπική απόκλιση να είναι μεγαλύτερες ή ίσες με το μηδέν. Όσον αφορά την διακύμανση και τυπική απόκλιση δείγματος στη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999) αναφέρεται ο παρακάτω μαθηματικός τύπος (29):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}(\sum_{i=1}^{n} n\bar{x}^2) \tag{29}$$

4.1.2.4 Κανονική κατανομή και τυποποιημένη τιμή Z

Όταν ο πληθυσμός υπό μελέτη είναι κανονικής κατανομής, όπως αναφέρει ο Zar (2010), μπορεί να οριστεί ένας μέσος, σύμφωνα με τον οποίο ο μισός πληθυσμός είναι μικρότερος από αυτόν και ο άλλος μισός μεγαλύτερος λόγω της συμμετρίας που διέπει την κανονική κατανομή.

Για τον υπολογισμό ποσοστού του πληθυσμού που είναι μεγαλύτερο, για κάποια ποσότητα μεγαλύτερη ή μικρότερη του μέσου, τότε είναι απαραίτητη η κανονική απόκλιση σ του πληθυσμού. Μέσω **της τυποποιημένης τιμής Z** (Zar 2010) **φαίνεται κατά πόσες κανονικές αποκλίσεις απέχει από τον μέσο** (σε ποια θέση βρίσκεται σε σχέση με το μέσο δηλαδή) **μια τιμή** x_i, του πληθυσμού και υπολογίζεται από τον παρακάτω μαθηματικό τύπο (30):

$$Z = \frac{x_i - \mu}{\sigma} \tag{30}$$

Ο παραπάνω υπολογισμός αναφέρεται (Zar 2010) ως κανονικοποίηση (normalizing) ή **τυποποίηση (standardizing)** για κάποια τιμή x_i, και η τιμή Z αναφέρεται ως **κανονική απόκλιση (normal deviate) ή τυποποιημένη τιμή (standard score)**. Η μέση τιμή ενός συνόλου τυποποιημένων τιμών είναι ίση με 0 και η διακύμανση είναι ίση με 1.

4.1.2.4 Μη κανονική κατανομή

Σύμφωνα με τον Zar (2010) όταν ο υπό μελέτη πληθυσμός προέρχεται από μη κανονική κατανομή, η κατανομή θα τείνει προς την κανονικότητα, όσο περισσότερο αυξάνεται το δείγμα n σύμφωνα με το **θεώρημα κεντρικού ορίου** (central limit theorem). Ήδη στην παραπάνω ενότητα (βλέπε ενότητα 4.1.2) έχει παρουσιαστεί η παράμετρος διακύμανσης σ^2. Έτσι για τη **διακύμανση μέσου** $\sigma_{\bar{x}}^2$ (δηλαδή η διακύμανση για την μέση τιμή, όπου στην συγκεκριμένη περίπτωση αφορά μόνο το δείγμα n και όχι τον πληθυσμό) παρατηρείται η μείωση της όσο μειώνεται το n όπως φαίνεται στον παρακάτω μαθηματικό τύπο (31):

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \tag{31}$$

Όπως ήδη έχει αναφερθεί σε παραπάνω ενότητα η τυπική απόκλιση συμβολίζεται με το σύμβολο σ (βλέπε ενότητα 4.1.2). Έτσι η τυπική απόκλιση μέσου $\sigma_{\bar{x}}$ αναφέρεται στη βιβλιογραφία (Zar 2010) και ως τυπικό σφάλμα (μπορεί να είναι σφάλμα τύπου I ή σφάλμα τύπου II) και ορίζεται από την παρακάτω μαθηματική έκφραση (32):

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \tag{32}$$

Σύμφωνα με το μαθηματικό τύπου υπολογισμού της τυποποιημένης τιμής Z (βλέπε εξίσωση (30)), η τυποποιημένη τιμή κατανομής μέσων που αφορά την x_i εκφράζεται με τον παρακάτω μαθηματικό τύπο (33) (Zar 2010):

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \tag{33}$$

Για τον υπολογισμό της τυποποιημένης τιμής πρέπει να είναι γνωστή η διακύμανσης σ^2 για ολόκληρο τον πληθυσμό, κάτι το οποίο είναι δύσκολο να υπολογιστεί. Συνεπώς αντ' αυτού υπολογίζεται η διακύμανση μέσου για τυχαίο δείγμα.

4.1.3 Επαγωγική Στατιστική και θεωρία πιθανοτήτων

Μέχρι τώρα αναλύθηκαν οι βασικοί ορισμοί της Στατιστικής και οι έννοιες και τα μέτρα που αφορούν ειδικότερα την Περιγραφική Στατιστική. Όσον αφορά την Επαγωγική Στατιστική, όπως ήδη έχει αναφερθεί παραπάνω, αποτελεί το τμήμα της Στατιστικής που ασχολείται με τη **διεξαγωγή συμπερασμάτων** για το σύνολο του πληθυσμού, μέσω της λήψης τυχαίου δείγματος. Μάλιστα όπως αναφέρεται στη βιβλιογραφία (Panik 2012) η Επαγωγική Στατιστική κατηγοριοποιείται σε δύο υπό-κλάδους: την Εκτιμητική και τον Έλεγχο Στατιστικών Υποθέσεων, στον οποίο και χρησιμοποιείται η θεωρία των Πιθανοτήτων.

Έτσι στο σημείο αυτό μπορεί να πραγματοποιηθεί μια πιο εκτεταμένη αναφορά στις θεωρίες Συνόλων και Πιθανοτήτων, καθώς κάποιοι βασικοί ορισμοί που τις αφορούν έχουν επεξηγηθεί στις παραπάνω ενότητες.

Σύμφωνα με τον Μπούτσικα (2003) το σύνολο των δυνατών αποτελεσμάτων ενός πειράματος ονομάζεται **δειγματικός χώρος** και τα μεμονωμένα αυτά δυνατά αποτελέσματα ονομάζονται **ενδεχόμενα ή γεγονότα**. Μάλιστα, όπως αναφέρουν οι Δαμιανού, Παπαδάτος και Χαραλαμπίδης (2003) ένας δειγματικός χώρος μπορεί να είναι αριθμήσιμος (πεπερασμένος ή άπειρος αριθμήσιμος) ή μη αριθμήσιμος. Ο αριθμήσιμος δειγματικός χώρος μπορεί να είναι διακριτός ή συνεχής (βλέπε ενότητα 4.1.2.1). Κάποιες βασικές έννοιες της θεωρίας Συνόλων και Πιθανοτήτων, οι οποίες έχουν αντληθεί από τη βιβλιογραφία (Δαμιανού, Παπαδάτος και Χαραλαμπίδη 2003) είναι οι ακόλουθες: ο όρος **σύνολο** αναφέρεται σε μια συλλογή, η οποία περιέχει

κάποια στοιχεία. Τα σύνολα συμβολίζονται με κεφαλαία γράμματα ενώ τα στοιχεία τους με πεζά. Η σχέση μεταξύ συνόλου - υποσυνόλου αφορά ένα σύνολο το οποίο περιλαμβάνει μέσα του ένα άλλο σύνολο. Στη θεωρία Συνόλων και Πιθανοτήτων χρησιμοποιούνται απεικονίσεις για την αναπαράσταση των συνόλων και των σχέσεων τους με άλλα σύνολα, τα γνωστά διαγράμματα Venn. Ο όρος **πιθανότητα** χρησιμοποιείται ώστε να υπολογιστεί το πόσο πιθανή είναι η πραγματοποίηση ενός ενδεχομένου του δειγματικού χώρου. Τα ενδεχόμενα (αναφέρονται επίσης και ως δειγματικά σημεία) και μπορεί να είναι ποσοτικής ή ποιοτικής φύσεως (π.χ. αριθμοί ή περιγραφές αντίστοιχα).

Ειδικότερα σε σχέση με όσα έχουν αναφερθεί στην ενότητα 4.1.2 σχετικά με τις τυχαίες μεταβλητές, όμοια για την εξέταση των ποιοτικών **ενδεχομένων** είναι δυνατή η αντιστοίχιση κάθε ενδεχομένου με έναν πραγματικό αριθμό. Ειδικότερα, σύμφωνα με Αγγελή και Δημάκη (2011), η **αντιστοίχιση αυτή πραγματοποιείται μέσω της συνάρτησης που εκφράζει την τυχαία μεταβλητή.** Σχετικά με τις τυχαίες μεταβλητές έχει γίνει ήδη αναφορά (βλέπε ενότητα 4.1.2.1). Όπως ήδη έχει αναφερθεί η συνάρτηση αυτή διαφέρει ανάλογα το είδος της τυχαίας μεταβλητής και η αντιστοίχιση μιας πιθανότητας για τα ενδεχόμενα της τυχαίας μεταβλητής ονομάζεται συνάρτηση πιθανότητας ή αλλιώς κατανομή πιθανότητας. Τέλος να τονιστεί ξανά πως η παρουσίαση των κατανομών πιθανοτήτων γίνεται κλασικά μέσω πινάκων ή γραφικών παραστάσεων (διάγραμμα πιθανότητας).

4.1.3.1 Έλεγχος Στατιστικών Υποθέσεων

Στην παρούσα διατριβή έχει χρησιμοποιηθεί ο υπο-κλάδος της Επαγωγικής Στατιστικής που αφορά τον **Έλεγχο Στατιστικών Υποθέσεων** για το προτεινόμενο μοντέλο. Έτσι, θα παρουσιαστούν τα βασικά σημεία του υπο-κλάδου αυτού, για την καλύτερη κατανόηση των διεργασιών που πραγματοποιήθηκαν για την ανάπτυξη του προτεινόμενου μοντέλου της διατριβής.

Σύμφωνα με τους Κατσάνος και Αβούρης (2008), ο Έλεγχος Στατιστικών Υποθέσεων (Statistical Hypothesis Testing) αφορά στην επαλήθευση **πρόβλεψης** που αφορά την μεταβλητή υπό μελέτη, η οποία ονομάζεται **υπόθεση (hypothesis)**. Έτσι, η υπόθεση μπορεί να θεωρηθεί ως ένας ισχυρισμός ή μια μη αποδεδειγμένη θεωρία όπως αναφέρει ο Panik (2012). Οι Κατσάνος και Αβούρης (2008) συνοψίζουν τη διαδικασία ελέγχου στατιστικής υπόθεσης στα εξής βήματα:

1. **Διατύπωση υποθέσεων**: στο στάδιο αυτό καθορίζεται ένα **ζευγάρι υποθέσεων**, **η μηδενική υπόθεση H_0 (null hypothesis) και η εναλλακτική H_1 ή H_A (alternative hypothesis).** Σύμφωνα με τον Zar (2010) όταν πρόκειται για διεξαγωγή συμπερασμάτων σε σχέση με το μέσο όρο κάποιου πληθυσμού η **μηδενική υπόθεση συνήθως αποτελεί μια συνοπτική δήλωση που αφορά το μέσο του πληθυσμού** (π.χ. μ=0). Η μηδενική υπόθεση δηλώνεται έχοντας υπόψη ότι μπορεί να είναι ψευδής. Για το λόγο αυτό δηλώνεται και η αντίστοιχη εναλλακτική υπόθεση, η οποία μπορεί να είναι αληθής. Μάλιστα, για την απόρριψη της μηδενικής υπόθεσης θα πρέπει να ισχύει κάποιος βαθμός βεβαιότητας (που συνδέεται παρακάτω με το επίπεδο σημαντικότητας), όπως συμπληρώνουν οι Κατσάνος και Αβούρης (2008).
2. **Διαμόρφωση κριτηρίων**: όπως αναφέρουν οι Κατσάνος και Αβούρης (2008), στον Έλεγχο Στατιστικών Υποθέσεων το βασικό ζήτημα προς επίλυση είναι να διευκρινιστεί εάν οι παρατηρήσεις στις αλλαγές των τιμών αφορούν κάποιο **σφάλμα δειγματοληψίας** (δηλαδή απόκλιση των πραγματικών τιμών του πληθυσμού από αυτούς που εκτιμώνται βάση ενός αντιπροσωπευτικού δείγματος) ή στην **επίδραση της μεταβλητής υπό-μελέτη**. Για το λόγο αυτό τίθενται κριτήρια που ορίζουν την ακριβή διαφορά μεταξύ των παραπάνω περιπτώσεων, ώστε να υπάρχει σαφές πεδίο τιμών που οδηγεί στην απόρριψη της μηδενικής υπόθεσης. Πιο συγκεκριμένα, σύμφωνα με Panik (2012), ορίζεται ένα εύρος τιμών για τη μηδενική υπόθεση, το οποίο θα διασπαστεί σε δύο ξεχωριστά υπό-σύνολα έστω υπο-σύνολο R (αντιστοιχεί ουσιαστικά στην εναλλακτική υπόθεση εφόσον η μηδενική απορρίπτεται) και το συμπλήρωμα του $\bar{R}$. Τα υπο-σύνολα αυτά μπορούν να χαρακτηριστούν ως περιοχές απόρριψης και μη-απόρριψης αντίστοιχα. Έτσι, θα πρέπει να καθοριστεί από τη μια η περιοχή απόρριψης και από την άλλη ένα **μέγεθος *α*,** το οποίο ονομάζεται **επίπεδο σημαντικότητας** (level of significance). Σύμφωνα με Κατσάνος και Αβούρης (2008) αυτό «*καθορίζει τη μέγιστη πιθανότητα το αποτέλεσμα μιας στατιστικής ανάλυσης να οφείλεται σε σφάλματα ή τυχαίους παράγοντες*». Σύμφωνα με τον Zar (2010) το κριτήριο απόρριψης αφορά στο **πόσο μικρή πιθανότητα αποδέχεται ο ερευνητής ώστε να απορρίψει τη μηδενική υπόθεση**. Συνηθισμένη τιμή πιθανότητας για την απόρριψη της μηδενικής υπόθεσης αποτελεί οποιαδήποτε τιμή αγγίζει το 5% και κάτω. Ακόμη το επίπεδο σημαντικότητας α συνδέεται και με την τιμή τυποποίησης

Z (βλέπε ενότητα 4.1.2.4), η οποία όσο μεγαλύτερη είναι, τόσο μικρότερη είναι και η πιθανότητα η μηδενική υπόθεση να είναι αληθής

3. **Συλλογή δεδομένων**: αφορά τη συλλογή αντιπροσωπευτικού δείγματος και εφαρμογή στατιστικών μέτρων (υπολογισμοί και μετρήσεις).
4. **Αξιολόγηση μηδενικής υπόθεσης**: το τελικό αυτό στάδιο αφορά τον έλεγχο του κατά πόσο η μηδενική υπόθεση ισχύει ή όχι και αυτό είναι εφικτό μέσω διάφορων τύπων στατιστικών δεικτών ελέγχου όπως ονομάζονται. Μόλις υπολογιστούν οι τιμές των στατιστικών μέτρων σε συνδυασμό με τα κριτήρια που διαμορφώθηκαν πραγματοποιείται η λήψη απόφασης. Σύμφωνα με Panik (2012) τα είδη ελέγχου χωρίζονται σε μονόπλευρους ελέγχους (one-tail test) ή όπως απαντάται στη βιβλιογραφία κατευθυνόμενους και σε αμφίπλευρους (two-tailed test) ή μη-κατευθυνόμενους. Οι δε μονόπλευροι διακρίνονται σε δεξιόπλευρους και αριστερόπλευρους. Όπως αναφέρουν οι Κατσάνος και Αβούρης (2008) μέσω των αμφίπλευρων ελέγχων μπορεί να παρατηρηθεί «*αν η αλλαγή της ελεγχόμενης μεταβλητής έχει οποιαδήποτε επίδραση στην παρατηρούμενη μεταβλητή*», ενώ μέσω των μονόπλευρων μπορεί να διερευνηθεί και ο τρόπος επίδρασης. Ακόμη, σύμφωνα με τον Zar (2010), μέσω των μονόπλευρων ελέγχων ο ερευνητής μπορεί να εστιάσει στην διαφορά που αφορά σε μια συγκεκριμένη κατεύθυνση προς μια παράμετρο πληθυσμού για μια υποτιθέμενη τιμή.

Καθότι ο καθορισμός κριτηρίου για την αποδοχή ή μη της μηδενικής της υπόθεσης (βήμα 2) αποτελεί κρίσιμο στάδιο για τον Στατιστικό Έλεγχο Υποθέσεων θα πραγματοποιηθεί περαιτέρω ανάλυση όσον αφορά το **επίπεδο σημαντικότητας *α***.

Πιο συγκεκριμένα, το γεγονός ότι η μηδενική υπόθεση δύναται να είναι ψευδής οδηγεί στο συμπέρασμα διεξαγωγής κάποιου τυπικού σφάλματος, όπως αναφέρθηκε και παραπάνω, για το δείγμα προς εξέταση. Το τυπικό σφάλμα μπορεί να ανήκει σε δύο κατηγορίες όπως αναφέρει ο Zar (2010):

1. **Σφάλμα τύπου Ι**: απόρριψη μηδενικής υπόθεσης όταν όντως είναι αληθής. Αντιστοιχεί στο **επίπεδο σημαντικότητας α**.
2. **Σφάλμα τύπου ΙΙ**: μη απόρριψης της μηδενικής υπόθεσης, όταν όντως είναι ψευδής. Η πιθανότητα αυτού του σφάλματος αντιστοιχεί στο σύμβολο **β**. Ως **δύναμη** (power) **στατιστικής υπόθεσης ορίζεται η τιμή 1-β**, η οποία εκφράζει την πιθανότητα **απόρριψης της μηδενικής υπόθεσης όταν είναι πραγματικά ψευδής**.

Όπως αναφέρει ο Zar (2010), οι δύο τύποι σφάλματος **α και β συνδέονται μεταξύ τους αντιστρόφως ανάλογα**, χαμηλές πιθανότητες του α, οδηγούν σε υψηλές πιθανότητες του β, αλλά και τα δύο μπορούν να μειωθούν ταυτόχρονα όταν αυξάνεται το μέγεθος n του δείγματος. Στον Έλεγχο Στατιστικών Υποθέσεων έχει ιδιαίτερη σημασία ο **επιθυμητός συνδυασμός μεταξύ α και β**. Γενικά όταν **α=0.05, τότε υπάρχει μικρή πιθανότητα σφάλματος τύπου Ι και μεγάλη πιθανότητα σφάλματος τύπου ΙΙ**.

Το **επίπεδο σημαντικότητας, ο τύπος ελέγχου στατιστικών υποθέσεων που χρησιμοποιείται και η αντίστοιχη πιθανότητα, θα πρέπει να δηλώνονται στα αποτελέσματα**. Ειδικότερα ο Zar (2010) αναφέρει πως όταν πραγματοποιείται απόρριψη της μηδενικής υπόθεσης με 5% επίπεδο σημαντικότητας μπορεί κανείς να αναφερθεί σε **σημαντική διαφορά** (significant difference), ενώ με 1% σε **πολύ σημαντική διαφορά** (highly significant difference).

Το επίπεδο σημαντικότητας *α*, αντικατοπτρίζεται μέσω της **τιμής p (p value)** της πιθανότητας. Ειδικότερα λοιπόν το επίπεδο σημαντικότητας *α* πρέπει να επιλεγεί από τον ερευνητή. Εναλλακτικά όμως υπάρχει η δυνατότητα καθορισμού του επιπέδου αυτού μέσω των ίδιων των δεδομένων του δείγματος, τα οποία δίνουν το παρατηρούμενο επίπεδο σημαντικότητας, το οποίο αναφέρεται στη βιβλιογραφία (Panik, 2012) ως τιμή p (από το αγγλικό probability). Πιο συγκεκριμένα σύμφωνα με τον Panik (2012), μέσω της τιμής p, υποδηλώνεται η πιθανότητα να «*λάβουμε μια υπολογισμένη τιμή από τον στατιστικό έλεγχο τουλάχιστον τόσο μεγάλη όσο αυτή που έχει παρατηρηθεί για τη μηδενική υπόθεση αν αυτή είναι αληθής*».

Σύμφωνα με τον Panik (2012) **με την τιμή p, μπορούν να διεξαχθούν συμπεράσματα για το αν η μηδενική υπόθεση είναι αληθής ή όχι**. Η τιμή p υπολογίζει την περίπτωση που η μηδενική υπόθεση ισχύει. Πιο συγκεκριμένα, όσο πιο μεγάλη είναι η τιμή p τόσο πιο αληθής θεωρείται η μηδενική υπόθεση. Όσο πιο μικρή είναι η τιμή, δηλαδή όσο πιο σπάνιο θα θεωρούνταν ένα αποτέλεσμα δείγματος για τη μηδενική υπόθεση, οδηγεί στο συμπέρασμα ότι είναι ψευδής. Η τιμή p αφορά την παραπάνω απόδειξη. Πόσο πιθανή ή απίθανη είναι η τιμή σε σχέση με τη μηδενική υπόθεση, εάν αυτή είναι αληθής. Όσο πιο μικρή τιμή p λαμβάνεται, τότε τόσο πιο ψευδής μπορεί να θεωρηθεί η μηδενική υπόθεση.

Όπως αναφέρουν οι Κατσάνος και Αβούρης (2008), η Επαγωγική Στατιστική περιλαμβάνει μια μεγάλη ποικιλία στατιστικών μεθόδων για τον έλεγχο υποθέσεων και η επιλογή της κατάλληλης μεθόδου έγκειται στα εξής:

1. Στόχος διεξαγωγής πειράματος.
2. Πλήθος, είδος και τιμές μεταβλητών (ανεξάρτητες-εξαρτημένες).
3. Πλήθος δειγμάτων.
4. Είδος δεδομένων.
5. Είδος στατιστικού ελέγχου με βάση τις προϋποθέσεις χρήσης του.

Όπως είναι κατανοητό για τόσο μεγάλο εύρος στατιστικών μεθόδων η ταξινόμηση των ειδών ελέγχων είναι αρκετά δύσκολη. Ορισμένες κατηγοριοποιήσεις θα αναφερθούν στη συνέχεια.

Σύμφωνα με τη βιβλιογραφία (Gries to appear in *International Encyclopedia of the Social and Behavioral Sciences*), σχεδόν **όλα τα είδη στατιστικών ελέγχων χρησιμοποιούν τα παρακάτω για τη διεξαγωγή ελέγχου υποθέσεων**: κατανομές, συχνότητες, μέσους όρους (π.χ. μέση τιμή), διασπορά και συσχετίσεις (π.χ. δείκτης συσχέτισης w του Kendall, βλέπε ενότητα 4.2.3). Τα εργαλεία αυτά έχουν περιγραφεί συνοπτικά στην ενότητα 4.1.2.

Μια πρώτη κατηγοριοποίηση των ελέγχων στατιστικών υποθέσεων αναφέρθηκε ήδη στην ενότητα 1.3, με την οποία οι στατιστικοί έλεγχοι διακρίνονται **σε καλής προσαρμογής,** ασχολούνται με την σύγκριση κατανομών και απόκλιση από το αναμενόμενο **και σε ελέγχους ανεξαρτησίας**. Ειδικότερα για τους ελέγχους ανεξαρτησίας σύμφωνα με τη βιβλιογραφία (Gries to appear in *International Encyclopedia of the Social and Behavioral Sciences*) διακρίνονται σε:

a. Μονοπαραγοντικούς (monofactorial), περιέχουν μόνο μια ανεξάρτητη μεταβλητή.
b. Πολυπαραγοντικούς (multifactorial), περιέχουν περισσότερες από μια μεταβλητές:
    i. Μεταβλητές που σχετίζονται αλλά αλληλεπιδρούν μεταξύ τους.
    ii. Μεταβλητές που σχετίζονται αλλά δεν αλληλεπιδρούν μεταξύ τους.

Σύμφωνα με SAS Institute (1999) οι έλεγχοι στατιστικών υποθέσεων χωρίζονται σε **παραμετρικούς** και **μη-παραμετρικούς**. Οι παραμετρικοί έλεγχοι εξαρτώνται από τις προδιαγραφές που φέρει μια κατανομή πιθανότητας και από υποθέσεις σχετικές με την κατανομή ενώ οι μη-παραμετρικοί έλεγχοι όχι. Καθώς οι παραμετρικοί έλεγχοι συχνά υποθέτουν μια κανονική κατανομή πληθυσμού, αν αυτή η υπόθεση δεν ισχύει για το πείραμα τότε συνίσταται η χρήση μη παραμετρικών ελέγχων. Όπως αναφέρεται στη βιβλιογραφία (Neideen και Brasel 2007) ορισμένοι

συνήθεις παραμετρικοί έλεγχοι είναι οι student t-Test, z-Test, ANOVA και ορισμένοι μη παραμετρικοί είναι οι χ^2, Spearman Rank Coefficient, Mann-Whitney U Test, Kruskal-Wallis Test.

4.2 Στατιστικά εργαλεία που χρησιμοποιήθηκαν στην διατριβή

Μετά τη συνοπτική αλλά εμπεριστατωμένη παρουσίαση στατιστικών θεμάτων, στην ενότητα αυτή θα παρουσιαστούν μεμονωμένα στατιστικά εργαλεία που χρησιμοποιήθηκαν στο πλαίσιο της διατριβής και πιο συγκεκριμένα εμφανίζονται στον 5ο κεφάλαιο, στο οποίο και παρουσιάζεται το προτεινόμενο μοντέλο.

4.2.1 Το κριτήριο χ^2 (chi square test)

Το κριτήριο αυτό ονομάζεται επίσης και κριτήριο ελέγχου ανεξαρτησίας (chi square test of independence) ή κριτήριο ελέγχου πινάκων συνάφειας (contingency tables). Σύμφωνα με Εμβαλωτής, Κατσής και Σιδερίδης (2006) το κριτήριο χ^2, είναι ένας στατιστικός έλεγχος μέσω του οποίου μπορεί να ελεγχθεί αν δύο μεταβλητές ενός πίνακα συνάφειας με κ γραμμές και λ στήλες είναι ανεξάρτητες μεταξύ τους. Το κριτήριο αυτό χρησιμοποιεί την κατανομή χ^2, και το επίπεδο σημαντικότητας α. Για να εφαρμοστεί ο στατιστικός αυτός έλεγχος, η διαδικασία ελέγχου απαιτεί τον ορισμού του ζεύγους υποθέσεων στατιστικού ελέγχου, όπως έχει αναφερθεί παραπάνω και έπειτα τις συχνότητες που αναμένονται για τη μηδενική υπόθεση. Έπειτα πραγματοποιείται η σύγκριση με τις πραγματικές συχνότητες. Συνεπώς μέσω του κριτηρίου χ^2 διευκρινίζεται η ανεξαρτησία μεταξύ δύο μεταβλητών.

Όπως αναφέρεται στη βιβλιογραφία (Everitt 2006), οι βαθμοί ελευθερίας (degrees of freedom) αποτελούν τον αριθμό ανεξάρτητων μονάδων πληροφορίας σε ένα δείγμα για τον υπολογισμό μιας στατιστικής, ενώ μπορεί να αντιστοιχεί και στον αριθμό παραμέτρων. Μάλιστα σύμφωνα με τους Εμβαλωτής, Κατσής και Σιδερίδης (2006), οι βαθμοί ελευθερίας ορίζονται από το παρακάτω αριθμητικό γινόμενο (34):

$$(\kappa - 1) \cdot (\lambda - 1) \tag{34}$$

Γενικότερα σύμφωνα με τους Εμβαλωτής, Κατσής και Σιδερίδης (2006) το κριτήριο χ^2 μπορεί να εφαρμοστεί σε ένα σύνολο περιπτώσεων που εξαρτώνται από το είδος της εξαρτημένης και της ανεξάρτητης μεταβλητής. Όλες αυτές οι περιπτώσεις έχουν συγκεντρωθεί στο πίνακα 3:

Πίνακας 3. Περιπτώσεις χρήσης κριτηρίου χ^2 με βάση τα είδη των μεταβλητών που εμπλέκονται

Εξαρτημένη μεταβλητή	**Ανεξάρτητη μεταβλητή**
διατακτική	κατηγορική
κατηγορική	Συνεχής
κατηγορική	διατακτική
κατηγορική	κατηγορική

Σύμφωνα με τον Zar (2010) υπάρχουν πολλές παραλλαγές των τύπων που καθορίζουν τον έλεγχο μέσω του κριτηρίου χ^2.

4.2.2 Σχέσεις μεταξύ μεταβλητών

Πολλές φορές στη Στατιστική διερευνώνται οι σχέσεις μεταξύ των μεταβλητών. Σύμφωνα με τον Zar (2010) για το λόγο αυτό υπάρχουν οι τεχνικές της παλινδρόμησης (regression) και της συσχέτισης (correlation). Αυτές διακρίνονται σε απλή (simple) και πολλαπλή (multiple) παλινδρόμηση/συσχέτιση, όπου αφορούν αντίστοιχα είτε σχέσεις μεταξύ δύο μεταβλητών είτε σχέσεις από τρείς μεταβλητές και πάνω.

Σύμφωνα με τον Zar (2010), ειδικότερα η παλινδρόμηση αφορά τη σχέση μεταβλητών όπου μια εξ αυτών μπορεί να θεωρηθεί εξαρτημένη (συνήθως συμβολίζεται με το Y) και πως το μέγεθος της εξαρτάται από το μέγεθος της άλλης (συνήθως συμβολίζεται με X) ή των άλλων μεταβλητών, δηλαδή των ανεξάρτητων. Πρόκειται λοιπόν για σχέση εξάρτησης μεταξύ μεταβλητών. Για την απεικόνιση της παλινδρόμησης συνήθως χρησιμοποιείται διάγραμμα διασποράς (scatter diagram) το οποίο παρουσιάζει τη διασπορά με τις τιμές της εξαρτημένης μεταβλητής στον άξονα y και της ανεξάρτητης στον x για την απλή παλινδρόμηση.

Σύμφωνα με τον Zar (2010) στην περίπτωση που στις μεταβλητές δεν υπάρχει σχέση εξάρτησης (εξαρτημένη και ανεξάρτητη μεταβλητή) αλλά καθώς αλλάζουν οι τιμές της μιας, αλλάζουν και της άλλης, η σχέση αυτή ονομάζεται συσχέτιση. Για τον υπολογισμό της συσχέτισης θα πρέπει να υποτεθεί πως κάθε μεταβλητή υπακούει σε μια κανονική κατανομή για κάθε συνδυασμό των υπόλοιπων μεταβλητών. Όπως αναφέρουν οι Κατσάνος και Αβούρης (2008) μέσω του υπολογισμού ενός δείκτη συσχέτισης είναι δυνατή η μέτρηση της σχέσης μεταξύ των μεταβλητών. Οι τιμές του συντελεστή συσχέτισης μπορεί να είναι από -1 έως +1 και μετράται μόνο η αύξηση του, χωρίς να ενδιαφέρει αν είναι αρνητική ή θετική η τιμή του συντελεστή.

Μάλιστα, όσο πιο πολύ αυξάνεται τόσο πιο πολύ συνδέονται οι μεταβλητές μεταξύ τους ενώ στην περίπτωση που παίρνει την τιμή 0 τότε οι μεταβλητές δεν συσχετίζονται. Ο πιο διαδεδομένος συντελεστής συσχέτισης είναι ο συντελεστής συσχέτισης Pearson, με το σύμβολο *r*.

4.2.3 Δείκτης συμφωνίας w του Kendall

Ο δείκτης συμφωνίας w του Kendall ή αλλιώς βαθμός συμφωνίας μεταξύ κριτών αποτελεί στατιστική που υπάγεται στα μη παραμετρικά τεστ σύμφωνα με τον Legendre (2010) και αποτελεί ένα μέτρο συμφωνίας μεταξύ μεταβλητών (Legendre 2005). Σύμφωνα με τον Zar (2010) η πολλαπλή συσχέτιση μεταβλητών είναι εφικτή μέσω του δείκτη συμφωνίας w του Kendall. Ο μαθηματικός τύπος που ορίζει τον υπολογισμό του δείκτη συμφωνίας w του Kendall έχει πολλές παραλλαγές. Ένας εξ αυτών είναι ο (35):

$$W = \frac{\sum R_i^2 - \frac{(\sum R_i)^2}{n}}{\frac{M^2(n^3-1)}{n}} \tag{35}$$

Στον παραπάνω τύπο το M αντιστοιχεί στον αριθμό μεταβλητών που συσχετίζονται, το n αντιστοιχεί στον αριθμό δεδομένων για κάθε μεταβλητή, ενώ το R αποτελεί την κατάταξη των τιμών για κάθε μεταβλητή. Η τιμή του δείκτη συμφωνίας w μπορεί να κυμαίνεται από 0 σε 1. Η τιμή μηδέν αναπαριστά την μη συσχέτιση, ή αλλιώς σύμφωνα με τον Έλεγχο Στατιστικών Υποθέσεων αντιστοιχεί στην μηδενική υπόθεση H_0 ότι οι μεταβλητές δε συσχετίζονται. Η τιμή 1 αναπαριστά την πλήρη συσχέτιση, ή αλλιώς την εναλλακτική υπόθεση H_A, ότι οι μεταβλητές συσχετίζονται.

Σύμφωνα με τον Zar (2010) είναι δυνατός ο υπολογισμός της σημαντικότητας του δείκτη συμφωνίας w του Kendall μέσω της παραλλαγής του κριτηρίου χ_r^2 του Friedman και ειδικότερα αντιστοιχίζοντας στον w το ισοδύναμο χ_r^2, όπως φαίνεται στον παρακάτω μαθηματικό τύπο (36):

$$\chi_r^2 = M(n-1)W \tag{36}$$

Τέλος, σύμφωνα με τον Zar (2010) *«αν οι κριτικές τιμές του πίνακα για τα n, M, είναι μεγαλύτερες από τον πίνακα τότε μπορούμε να υποθέσουμε ότι το* χ_r^2 *προσεγγίζεται από το* χ^2*, με n-1 βαθμούς ελευθερίας»*.

ΚΕΦΑΛΑΙΟ 5^{Ο}
ΠΡΟΤΕΙΝΟΜΕΝΟ ΜΟΝΤΕΛΟ ΑΝΑΚΤΗΣΗΣ ΣΗΜΑΙΝΟΝΤΩΝ ΟΡΩΝ ΕΓΓΡΑΦΩΝ

5.1 Εισαγωγή στο προτεινόμενο μοντέλο

Η παρούσα διδακτορική διατριβή ξεκίνησε με μια θεωρητική μελέτη των ιδιοτήτων που πιθανόν παρουσιάζει ένα κείμενο και εστιάσθηκε στη δυνατότητα να εξαχθούν μέσω τέτοιων ιδιοτήτων σημαίνουσες λέξεις, οι οποίες δυνητικά θα αποτελούσαν τη δομική ύλη κατασκευής μιας οντολογίας. Η οντολογία αυτή θα αποτελούσε τη βάση κατασκευής ενός σημασιολογικού δικτύου. Προς αυτό το σκοπό χρησιμοποιήθηκε το διανυσματικό μοντέλο του Salton (βλέπε ενότητα 2.2.1).

Συγκεκριμένα, στο μοντέλο VSM ο κάθε όρος – λέξη αποτελεί διάσταση σε έναν πολυδιάστατο χώρο, στο οποίο το διάνυσμα του εγγράφου αναπαρίσταται στις διαστάσεις των όρων. Οι συνιστώσες του διανύσματος αυτού βασίζονται στη συχνότητα εμφάνισης όρων ως προς το έγγραφο σε κανονικοποιημένη μορφή και σε σχέση με όλη τη συλλογή εγγράφων, σύμφωνα με το σχήμα απόδοσης βαρών σε όρους *tf - idf*.

Στο προτεινόμενο μοντέλο, η διατριβή χρησιμοποιεί την εξής παραδοχή: το διάνυσμα που αναπαριστά κάθε κείμενο κινείται σε τρισδιάστατο χώρο, όπου κάθε συνιστώσα του καθορίζεται **(α)** από τη θέση του όρου, **(β)** από την ένταση του όρου χρησιμοποιώντας την κωδικοποίηση ASCII και **(γ)** από τον αριθμό χαρακτήρων όρου. Το δε κριτήριο ομοιότητας ορίζεται ως το συνημίτονο (γωνία απόκλισης) του διανύσματος κάθε όρου ως προς τη συνολική συνισταμένη όλου του εγγράφου. Η στατιστική θεμελίωση της συσχέτισης μεταξύ των παραπάνω τριών μεταβλητών διεξήχθη μέσω του δείκτη συμφωνίας w του Kendall (βλέπε ενότητα 4.2.3), με σκοπό να διαπιστωθεί εάν και κατά πόσον αυτές οι τρεις μεταβλητές δύναται να συσχετιστούν μέσα σε ένα συγκεκριμένο αριθμό όρων. Η διαδικασία όλου αυτού του πειράματος μπορεί να περιγραφεί μέσω των δύο δημοσιεύσεων (Poulimenou S. et al. 2014, Poulimenou S. et al. *to appear in* 2016) που ακολουθούν:

5.2 Εξαγωγή λέξεων κλειδιών από τίτλους άρθρων για οντολογικές χρήσεις

Abstract— Σε αυτό το άρθρο παρουσιάζεται ένας καινοτόμος αλγόριθμος που έχει στόχο την εξαγωγή, με την έννοια της απομείωσης, βέλτιστης τριάδας λέξεων κλειδιών, οι οποίοι αποτελούν περιγραφικούς όρους κειμένου για τίτλους επιστημονικών άρθρων. Οι όροι αυτοί χρησιμοποιούνται ως ερώτημα (βλέπε ενότητα της διατριβής 2.1) που εισάγεται στη μηχανή αναζήτησης από έναν χρήστη ως λέξεις κλειδιά σε φυσική γλώσσα, για την αναζήτηση και ανάκτηση επιστημονικών άρθρων από ψηφιακές βάσεις δεδομένων ή/και αποθετήρια. Ο αλγόριθμος αυτός βασίζεται στο VSM (βλέπε ενότητα της διατριβής 2.2.1) ώστε να αναπαραστήσει τους τίτλους άρθρων αντιστοιχίζοντας κάθε όρο του τίτλου ως διάνυσμα, όπου κάθε διάνυσμα με τη σειρά του καθορίζεται από τρεις μεταβλητές, τον αριθμό χαρακτήρων του όρου, τον αριθμό κωδικοποίησης κάθε όρου και τη σειρά/θέση του όρου στον τίτλο. Βασισμένος στα βάρη, ο αλγόριθμος υπολογίζει κάθε μια από τις τρεις παραπάνω μεταβλητές για κάθε όρο και προτείνει το βαθμό καταλληλότητας κάθε όρου ως λέξη κλειδί, προκειμένου να ανακτήσει το αντίστοιχο άρθρο στην κορυφή της κατάταξης. Η πειραματική αξιολόγηση του αλγόριθμου με πραγματικά επιστημονικά δεδομένα αποδεικνύει την αποτελεσματικότητα του στην ανίχνευση περιγραφικών λέξεων κλειδιών κειμένου και επιβεβαιώνει την υπόθεση του άρθρου ότι στην περίπτωση των επιστημονικών εκδόσεων, όροι εξαγόμενοι από τίτλους μπορούν να εκφράζουν τα άρθρα.

I. ΕΙΣΑΓΩΓΗ

Η προσέγγιση μέσω λέξεων κλειδιών θεωρείται φιλική στο χρήστη και εύκολη στην εφαρμογή με αποδεκτά αποτελέσματα όσον αφορά την ακρίβεια (βλέπε ενότητα της διατριβής 2.3.3) στην ανάκτηση. Παράλληλα οι σημασιολογικά πλούσιες οντολογίες αντιμετωπίζουν την ανάγκη για πλήρης περιγραφές ανάκτησης κειμένου και βελτιώνουν την ακρίβεια της ανάκτησης [1]. Η εξαγωγή/απομείωση πληροφορίας, όπως είναι και οι λέξεις κλειδιά (βλέπε ενότητες της διατριβής 2.4, 2.4.1 και 3.1.4) είναι πολύ σημαντική για ανάκτηση κειμένου, ανάκτηση ιστοσελίδων, ομαδοποίηση κειμένου, παρουσιάσεις, εξόρυξη κειμένου κλπ. Εξάγοντας κατάλληλες λέξεις κλειδιά μπορεί κανείς εύκολα να διαλέξει ποιο κείμενο

να διαβάσει και να μάθει τη σχέση μεταξύ εγγράφων [2]. Ένα βασικό σχήμα απόδοσης βαρών σε όρους για την ευρετηρίαση τους είναι το *Term Frequency - Inverse Document Frequency* (*tf - idf*) [3], (βλέπε ενότητα της διατριβής 2.2.2), το οποίο εξάγει λέξεις κλειδιά που εμφανίζονται συχνά σε μεμονωμένα κείμενα αλλά όχι στο σύνολο της συλλογής [4,5]. Είναι υπολογιστικά εφικτό και αποδίδει αρκετά καλά [6]. Ο όρος «*εξαγωγή λέξεων κλειδιών*» χρησιμοποιείται στο πλαίσιο της εξόρυξης κειμένου [3]. Η εξαγωγή λέξεων κλειδιών έχει αντιμετωπιστεί επίσης ως εποπτευόμενο πρόβλημα μάθησης [4, 5, 7], όπου ένας ταξινομητής χρησιμοποιείται για την ταξινόμηση υποψήφιων λέξεων σε θετικά ή αρνητικά στιγμιότυπα, χρησιμοποιώντας κάποιο σύνολο χαρακτηριστικών. Σε άλλες έρευνες εξαγωγής λέξεων κλειδιών έχουν εκμεταλλευτεί σημασιολογικές πηγές [8], μετρική βασισμένη στον ιστό, όπως PMI score (point-wise mutual information) [7], ή αλγόριθμους βασισμένους σε γράφους (π.χ., [9] όπου αποπειράθηκαν να χρησιμοποιήσουν μια ενισχυμένη προσέγγιση όπου θα ήταν δυνατή η ταυτόχρονη εξαγωγή λέξεων κλειδιών και η ομαδοποίηση.) [6].

Σχετικές επιστημονικές εργασίες έχουν διεξαχθεί όπου οι συχνοί όροι (*ft*) θεωρούνται ζωτικής σημασίας [10], συνεπώς εξάγονται προκειμένου να είναι εφικτή η εξαγωγή λέξεων κλειδιών από ένα κείμενο. Σε μια εναλλακτική προσέγγιση, ο αλγόριθμος της έρευνας [11], όπου όχι μόνο διεξάγεται μέτρηση της συχνότητας όρων και άλλων στατιστικών αλλά εφαρμόζεται και η χρήση επαγγελματικών ευρετηριάσεων.

Επιπλέον, το σχήμα απόδοσης βαρών σε όρους *tf - idf* συνδυάζεται με το VSM για αυτόν το σκοπό [12]. Το *tf - idf* χρησιμοποιείται πολύ συχνά στην ΑΠ προκειμένου να συγκρίνει ένα διάνυσμα ερωτήματος με το διάνυσμα εγγράφου χρησιμοποιώντας εξίσωση ομοιότητας ή απόστασης όπως είναι η συνάρτηση ομοιότητας συνημίτονου. Παρόλα αυτά το πρόβλημα εστιάζεται στη δυσαρμονία με θεωρία της πληροφορίας του Shannon [13] (βλέπε ενότητα της διατριβής 3.3.2). Περισσότερες λεπτομέρειες δίνονται στην ενότητα *II.A*. Ο αλγόριθμος που προτείνεται σε αυτό το άρθρο μπορεί να θεωρηθεί καινοτόμος, καθώς δε βασίζεται στη συχνότητα με την οποία εμφανίζεται ένας όρος (*tf - idf*), όπου θεωρεί κάθε λέξη σε ένα κείμενο με ίση βαρύτητα, αλλά εισάγει τρεις μεταβλητές που προσδιορίζουν κάθε λέξη μοναδικά. Συγκρίνοντας τους δύο αλγόριθμους, στον *tf - idf* η γωνία ενός διανύσματος λέξης και η διανυσματική συνιστώσα αναπαριστούν τη συχνότητα εμφάνισης του όρου, ενώ στον προτεινόμενο αλγόριθμο η γωνία αναπαριστά τρεις

μοναδικές μεταβλητές που ορίζουν μια ξεχωριστή διανυσματική ταυτότητα για κάθε λέξη.

Είναι γνωστό ότι οι οντολογίες συσχετίζονται με ένα μοντέλο γνώσης, εννοώντας πληροφορίας. Συνεπώς είναι λογική η εισαγωγή της έννοιας της εντροπίας και αμοιβαίων πληροφοριών, όπως αυτά ορίζονται από τον Shannon για τη θεωρία πληροφοριών [14], στις οντολογίες. Η εντροπία (βλέπε ενότητα της διατριβής 3.3.2) και οι αμοιβαίες πληροφορίες δίνουν τη δυνατότητα να ορίσουμε επίσημα ένα μέσο μέτρησης απόστασης. Με αυτή την απόσταση ένα γερό θεμέλιο δίνεται για τη λήψη της εγγενούς δομής της οντολογίας. Κατά συνέπεια, στο άρθρο αυτό, γίνεται μια προσπάθεια χρήσης ενός νέου αλγόριθμου, ο οποίος έχει δημιουργηθεί βασισμένος στη φιλοσοφία του *tf - idf* δίνοντας όμως ταυτόχρονα μια λύση στη δυσαρμονία των μοντέλων τύπου Shannon.

Αυτό το άρθρο είναι οργανωμένο ως εξής:

- Ενότητα II, παρέχει θεωρητικές και πρακτικές εφαρμογές αυτής της μελέτης που σχετίζονται με τον αλγόριθμο *tf - idf*, τη βάση του αλγόριθμου και το κριτήριο κατάταξης.
- Ενότητα III, αναλύεται το πειραματικό σχέδιο, στάδιο συλλογής δεδομένων, η εφαρμογή του αλγόριθμου, η ανάκτηση και αξιολόγηση των αποτελεσμάτων.
- Ενότητα IV, πραγματοποιείται μια σύγκριση μεταξύ του *tf – idf* και του αλγόριθμου που παρουσιάζεται στο άρθρο αυτό.
- Ενότητα V, παρουσιάζονται τα συμπεράσματα και τα μελλοντικά σχέδια.

II. Μεθοδολογια

A. Tf – idf

Το σχήμα *tf - idf* αποτελεί το πιο ευρέως χρησιμοποιούμενο για την απόδοση βάρους στην ΑΠ ως τώρα. Βασίζεται σε τρία συστατικά:

- Document Frequency (DF)
- Inverse Document Frequency (IDF)
- Term Frequency (TF)

Και εφαρμόζεται από τη φόρμουλα

$$TF / IDF = Term\_Frequency\_X\_Inverse\_Document\_Frequency$$

Η μέσω της εξίσωσης (1).

$$w_{t,d} = \left(1 + \log tf_{t,d}\right) * \log_{10}\left(\frac{N}{df_t}\right) \quad (1)$$

Ένα χαρακτηριστικό του σχήματος *tf - idf* είναι πως η βαθμολογία (βάρος) που υπολογίζει για έναν όρο, αυξάνεται μαζί με την αύξηση συχνότητας εμφανίσεων του σε ένα έγγραφο (συστατικό *tf*). Ακόμη η βαθμολογία αυξάνεται όσο πιο σπάνια εμφανίζεται ένας όρος σε ολόκληρη τη συλλογή (συστατικό *idf*). Πρόκειται για μια διαρκή δυναμική διαδικασία. Επιπλέον, η τεχνική αυτή απαιτεί πολυάριθμες συνδέσεις ανάμεσα στους όρους και τα έγγραφα. Τα σημαντικότερα προβλήματα αναφέρονται για τη διαδικασία της αναζήτησης όταν χρησιμοποιούνται βάρη όρων [12, 13].

Πιο συγκεκριμένα, ο πυρήνας του προβλήματος εντοπίζεται στη δυσκολία αναγνώρισης ενός μεμονωμένου δειγματικού χώρου (βλέπε ενότητα της διατριβής 4.1.3) και αντίστοιχα μέτρων υπολογισμού πιθανοτήτων, μέσω των οποίων μπορούν να καθοριστούν όλες οι μεταβλητές (βλέπε ενότητα της διατριβής 4.1.2). Χωρίς ένα τέτοιο ενιαίο δειγματικό χώρο, κάθε διαδικασία που περιέχει αντιστοίχιση ή/και συγκερασμό διαφορετικών μέτρων (τοπικού ή παγκόσμιου επιπέδου) μπορεί να μην είναι αποδεκτή βάση της θεωρίας Shannon [13] (βλέπε ενότητα της διατριβής 3.3.2). Ένα σημαντικότερο πρόβλημα εμφανίζεται όταν γίνεται αναζήτηση χρησιμοποιώντας όρους που τους έχουν αποδοθεί βάρη. Καθώς η συνήθης πρακτική αποτελεί την επιλογή μονάχα των όρων που απαρτίζουν το ερώτημα του χρήστη, όλοι οι υπόλοιποι όροι του λεξιλογίου αγνοούνται. Πρόκειται για μια πρακτική χωρίς νόημα για τα μοντέλα τύπου Shannon διότι: κάθε όρος υποτίθεται πως μεταβιβάζει μια συγκεκριμένη ποσότητα πληροφορίας. Τότε, αν θεωρηθεί ότι η κάθε συλλογή *N* κειμένων αποτελείται από k_i όρους που αυτοί μπορεί να είναι λέξεις, φράσεις, προθέματα λέξεων ή μορφήματα τότε κάθε ερώτημα όρου παράγει μία ποσότητα πληροφορίας ίση με $\log P(k_i)$, ανεξάρτητα με το αν ο όρος αυτός υπάρχει στο ερώτημα. Έτσι στην περίπτωση του ερωτήματος, δεν υπάρχει σύνδεση για όλη την ποσότητα πληροφορίας με το συγκεκριμένο ερώτημα και δεν υπάρχει κάποια δικαιολογία γιατί να επιτρέπεται να συμβαίνει αυτό [13]. Παρόλα αυτά η ομοιότητα της βασικής *idf* διατύπωσης με το συστατικό της εντροπίας ενέπνευσε πολλούς

ερευνητές [14] να πραγματοποιήσουν συνδέσεις διαφορετικές από τις παραπάνω προτεινόμενες.

Β. Η βάση του αλγόριθμου

Η αναπαράσταση των μεταβλητών, οι οποίες ορίζουν μοναδικά το βάρος ενός όρου, βασίζεται στη θεωρία του VSM. Πιο συγκεκριμένα, θεωρήθηκε πως ο τίτλος ενός δημοσιευμένου εγγράφου μπορεί να αναπαρασταθεί από ένα ορισμένο διάνυσμα $\vec{V}_s$ το οποίο είναι μοναδικό. Αυτή η τεχνική δημιουργεί έναν μηχανισμό αυτόματης συσχέτισης, ο οποίος είναι μοναδικός για κάθε κείμενο και η πρακτική αυτή δίνει λύση για τα μοντέλα τύπου Shannon καθώς κάθε κείμενο υποβαθμίζεται σε κάθε όρο τοπικά. Συνεπώς για κάθε κείμενο δημιουργείται ένα σταθερός δειγματικός χώρος με υπολογίσιμο μέτρο πιθανότητας (βλέπε ενότητα της διατριβής 4.1.2). Λαμβάνοντας υπόψη τα παραπάνω, θεωρείται πως κάθε μια λέξη ενός συγκεκριμένου τίτλου αναπαρίσταται από ένα διάνυσμα $\vec{V}_w$ το οποίο αποτελείται από τρεις μεταβλητές, οι οποίες αναφέρονται ονομαστικά: η θέση/σειρά της λέξης στον τίτλο (*i*), ο αριθμός κωδικοποίησης (*s*) που αποτελεί έναν αριθμό ταυτότητας, και ο αριθμός των χαρακτήρων (*w*), (βλέπε εξίσωση 2).

$$\vec{V}_w^{(i)}\Big|_{i=0}^{n} = \begin{bmatrix} i & s_i & w_i \end{bmatrix} \tag{2}$$

Όπου,

$$s_0 = 0 \wedge w_0 = 0$$

Επειδή οι τιμές των τριών μεταβλητών έχουν μια μη κανονικοποιημένη μορφή, για το λόγο αυτό, κανονικοποιείται (βλέπε ενότητα της διατριβής 2.2.2) το διάνυσμα $\vec{V}_w$ και εξάγεται το ισοδύναμο του μέσα από την εξίσωση (3):

$$\vec{V}_{we}^{\,i}\Big|_{i=0}^{n} = \frac{\vec{V}_w^{\,i}\Big|_{i=0}^{n}}{\left\|\vec{V}_w^{\,i}\Big|_{i=0}^{n}\right\|} \tag{3}$$

Για τον αριθμό κωδικοποίησης *s* χρησιμοποιείται η κωδικοποίηση χαρακτήρων ASCII [15], περισσότερες λεπτομέρειες για αυτή τη διαδικασία παρέχονται στο πειραματικό μέρος.

Ακόμη, το διάνυσμα $\vec{V}_s$ αναπαρίσταται ως το διάνυσμα της συνισταμένης των διανυσμάτων λέξεων (άθροισμα των επιμέρους διανυσμάτων), βλέπε εξίσωση 4.

$$\vec{V}_s = \sum_{i=0}^{n} \vec{V}_{we}^{(i)} \tag{4}$$

C. Το κριτήριο κατάταξης

Οι κατατάξεις εγγύτητας εγγράφων για ένα ερώτημα χρήστη μπορούν να υπολογιστούν μέσω της υπόθεσης της θεωρίας Ομοιότητας Εγγράφων [16]. Έτσι, από την προηγούμενη επεξεργασία αλγορίθμου, όπου η διαδικασία κατάταξης πραγματοποιούνταν σύμφωνα με την θέση *i*, γίνεται αντικατάσταση που βασίζεται στην θεωρία Ομοιότητας Εγγράφων. Συγκρίνεται λοιπόν το συνημίτονο γωνίας (βλέπε ενότητες της διατριβής 2.2 και 3.2.3.2) ανάμεσα στο διάνυσμα εγγράφου και στο διάνυσμα ερωτήματος, το οποίο αποτελεί την συνισταμένη $\vec{V}_s$ και αναπαριστά τον κυρίαρχο προσανατολισμό του διανύσματος του τίτλου. Ο υπολογισμός του συνημίτονου της γωνίας μεταξύ των διανυσμάτων ορίζεται από την εξίσωση (5).

$$\cos\theta^{(i)} = \frac{\vec{V}_{we}^{(i)} \vec{V}_s}{\left\|\vec{V}_{we}^{(i)}\right\| \left\|\vec{V}_s\right\|} \tag{5}$$

Η νέα κατάταξη του

$$r_{j+1}\Big|_{j=0}^{n} = \min\left[\cos\theta^{(\imath)} - r_0\right] \tag{6}$$

Όπου

$$r_0 = \min\left[\cos\theta^{(i)}\right]$$

D. Η πρωτοτυπία του Αλγόριθμου

Λαμβάνοντας υπόψη τις ενότητες *II.A* και *II.B* ο υπολογισμός των διανυσμάτων των εγγράφων μπορεί να υπολογιστεί με δύο διαφορετικές τεχνικές.

Στην κλασσική θεωρία του VSM χρησιμοποιώντας το σχήμα απόδοσης βάρους σε όρους *tf - idf*, το διάνυσμα του εγγράφου $\vec{U}_s$ που εξάγεται χρησιμοποιεί πολυμεταβλητά συστατικά *i*.

$$\vec{U}_s = \left[w_1, w_{2,}..., w_i\right] \tag{7}$$

Στην προτεινόμενη μέθοδο, το διάνυσμα εγγράφου $\vec{V}_s$ εξάγεται

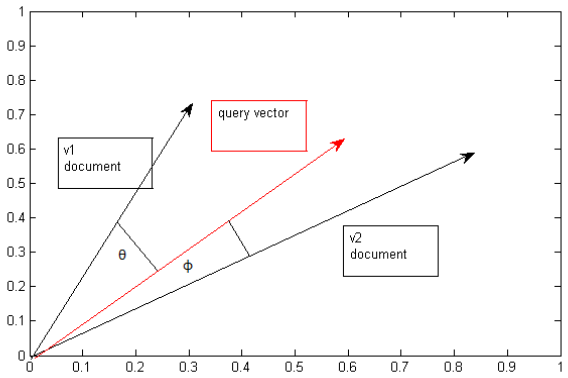
χρησιμοποιώντας τις τρεις προαναφερθείσες μεταβλητές και ένα $\vec{V}_i = [i \quad s_i \quad w_i]$. Η προτεινόμενη μέθοδος χρησιμοποιεί σταθερό μέγεθος για το διάνυσμα εγγράφου (αναπαρίσταται σε τρεις διαστάσεις).

Ομοίως, το διάνυσμα του ερωτήματος, με βάση το προτεινόμενο διάνυσμα αναπαρίσταται σε ίσο αριθμό διαστάσεων με αυτό, ενώ το κλασσικό διάνυσμα αναπαρίσταται σε *i* διαστάσεις.

Υιοθετείται ένα διάνυσμα $\vec{V}_i$ αντί του $\vec{U}_s$ όπου η *i* μεταβλητή αντικαθίσταται από την *r* (βλέπε εξίσωση 6). Ο λόγος της αντικατάστασης είναι διότι η μεταβλητή *r* θεωρείται πολύ σημαντική στη θεωρία της Σημασιολογίας [17]. Η αντικατάσταση έγινε σκόπιμα καθώς το διάνυσμα $\vec{V}_i$ αναπαριστά το βαθμό επιρροής μέσω της μεταβλητής *r*, και πρόκειται να οδηγήσει σε μια νέα διαδικασία κατάταξης βάση της επιρροής.

$$\vec{U}_i = [r_i \quad s_i \quad w_i] \tag{8}$$

Τέλος και στις δύο υποθέσεις, σε μια αναζήτηση με ερώτημα χρήστη που αποτελείται από λέξεις κλειδιά, μπορεί να υπολογιστεί η κατάταξη εγγύτητας εγγράφων, χρησιμοποιώντας τις υποθέσεις της θεωρίας Ομοιότητας Εγγράφων (βλέπε ενότητα της διατριβής 3.2.3.2), μέσω της σύγκρισης των γωνιών απόκλισης ανάμεσα στα διανύσματα των εγγράφων και της συνιστάμενης του διανύσματος ερωτήματος (βλέπε εικ.1).



Εικ. 1. Η διαδικασία ερωτήματος ανάμεσα σε διάνυσμα ερωτήματος και διάνυσμα εγγράφου

Η αντιστοίχιση αυτή πραγματοποιείται με τον υπολογισμό του συνημίτονου των γωνιών (θ, φ παράδειγμα βλέπε εικ. 1) ανάμεσα στα διανύσματα, αντί για την ίδια την μεταξύ τους γωνία, βλέπε εξισώσεις 5 και 6.

ΙΙΙ. ΠΕΙΡΑΜΑΤΙΚΟ ΣΤΑΔΙΟ

Α. *Η συλλογή δεδομένων*

Αρχικά χρησιμοποιείται ένας φυλλομετρητής διαδικτύου για την πρόσβαση στη μηχανή αναζήτησης (βλέπε ενότητα της διατριβής 2.4.1) Google Scholar από όπου και γίνεται η συλλογή τίτλων όπου θεματικά ανήκουν σε 3 διαφορετικές τάξεις της ταξινόμησης Dewey. Πιο συγκεκριμένα από κάθε τάξη επιλέχθηκαν 10 τίτλοι προερχόμενοι από 10 διαφορετικές υπο-τάξεις, έτσι συλλέχθηκε ένα σύνολο από 181 τίτλους συνολικά, από 3 διαφορετικές τάξεις Dewey. Οι υπο-τάξεις αυτές είναι: *Knowledge, Systems, Bibliographies, Catalogs, Libraries, Biographies, Topology, Publishing, Manuscripts and Algebra*. Συνεπώς το συνολικό δείγμα τίτλων είναι 181.

Αφού ήδη εξηγήθηκε και παρουσιάστηκε εκτενώς ο αλγόριθμος ο οποίος εξάγει τις τρεις κυρίαρχες λέξεις κλειδιά και προτού να αρχίσει η εισαγωγή τίτλων για επεξεργασία από τον αλγόριθμο πρέπει να πραγματοποιηθεί η προ-επεξεργασία (βλέπε ενότητες της διατριβής 1.4 και 3.1.4) του τίτλου. Ουσιαστικά, πραγματοποιείται ένα φιλτράρισμα του τίτλου με αυτόματο τρόπο ώστε να αφαιρεθούν οι λέξεις που δεν θα μπορούσαν να αποτελέσουν λέξεις-κλειδιά.

Αφού ο τίτλος φιλτραριστεί, το επόμενο βήμα είναι η εφαρμογή του αλγόριθμου σε κάθε τίτλο. Όλα τα αποτελέσματα για κάθε ξεχωριστό τίτλο συγκεντρώθηκαν και καταχωρήθηκαν σε αρχείο. Το αποτέλεσμα που παράγεται κάθε φορά επιστρέφει τις τρεις κυρίαρχες λέξεις ως λέξεις κλειδιά για κάθε τίτλο.

Στο επόμενο βήμα γίνεται έλεγχος του αποτελέσματος προκειμένου να μετρηθεί η επιτυχία των αποτελεσμάτων του πειράματος. Γίνεται αναζήτηση στο Google Scholar (για την περιοχή του τίτλου) χρησιμοποιώντας τα ακριβή αποτελέσματα που επιστρέφει ο αλγόριθμος και γίνεται στατιστική αποτίμηση των συνολικών αποτελεσμάτων.

Για περισσότερη ακρίβεια θα παρουσιαστεί ένα παράδειγμα βήμα προς βήμα ακολουθώντας τη μεθοδολογία του πειράματος. Η έναρξη της διαδικασίας αφορά την επιλογή ενός τίτλου άρθρου, ο οποίος προκύπτει μέσα από αναζήτηση στο Google Scholar. Ας θεωρηθεί για το παράδειγμα ότι ο όρος αναζήτησης είναι “Knowledge” και ο τίτλος που ανακτάται “Advances in knowledge discovery and data mining”. Ο τίτλος αυτός προ-επεξεργάζεται πριν την εφαρμογή του αλγόριθμου. Με τη βοήθεια

λογισμικού POS (βλέπε ενότητα της διατριβής 3.2.2.1), διευκολύνεται η προ-επεξεργασία του τίτλου. Κρατούνται ή αφαιρούνται οι λέξεις, με βάση το αν μπορούν να θεωρηθούν πιθανές λέξεις κλειδιά, ανάλογα με το μέρος του λόγου στο οποίο ανήκουν. Όλες οι διακόπτουσες λέξεις (βλέπε ενότητες της διατριβής 2.4 και 2.4.1) αφαιρούνται και ακόμη μέρη του λόγου που έχουν τεθεί υπό περιορισμό για το πείραμα αυτό, αφαιρούνται και αυτά. Συνεπώς, από τον αρχικό τίτλο απομένει ο "Advances knowledge discovery data mining". Σε αυτό το σημείο πραγματοποιείται η εφαρμογή του αλγόριθμου στον εκκαθαρισμένο πλέον τίτλο ώστε να εξαχθούν οι τρεις κυρίαρχες λέξεις κλειδιά της πρότασης. Αφού εφαρμοστεί ο αλγόριθμος, συνεχίζεται το πείραμα ώστε να ελεγχθούν τα αποτελέσματα. Πάλι μέσω του Google Scholar και με αναζήτηση μόνο στην περιοχή του τίτλου εισάγονται οι τρεις κυρίαρχες λέξεις κλειδιά και ελέγχονται τα αποτελέσματα που επιστρέφονται. Διιαπιστώνεται για το παράδειγμα ότι το πρώτο σε κατάταξη αποτέλεσμα είναι ο αρχικός τίτλος στο δείγμα 181 τίτλων.

B. Εφαρμογή του αλγόριθμου

Σύμφωνα με τις ενότητες (*II.A, II.B, III.A*) εφαρμόζεται η προτεινόμενη διαδικασία που περιγράφηκε χρησιμοποιώντας το παραπάνω παράδειγμα σύμφωνα με τα παρακάτω βήματα:

Βήμα 1. Στην προ-επεξεργασία του τίτλου "Advances in knowledge discovery and data mining" το φιλτραρισμένο αποτέλεσμα είναι το εξής "Advances knowledge discovery data mining" σύμφωνα με την ενότητα (*III.A*)

Βήμα 2. Εφαρμόστηκε ο αλγόριθμος χρησιμοποιώντας τις εξισώσεις 2 και 3 για την εξαγωγή των μεταβλητών του V_w. Η εφαρμογή πραγματοποιήθηκε από τον παρακάτω κώδικα (βλέπε παράρτημα Α) μέσω του προγράμματος Matlab και τα αποτελέσματα παρουσιάζονται στον πίνακα Ι που ακολουθεί:

```
num1=double('Advances knowledge discovery data mining');
k1=find (num1==32);
k2=length(k1);
sol1=[];
for i=2:1:k2;
```

```
j=i-1;
t=num1(k1(j)+1:k1(j+1)-1);
sol=[i,length(t),sum(t)];
sol1=[sol1;sol];
end
t=num1(1:k1(1)-1);
solbeg=[1,length(t),sum(t)];
t=num1(k1(k2)+1:length(num1));
sollast=[k2+1,length(t),sum(t)];
sol1=[solbeg;sol1;sollast];
% sol1=Vw word vector
% word normalization vector Vwe
d1=sol1/norm(sol1);
```

Πίνακας Ι. Οι μεταβλητές του V_w

Λέξεις	*Θέση*	*Αριθμός χαρακτήρων*	*Ένταση*
Advances	1	8	805
Knowledge	2	9	960
Discovery	3	9	984
Data	4	4	410
Mining	5	6	642

Βήμα 3. Εφαρμόστηκε ο αλγόριθμος χρησιμοποιώντας την εξίσωση 4 για την εξαγωγή των μεταβλητών της συνισταμένης ($\mathbf{v}_s$). Αυτό έγινε μέσω του παρακάτω κώδικα στο πρόγραμμα Matlab " u=sum(d1)" (βλέπε παράρτημα Α) και τα αποτελέσματα παρουσιάζονται στην εξίσωση (9):

$$\vec{V}_s = \sum_{i=0}^{5} \vec{V}_{we}^{(i)} \vec{V}_s^{(i)} \Big|_{i=0}^{5} = \begin{bmatrix} 0.0085 & 0.0204 & 2.1524 \end{bmatrix} \quad (9)$$

Βήμα 4. Η παραπάνω εφαρμογή, σύμφωνα με τις εξισώσεις 5 και 6 δίνει τα ακόλουθα αποτελέσματα που παρουσιάζονται στον πίνακα ΙΙ.

Πίνακας ΙΙ. Η γωνία απόκλισης ανάμεσα σε συνισταμένη ($\mathbf{v}_s$) και λέξη ($\mathbf{v}_w$)

Λέξεις	*Cosθ*	*Κατάταξη*
Advances	0.1572	3
Knowledge	0.1069	2
Discovery	0.0547	1
Data	0.3332	5
Mining	0.2202	4

Ως αποτέλεσμα αυτής της διαδικασίας οι λέξεις "*discovery, knowledge and Advances"* επιλέγονται ως η βέλτιστη τριάδα λέξεων κλειδιών.

C. Τα αποτελέσματα ανάκτησης

Σύμφωνα με την ενότητα *Εφαρμογής του Αλγορίθμου (IIIB)*, πλέον έχουν εξαχθεί 181 τριάδες λέξεων κλειδιών που αντιστοιχούν σε ισάριθμους τίτλους άρθρων, οι οποίοι ανακτήθηκαν μέσω της μηχανής αναζήτησης Scholar Google. Κατά την διαδικασία της ανάκτησης κάθε μια από τις 181 τριάδες λέξεων κλειδιών εισάγονται ως ερωτήματα στην μηχανή αναζήτησης Scholar Google ώστε να ληφθούν τα αντίστοιχα αποτέλεσμα σε μορφή καταταγμένης λίστας. Επιπλέον ερευνάται εάν ο πρώτος τίτλος που εμφανίζεται στα αποτελέσματα περιέχει την τριάδα λέξεων που έχει εξαχθεί μέσω του αλγορίθμου, στην παραπάνω ενότητα. Περισσότερες λεπτομέρειες για τη σειρά κατάταξης των αποτελεσμάτων και τη βαθμολογία των δοκιμών που θεωρούνται επιτυχημένες για το πείραμα που διεξήχθη κατά τη διαδικασία της ανάκτησης, δίνονται στον πίνακα ΙΙΙ. Να διευκρινιστεί ότι, η επιτυχία στην αναζήτηση αφορά την εμφάνιση του αντίστοιχου τίτλου που αναζητείται και η σειρά αφορά τη θέση στην λίστα αποτελεσμάτων.

Πίνακας ΙΙΙ. Αποτελέσματα ανάκτησης σχετικά με τις 181 τριάδες λέξεων κλειδιών

Σειρά επιτυχίας	*Ερωτήματα*	*Παρουσία κυρίαρχων λέξεων κλειδιών*
1	156	155
2	14	12
3	3	2
4	0	0
5	1	1

6	0	0
7	0	0
8	0	0
9	0	0
10	1	0
11	2	0
12	1	0
13	1	0
14	0	0
15	0	0
16	0	
17	1	

Επιπλέον, αν θεωρηθεί ότι το x συμβολίζει τη σειρά της επιτυχημένης ανάκτησης μέσω της εκάστοτε τριάδας λέξεων και y συμβολίζει τον αριθμό ερωτημάτων, τότε υπάρχει ο πολυωνυμικός βαθμός n:

$$p(x) = p_1 x^n + p_2 x^{n-1} + \ldots + p_n x + p_{n+1} \tag{10}$$

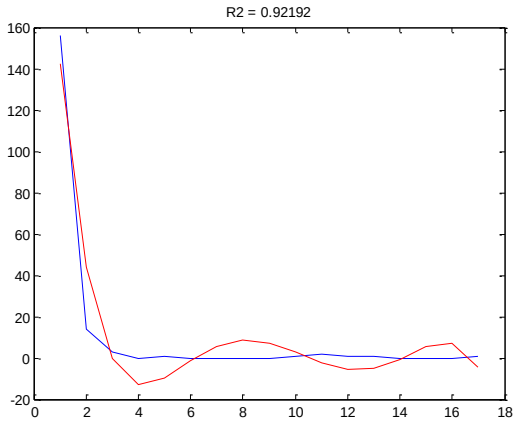
Όπου

$$\min\left(\sum_i \left(p(x_i) - y_i \right)^2 \right) \tag{11}$$

Έπειτα χρησιμοποιώντας τον παρακάτω κώδικα στο πρόγραμμα υπολογιστή Matlab (βλέπε εικ. 2) (<u>βλέπε παράρτημα Α</u>):

```
c = polyfit(x,y,5);
xfit = linspace(min(x),max(x),length(x));
yfit = polyval(c,xfit);
plot(x,y,'o',xfit,yfit,'--')
```

Υπολογίστηκε ότι ο βαθμός προσαρμογής (n=5) του πολυωνύμου στα δεδομένα της μεταβλητής αποτελεί τη βέλτιστη προσαρμογή σύμφωνα με τη μέθοδο ελαχίστων τετραγώνων (βλέπε εικ. 2).



Εικ. 2. Η πολυωνυμική εφαρμογή παλινδρόμησης με $r^2 = 0.9219$

Τότε το πολυώνυμο αυτού του πειράματος καθορίζεται από την εξίσωση (12):

$$p(x) = -0.0071x^5 + 0.3566x^4 - 6.7615x^3 + 59.1588x^2 - 233.5057x + 323.0271 \quad (12)$$

D. Αξιολόγηση αποτελεσμάτων

Για την αξιολόγηση του συνόλου δεδομένων, χρησιμοποιείται η συνάρτηση πυκνότητας πιθανότητας (βλέπε ενότητα της διατριβής 4.1.2). Ο λόγος για τον οποίο χρησιμοποιείται η συνάρτηση αυτή είναι διότι εστιάζει στον καθορισμό των τιμών του μέγιστου πλάτους του διαστήματος $x = [1, \dots, 17]$. Αυτές οι τιμές αντιστοιχούν στη σειρά επιτυχούς ανάκτησης άρθρων μέσω αναζήτησης στο Scholar Google. Με τον τρόπο αυτό, επιδιώκεται ο καθορισμός του παράθυρου διαστήματος για τις τιμές με την υψηλότερη πιθανότητα βαθμολογιών κατά την ανάκτηση. Συνεχίζοντας γίνεται διερεύνηση του διαγράμματος (βλέπε εικ. 2) των δεδομένων με βάση την απαίτηση αντιστοίχισης του με τη μορφή γκαουσιανής συνάρτησης πυκνότητας πιθανότητας (βλέπε ενότητα της διατριβής 4.1.2), μιας που να είναι σχετικά εύκολη η εφαρμογή της. Η γκαουσιανή κατανομή είναι συνήθως χρήσιμη όταν οι τιμές μιας τυχαίας μεταβλητής είναι συγκεντρωμένες κοντά στον αριθμητικό μέσο της, με την πιθανότητα η τιμή να πέφτει κάτω του μέσου να αντιστοιχεί στην πιθανότητα να πέφτει πάνω από το μέσο [18]-[19].

Σε αυτό το στάδιο απεικονίζονται η πυκνότητα πιθανότητας των ακέραιων τιμών του πολυώνυμου με $x = [1, \dots, 17]$ και παρουσιάζονται στον πίνακα IV:

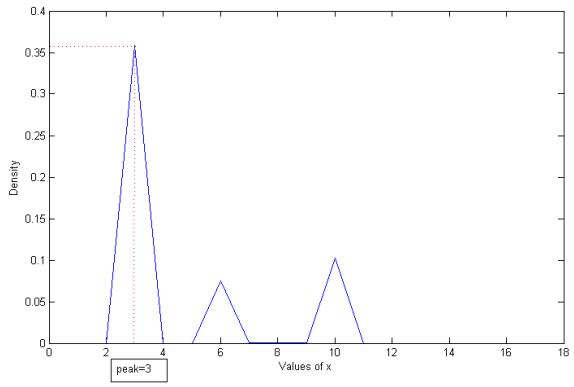
Πίνακας IV. Υπολογισμός αξιών πολυώνυμου

Τιμές x	*Τιμές P*
1	142.2682
2	44.0373
3	-0.4620
4	-13.1717
5	-10.0314
6	-1.8303
7	4.9408
8	7.2375
9	4.6098
10	-1.6499
11	-9.3588
12	-16.2957
13	-21.0530
14	-23.8887
15	-27.5784
16	-38.2673
17	-66.3222

Εάν η $f(x)$ είναι η κανονική κατανομή $\chi \in \{p_1, p_{2,\ldots,} p_{17}\}$ τότε υπολογίζεται η συνάρτηση κανονικής πιθανότητας πυκνότητας (βλέπε εξίσωση 13):

$$\Pr[a \le X \le b] = \int_a^b f_X(x)dx \qquad (13)$$

Για a=1 και b=17 (βλέπε εικ. 3), διαπιστώνεται ότι η πιο πιθανή τιμή της x είναι 3.



Εικ. 3. Η συνάρτηση πυκνότητας πιθανότητας των p τιμών

Από την ερμηνεία του αποτελέσματος αυτού διεξάγεται το συμπέρασμα πως η προτεινόμενη μέθοδος ερωτήματος – τριάδας λέξεων κλειδιών δίνει ως αποτέλεσμα έγκυρη κατάταξη ανάκτησης στις τρεις πρώτες θέσεις.

Επιπλέον, ώστε να αξιολογηθεί η ευαισθησία και η ειδικότητα (βλέπε ενότητα της διατριβής 2.3.3.1) της μεθόδου του άρθρου [20] θεωρείται ότι οι επιτυχημένες ανακτήσεις - βαθμολογίες κυμαίνονται σε διάστημα ανάμεσα από 1 και 3 (βλέπε ενότητα της διατριβής 2.3.4), όπως φαίνεται και από τη συνάρτηση πυκνότητας πιθανότητας (βλέπε εικ. 2). Στην περίπτωση του άρθρου, θεωρούνται ως αληθινές θετικές βαθμολογίες, όλα τα άρθρα τα οποία είχαν θέση ανάκτησης πρώτη, δεύτερη και τρίτη μέσω του ερωτήματος – τριάδας λέξεων κλειδιών. Ως ψεύτικες θετικές βαθμολογίες θεωρούνται τα άρθρα όπου οι κυρίαρχες λέξεις κλειδιά δεν εμφανίζονται στον ανακτημένο τίτλο. Τέλος, αληθινές αρνητικές βαθμολογίες θεωρούνται όλες οι υπόλοιπες βαθμολογίες πέρα από τις πρώτες τρεις θέσεις. Έτσι, ο πίνακας IV μετατρέπεται στον πίνακα V.

Πίνακας V. Αποτελέσματα επιτυχούς ανάκτησης σχετικά με τις 181 τριάδες λέξεων

Αληθινές θετικές	*Ψεύτικες θετικές*	*Αληθινές αρνητικές*
151	11	8
sensitivity	0.90	
specificity	0.58	

Η ευαισθησία σχετίζεται με την ικανότητα του ελέγχου να αναγνωρίζει θετικά αποτελέσματα. Αυτό μπορεί επίσης να γραφτεί ως εξής:

$$sensitivity = \frac{number\_of\_true\_positives}{number\_of\_true\_positives + number\_of\_false\_positives} = \frac{151}{168} = 0.90$$

Η ειδικότητα σχετίζεται με την ικανότητα του ελέγχου να αναγνωρίζει αρνητικά αποτελέσματα. Αυτό μπορεί επίσης να γραφτεί ως εξής:

$$specificity = \frac{number\_of\_true\_negatives}{number\_of\_true\_negatives + number\_of\_false\_positives} = \frac{11}{19} = 0.58$$

IV. ΣΥΓΚΡΙΣΗ TF-IDF ΜΕ ΤΟΝ ΑΛΓΟΡΙΘΜΟ

Στην ενότητα αυτή διεξάγεται ένα πείραμα με σκοπό τη σύγκριση του γνωστού σχήματος απόδοσης βαρών *tf–idf* με τον αλγόριθμο που έχει παρουσιαστεί στις παραπάνω ενότητες.

Αρχικά, για να πραγματοποιηθεί η σύγκριση αυτή, συλλέγεται ένα δείγμα 10 τίτλων επιστημονικών άρθρων. Στη συνέχεια οι τίτλοι αυτοί προ-επεξεργάζονται προκειμένου να αφαιρεθούν οι διακόπτουσες λέξεις, δίνοντας ως αποτέλεσμα προ-επεξεργασμένους τίτλους με πλήθος λέξεων ίσο με εννέα λέξεις για κάθε τίτλο, ώστε να υπάρχει ομοιογένεια στο δείγμα. Στη συνέχεια μέσω του προγράμματος MATLAB, εφαρμόζεται ο αλγόριθμος ώστε να συγκεντρωθούν οι τιμές για το δείγμα και πιο συγκεκριμένα, όλα τα βήματα όπως αναφέρονται στις ενότητες III. Α και III. Β. Έτσι τα αποτελέσματα φαίνονται στον πίνακα VI:

Πίνακας VI. Αποτελέσματα εφαρμογής αλγόριθμου για το δείγμα

Τίτλοι	*Τριάδες αλγόριθμου*	*Cosθ*
T1	applications nanoparticles assembly	0.0283 0.0319 0.0602
T2	production agriculture tools	0.0254 0.0419 0.0693
T3	advances diagnosis applications	0.0539 0.1102 0.1118
T4	synthesis applications properties	0.0179 0.0343 0.0361
T5	life itself subjectivity	0.0989 0.1062 0.1111
T6	throughput molecular techniques	0.0418 0.0631 0.0776
T7	cells technological opportunities	0.0168 0.0274 0.0376
T8	scientific legislative stem	0.0533 0.0884 0.1409
T9	mammography breast diagnosis	0.0226 0.0887 0.1200
T10	hypoxia implications tumor	0.0278 0.0356 0.0635

Στη συνέχεια πραγματοποιείται ο υπολογισμός των τιμών με βάση το σχήμα απόδοσης βάρους *tf-idf*, με τη βοήθεια του προγράμματος MATLAB. Δημιουργείται η μήτρα όρων εγγράφων και συλλέγονται οι όροι (βλέπε ενότητα της διατριβής 2.4.1.1). Έπειτα υπολογίζεται η βαρύτητα κάθε όρου ανά τίτλο, με βάση τους

παράγοντες απόδοσης βάρους του *tf-idf* (βλέπε ενότητα της διατριβής 2.2.2). Στον παρακάτω πίνακα VII, φαίνονται τα αποτελέσματα:

Πίνακας VII. Αποτελέσματα εφαρμογής tf-idf για το δείγμα

Τίτλοι	*Όροι με βάση τη βαρύτητα*	*Συχνότητα*
T1	FePt, assembly, century, nanobiotechnology, scientific / advances	2,05 / 0,25
T2	agriculture, biotechnology, future, new, pig, production, tools / advances	2,05 / 0,25
T3	accuracy, genetic, improve, preimplantation, range/ advances	2,05 / 0,25
T4	aliphatic, nanotechnology, polyester, polymer, properties, stars/ biomedicine	2,05 / 0,41
T5	century, first, itself, life, politics, power, subjectivity, twenty / biomedicine	2,05 / 0,41
T6	high, molecular, nanobiotechnology, systems, techniques, throughput / advances	2,05 / 0,25
T7	aging, concepts, opportunities, regenerative / advances	2,05 / 0,25
T8	admixed, embryos, ethical, human, legislative, scientific / advances	2,05 / 0,25
T9	aided, breast, cancer, computer, detection, mammography / advances	2,05 / 0,25
T10	assessing, hypoxia, implications, methods, planning, treatment, tumor, vivo / advances	2,05 / 0,25

Τέλος, προχωρώντας στη σύγκριση του αλγόριθμου με το *tf-idf* καταλήγουμε στα εξής αποτελέσματα, όπως συνοψίζονται στον πίνακα VIII:

Πίνακας VIII: Κοινοί όροι ανά τίτλο για το δείγμα

Τίτλοι	*Κοινοί όροι*
T1	assembly
T2	production agriculture tools
T3	advances
T4	properties
T5	life itself subjectivity
T6	throughput molecular techniques
T7	opportunities
T8	scientific legislative
T9	mammography breast
T10	hypoxia implications tumor

Για όλο το δείγμα συνολικά, οι όροι που έχουν εξαχθεί με βάση τον αλγόριθμο εμπεριέχονται στο σύνολο των αποτελεσμάτων σύμφωνα με το *tf-idf*. Πιο συγκεκριμένα:

- 40% των αποτελεσμάτων με βάση το *tf-idf*, εμπεριέχουν και τους τρείς όρους εξαγωγής με βάση τον αλγόριθμο.
- 20% των αποτελεσμάτων με βάση το *tf-idf*, εμπεριέχουν τους δύο όρους εξαγωγής με βάση τον αλγόριθμο.
- 40% των αποτελεσμάτων με βάση το *tf-idf*, εμπεριέχουν και έναν από τους όρους εξαγωγής με βάση τον αλγόριθμο.

V. ΣΥΜΠΕΡΑΣΜΑΤΑ

Παρουσιάστηκε ένας καινοτόμος αλγόριθμος εξαγωγής λέξεων κλειδιών που ανιχνεύει τους πιο σημαντικούς όρους στους τίτλους επιστημονικών άρθρων, οι οποίοι στη συνέχεια χρησιμοποιούνται ως λέξεις κλειδιά προκειμένου να ανακτηθούν άρθρα από ψηφιακές επιστημονικές πηγές του διαδικτύου. Πιο συγκεκριμένα, πυρήνα του προβλήματος αποτελεί η δυσκολία αναγνώρισης ενός μεμονωμένου δειγματικού χώρου και μέτρησης πιθανότητας όπου όλες οι μεταβλητές να μπορούν να καθοριστούν. Ο αλγόριθμος του άρθρου βασίζεται σε μια παραλλαγή του VSM για να αναπαραστήσει τίτλους άρθρων και αντιμετωπίζει κάθε όρο σε έναν τίτλο ως διάνυσμα, όπου κάθε διάνυσμα αποτελείται από τρεις μεταβλητές οι οποίες ονομαστικά είναι ο αριθμός των χαρακτήρων που αποτελείται ο όρος, ο αριθμός κωδικοποίησης του όρου στον τίτλο και η σχετική σειρά του όρου στον τίτλο. Στην κλασσική θεωρία του VSM, χρησιμοποιώντας τον σχήμα απόδοσης βαρών σε όρους *tf - idf*, το διάνυσμα του εγγράφου εξάγεται μέσω πολυμεταβλητών συστατικών *j*, ενώ το κλασσικό διάνυσμα έχει διάσταση *j*.

Βασισμένος στα βάρη τα οποία υπολογίζει για κάθε μια από τις παραπάνω μεταβλητές, ο αλγόριθμος του άρθρου αναθέτει σε κάθε όρο στον τίτλο του άρθρου μια βαθμολογία που υποδεικνύει την καταλληλότητα του όρου αυτού ως όρου αναζήτησης και ανάκτησης – λέξης κλειδί που θα μπορούσε να ανακτήσει το αντίστοιχο άρθρο στην κορυφαία θέση κατάταξης. Η πειραματική αξιολόγηση του προτεινόμενου αλγόριθμου σχετικά με πραγματικά δεδομένα απέδειξε την αποτελεσματικότητα του στην ανίχνευση των καταλληλότερων λέξεων κλειδιών από τίτλους άρθρων και υποδεικνύει ότι οι όροι εξαγόμενοι από τίτλους άρθρων επαρκούν για τη σημασιολογική αναπαράσταση ενός άρθρου, όσον αφορά τις επιστημονικές εκδόσεις άρθρων, υπόδειξη η οποία θα μελετηθεί περισσότερο στο μέλλον. Αυτό το διάστημα, γίνεται προσπάθεια εμπλουτισμού του αλγόριθμου με πρόσθετα

χαρακτηριστικά σχετικά με το περιεχόμενο των άρθρων, όπως το abstract και τις λέξεις κλειδιά που υποδεικνύουν οι συγγραφείς για τα άρθρα τους. Ακόμη, διερευνάται η αντιμετώπιση του ζητήματος των συνώνυμων λέξεων και το φαινόμενο της πολυσημίας (βλέπε ενότητα της διατριβής 3.2.3), τα οποία δεν έχουν αντιμετωπιστεί σε αυτό το πείραμα. Επιπλέον, μια επέκταση του παρόντος πειράματος έχει να κάνει με τη χρήση των αναγνωρισμένων λέξεων κλειδιών ώστε να δημιουργηθεί μια οντολογία που θα μπορούσε να χρησιμοποιηθεί ως ψηφιακό αποθετήριο όρων αναζήτησης, ώστε να διεξάγονται εργασίες ανάκτησης κειμένων σε διαδικτυακά επιστημονικά αποθετήρια.

Επιπλέον, προκειμένου να αξιολογηθεί η ευαισθησία της μεθόδου, θεωρήθηκε ότι οι βαθμολογίες επιτυχίας κυμαίνονται σε διάστημα από 1 έως 3, όπως προέκυψαν από τη συνάρτηση πιθανότητας πυκνότητας.

Ακόμη, προκειμένου να τεκμηριωθεί καλύτερα η αξιολόγηση για την αποτελεσματικότητα του αλγόριθμου, πραγματοποιήθηκε ένα πείραμα με σκοπό τη σύγκριση του με το γνωστό σχήμα απόδοσης βαρών *tf-idf*. Ως συμπέρασμα προέκυψε πως το σχήμα *tf-idf* εξαρτάται άμεσα από το συνολικό δείγμα του πληθυσμού, προκειμένου να το αξιολογήσει και να αποδώσει τη βαρύτητα στους όρους, σε αντίθεση με τον αλγόριθμο εξαγωγής λέξεων κλειδιών, ο οποίος θεωρείται αυτάρκης, καθώς μπορεί να λειτουργήσει επιτυχημένα ανά τίτλο, χωρίς να χρειάζεται το σύνολο του δείγματος.

Τέλος, στα μελλοντικά σχέδια του άρθρου είναι η ανακοίνωση της οντολογίας που πρόκειται να αναπτυχθεί, ώστε να μπορεί να είναι αξιοποιήσιμη και από άλλους ερευνητές που ασχολούνται με εργασίες σημασιολογικής ανάκτησης. Η εντροπία και οι σχετικές πληροφορίες δίνουν τη δυνατότητα να καθορίσουμε επίσημα ένα μέτρο απόστασης. Με αυτό το μέτρο απόστασης δίνεται ένα στέρεο θεμέλιο για την αποτύπωση της εγγενούς δομής της οντολογίας. Ως αποτέλεσμα, μέσω της εργασίας αυτής γίνεται απόπειρα να χρησιμοποιηθεί ένας νέος αλγόριθμος, δημιουργημένος βάση της φιλοσοφίας του *VSM* αλγόριθμου, ταυτόχρονα όμως δίνοντας λύση στη δυσαρμονία που παρουσιάζουν τα μοντέλου τύπου Shannon.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] H. Wang, S. Liu, L.T. Chia. “Does ontology help in image retrieval? A comparison between keyword, text ontology and multi-modality ontology

approaches". In *Proceedings of the 14th Annual ACM International Conference on Multimedia*. 2006. pp. 109–112.
[2] Y. Matsuo, M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information". *International Journal on Artificial Intelligence Tools*. 2004. Vol. 13, pp. 157–169.
[3] M. Rajman, R. Besancon "Text mining – knowledge extraction from unstructured textual data". In *Proceedings of the 6th Conference of International Federation of Classification Societies*. 1998.
[4] E. Frank, G.W. Paynter, I.H. Witten, C. Gutwin, C.G. Nevill-Manning. "Domain-specific keyphrase extraction". In *Proceedings of IJCAI*, 1999, pp. 688–673.
[5] Y.H. Kerner, Z. Gross, A. Masa, "Automatic extraction and learning of keyphrases from scientific articles". *Computational Linguistics and Intelligent Text Processing*, 2005, pp. 657–669.
[6] Liu F. Liu, Y Liu. "Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion". *In Proceedings of IEEE SLT*.
[7] P. Turney, "Coherent keyphrase extraction via web mining". In *Proceedings of IJCAI*, 2003, pp. 434–439.
[8] G. Carenini, R.T. Ng, X. Zhou, "Summarizing emails with conversational cohesion and subjectivity," in *Proceedings of ACL/HLT*, 2008.
[9] A. Janin, D. Baron, J. Edwards, D. Ellis, G. Gelbart, N. Norgan, B. Pe- skin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," in *Proceedings of ICASSP*, 2003.
[10] Matsuo, Y., Ishizuka, M. "Keyword extraction from a single document using word co-occurrence statistical information". International Journal on Artificial Intelligence Tools. 2004. Vol 13, pp. 157–169.
[11] Hulth, A., "Improved automatic keyword extraction given more linguistic knowledge". In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. 2003. Association for Computational Linguistics, pp. 216–223
[12] P. Soucy και G. W. Mineau, 'Beyond TFIDF weighting for text categorization in the vector space model'. IJCAI, 2005, Vol. 5, pp 1130–1135.
[13] Robertson, S. "Understanding Inverse Document Frequency: On theoretical arguments for IDF". Journal of Documentation, 2004, Vol. 60 (5), pp. 503–520.
[14] C. Blake, 'A comparison of document, sentence, and term event spaces', In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006, pp. 601–608.
[15] M. Poulos, S. Papavlasopoulos, V. Chrissikopoulos, "A text categorization technique based on a numerical conversion of a symbolic expression and an onion layers algorithm". Journal of Digital Information, 2006, Vol. 1 (6).
[16] Salton, G., Wong, A., Yang, C. S. "A vector space model for automatic indexing". Communications of the ACM. 1975. Vol. 18, pp. 613–620.
[17] Harispe S. et al. "Semantic measures for the comparison of units of language, concepts or entities from text and knowledge base analysis". Arxiv. 2013. Vol. 1310, 1285. pp. 1-159.
[18] Parzen, E. "On estimation of a probability density function and mode". *The annals of mathematical statics*. 1962. Vol. 33, pp. 1065-1076
[19] Abramowitz, M., Stegun, I. A. (Eds.). "Probability Functions." Ch. 26 in Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing. New York: Dover, 1972, pp. 925-964,

[20] M. Poulos, G. Bokos, N. Kanellopoulos, S. Papavlasopoulos and M. Avlonitis. “Specific selection of FFT amplitudes from audio sports and news broadcasting for classification purposes”. *Journal of Graph Algorithms and Applications*. 2007. Vol. 11(1). pp. 277–307

5.3 Υπόθεση συνεκτικότητας μικρών κειμένων

Abstract: Σε αυτό το άρθρο, μελετάται πειραματικά ο βαθμός στον οποίο το μήκος ενός μικρού κειμένου επηρεάζει την κατανόηση και την αναγνωσιμότητα (βλέπε ενότητα της διατριβής 3.1.5) του, στο πλαίσιο της Ποσοτικής Γλωσσολογίας (ΠΓ) (βλέπε ενότητα της διατριβής 3.3). Η ΠΓ εστιάζει κυρίως στην ανάλυση συλλογών αποτελούμενων από μεγάλα κείμενα και μια από τις κυριότερες επιστημονικές θεωρίες που χρησιμοποιούνται είναι αυτή του νόμου των Menzerath – Altmann (βλέπε ενότητα της διατριβής 3.3.1.1). Σε αυτό το άρθρο γίνεται μια απόπειρα ορισμού του πλαισίου ποσοτικής ανάλυσης για μικρά κείμενα, τα οποία αποτελούνται περίπου από μια ή δύο προτάσεις, λόγω του ότι θεωρούνται πολύ σημαντικά σε πολλά επιστημονικά πεδία. Για να επιτευχθεί ο στόχος του άρθρου αυτού δημιουργήθηκε μια στατιστική διαδικασία ελέγχου συνεκτικότητας, μέσω τριών μεταβλητών για μικρά κείμενα. Η εφαρμογή αυτού, πραγματοποιήθηκε μέσω πειραματικής και στατιστικής αξιολόγησης. Με την ολοκλήρωση της παραπάνω αναφερόμενης αξιολόγησης, τα στατιστικά αποτελέσματα απέδειξαν ότι η συνεκτικότητα μικρών κειμένων, η κατανόηση και η αναγνωσιμότητα επιτυγχάνονται πλήρως σε μικρά κείμενα αποτελούμενα από 14 λέξεις, όταν οι τρεις προκαθορισμένες μεταβλητές συσχετίζονται και αντιστρόφως. Για την απόδειξη της παραπάνω υπόθεσης χρησιμοποιήθηκαν: το μοντέλο αναπαράστασης εγγράφων VSM (βλέπε ενότητα της διατριβής 2.2.1) και ο δείκτης συμφωνίας w του Kendall (βλέπε ενότητα της διατριβής 4.2.3). Η αξιολόγηση των στατιστικών αποτελεσμάτων κατέληξε ότι η παραπάνω υπόθεση ισχύει για ένα σύνολο περιπτώσεων με πιθανότητα p>99%. Ακόμη στο πείραμα χρησιμοποιήθηκαν μικρά κείμενα στην αγγλική γλώσσα, ενώ στην συνέχεια αποδείχτηκε μέσω πειράματος ότι η γλώσσα του κειμένου τελικά δεν παίζει σημαίνοντα ρόλο. Για την επιβεβαίωση αυτού λοιπόν διεξήχθη ένα μικρότερης κλίμακας πείραμα με μικρά κείμενα στη γερμανική γλώσσα και η παραπάνω υπόθεση επιβεβαιώθηκε, δηλαδή ότι το προτεινόμενο μοντέλο σε αυτό το άρθρο μπορεί να εφαρμοστεί σε όλα τα μικρά κείμενα ασχέτως της γλωσσικής τους προέλευσης.

Keywords: Short Text Processing, Vector Space Model, Lexical Coherence

Ι. ΕΙΣΑΓΩΓΗ

Στην θεωρία ΠΓ, η λεξιλογική συνεκτικότητα (βλέπε ενότητα της διατριβής 3.1.2) των κειμένων ασχέτως της κατανομής λέξεων θεωρείται ένα πολύ σημαντικό επιστημονικό πεδίο. Σύμφωνα με τον Carstens (2001), η γλωσσική συνεκτικότητα κειμένων ορίζεται ως «*οι τρόποι με τους οποίους τα συστατικά των προτάσεων ενός κειμένου δηλαδή, οι λέξεις που ακούμε και χρησιμοποιούμε είναι αμοιβαία συνδεδεμένες μεταξύ τους (γραμματικά και λεξιλογικά)*». Οι Haliday και Hasan (1976) υπογραμμίζουν τη συνοχή ως «*τη σημασιολογική σχέση μεταξύ ενός στοιχείου και ενός άλλου μέσα στο κείμενο και κάποιου άλλου που θεωρείται κρίσιμο για την ερμηνεία*

τους». Επιπλέον, όπως αναφέρεται από Fahnestock (1983), η συνεκτικότητα, η οποία πηγάζει από σωστή σύνθεση κειμένου, είναι ένα κρίσιμο στοιχείο ώστε η ροή των ιδεών να πραγματοποιείται ομαλά σε όλο το κείμενο και ώστε η αναγνωσιμότητα να διατηρεί υψηλά επίπεδα για την κατανόηση του νοήματος του κειμένου. Με βάση τους Richards et al. (1992), ως αναγνωσιμότητα ορίζεται: «*πόσο εύκολα το γραπτό υλικό μπορεί να αναγνωσθεί και να κατανοηθεί. Αυτό εξαρτάται από πολλούς παράγοντες συμπεριλαμβανομένου και του μέσου μήκους μιας πρότασης, τον αριθμό καινούριων λέξεων που περιλαμβάνονται σε αυτά και την γραμματική πολυπλοκότητα της γλώσσας που χρησιμοποιείται στο απόσπασμα κειμένου»*. Ακόμη η έννοια της αναγνωσιμότητας συνδέεται με αυτήν της κατανόησης. Ο Sparks (2012) αναφέρει ότι η κατανόηση λόγου εμπεριέχει την κατασκευή νοήματος από εκτεταμένα τμήματα της γλώσσας. Επιπλέον ότι η επιτυχής κατανόηση μεγαλύτερων μονάδων κειμένου απαιτεί τη λήψη συμπερασμάτων ώστε να συνδεθούν οι ιδέες τόσο στο κείμενο τοπικά όσο και σε παγκόσμιο πλαίσιο λόγου.

Όπως αναφέρεται από τον Eroglu (2013) η γλωσσολογική οργάνωση στα κείμενα μπορεί να επιτευχθεί όταν ισχύει ο νόμος των Menzerath-Altmann (νόμος ΜΑ) (Altmann 1980). Ο νόμος ΜΑ θεωρείται βασικός νόμος στην ΠΓ, όπου σύμφωνα με Baixeries et al. (2013) παρατηρούνται οι σχέσεις ανάμεσα στο μέγεθος του συνόλου και στο μέγεθος των επιμέρους μερών της γλώσσας. Έτσι σύμφωνα με Eroglu S. (2013), για τον νόμο ΜΑ, όσο μεγαλύτερο είναι το γλωσσικό δόμημα, τόσο μικρότερα θα είναι τα συστατικά που το αποτελούν, όπου δόμημα θεωρείται το σύνολο και συστατικό το μέρος του συνόλου. Ο νόμος ΜΑ είναι βασικός και σημαντικός νόμος στην ΠΓ και εστιάζει κυρίως στη στατιστική ανάλυση μεγάλων κειμένων. Ειδικότερα, σύμφωνα με τον Hrebicek (2002), ο νόμος ΜΑ δεν ισχύει σε εξαιρετικά μικρά κείμενα, όταν μικρά κείμενα θεωρούνται απλές ή περίπλοκες προτάσεις. Όμως τα μικρά κείμενα θεωρούνται πολύ σημαντικά σε πολλούς επιστημονικούς τομείς, όπως στην διαδικτυακή επικοινωνία και το ηλεκτρονικό εμπόριο αλλά και στην γρήγορη αναζήτηση στο διαδίκτυο. Σύμφωνα με τους Ge Song et al. (2014) θεωρείται μεγάλη πρόκληση η ταξινόμηση μικρών κειμένων καθώς ο περιορισμένος αριθμός λέξεων τους δεν μπορεί να αναπαραστήσει ούτε το χώρο χαρακτηριστικών ούτε τις πραγματικές σχέσεις μεταξύ λέξεων και εγγράφων. Καθώς τα μικρά κείμενα έχουν μικρό μήκος σύμφωνα με τον Xiaojun Quan (2009) τα κλασικά μέτρα ομοιότητας (βλέπε ενότητα της διατριβής 3.2.3.2) και οι αντίστοιχες αναπαραστάσεις των αντικειμένων τους (βλέπε ενότητα της διατριβής 2.2.1) δεν μπορούν να εφαρμοστούν με επιτυχία.

Είναι γνωστό πως οι άνθρωποι χρησιμοποιούν προτάσεις για να επικοινωνούν μεταξύ τους επιτυχημένα και ορισμένες παράμετροι πρέπει να ληφθούν υπόψη, καθώς η σωστή δόμηση προτάσεων αποτελεί ένα πολύ σημαντικό ζήτημα. Προκειμένου η επικοινωνία να είναι επιτυχής πρέπει να υπάρχει ένα μέσο μήκος πρότασης ώστε αυτή να μην προκαλεί σύγχυση ή να μην θεωρείται περίπλοκη. Η λεξιλογική συνεκτικότητα της πρότασης έχει ορισθεί εμπειρικά από Kornai (2008), στο δημοσιογραφικό λόγο με μέσο μήκος πρότασης να είναι πάνω από 15 λέξεις. Ακόμη η Tesitelova (1992) αναφέρει ότι κατά την δεκαετία του ‘80 υπήρξε ενδιαφέρον στην επιστημονική έρευνα της πρότασης ως συντακτικού φαινομένου.

Επιπλέον, η πολυπλοκότητα της σημασίας της πρότασης, εξαρτάται από το μήκος της πρότασης. Σύμφωνα με τον Cutts (2009) το μέσο μήκος πρότασης πρέπει να βρίσκεται ανάμεσα σε 15-20 λέξεις, ώστε να διατηρεί την αναγνωσιμότητα αν και το μέσο μήκος δεν είναι δυνατόν να επιτυγχάνεται πάντα. Οι Taskar et al. (2004) στην εργασία τους σχετικά με περιγραφική ανάλυση και τεχνικές δυναμικών προσεγγίσεων, επιτελούν τα πειράματα τους θέτοντας τον περιορισμό του μήκους πρότασης να είναι ίσο ή λιγότερο από 15 λέξεις. Άλλα επιστημονικά πεδία που εκφράζουν ενδιαφέρον σχετικά με το μήκος της πρότασης αφορούν την έρευνα για την ανθρώπινη μνήμη, όπως η Γνωστική Ψυχολογία και οι Νευροεπιστήμες. Σύμφωνα με τον Baddeley (2003), στο μοντέλο του για τη μνήμη εργασίας που αποτελείται από τρία τμήματα, αντιμετώπισε προβλήματα σχετικά με την αλληλεπίδραση με τη μακροπρόθεσμη μνήμη, όπου ο περιορισμός των 15 λέξεων αναφέρεται ξανά. Στα πειράματα τους οι Daveman και Carpenter (1980) υποστηρίζουν ότι χρειάζονται 5 δευτερόλεπτα ώστε ένας άνθρωπος να ολοκληρώσει την ανάγνωση μιας πρότασης. Επιπλέον, οι Anderson et al. (2001) στην μελέτη τους σχετικά με τη μνήμη προτάσεων, περιγράφουν διάφορα μοντέλα μνήμης και παρατηρούν ότι ο ανθρώπινος εγκέφαλος χρειάζεται μερικές εκατοντάδες χιλιοστά του δευτερολέπτου για να επεξεργαστεί μια λέξη. Στην ίδια έρευνα αναφέρεται η διεξαγωγή του πειράματος του Zimny (1987), όπου μια διαδικασία παρουσίασης λέξη προς λέξη ολοκληρώθηκε σε 300 χιλιοστά του δευτερολέπτου ανά λέξη. Συνεπώς από τις παραπάνω έρευνες μπορεί κανείς να συμπεράνει ότι μια τυπική πρόταση θα πρέπει να περιέχει 16-17 λέξεις. Λαμβάνοντας υπόψη όλα τα παραπάνω η ανάλυση της συνεκτικότητας (βλέπε ενότητες της διατριβής 3.1.1 και 3.1.2) στα μικρά κείμενα και ειδικότερα στις προτάσεις μπορεί να θεωρηθεί ένα πολύ σημαντικό επιστημονικό πεδίο για διερεύνηση.

Ο σκοπός αυτού του άρθρου είναι η καθιέρωση μιας στατιστικής υπόθεσης που επιβεβαιώνει τις εμπειρικές παρατηρήσεις σχετικά με τη συνεκτικότητα των μικρών κειμένων ή μιας πρότασης. Έτσι μπορεί να καλυφθεί το κενό που αφήνει ο νόμος ΜΑ σχετικά με τα μικρά κείμενα και η μελέτη αυτή μπορεί να αποτελέσει το θεμέλιο λίθο για την ανάλυση μικρών κειμένων.

Για την εφαρμογή των παραπάνω, ανιχνεύθηκαν μεταβλητές που θεωρούνται κρίσιμες για τη συνεκτικότητα κειμένου. Η συνεκτικότητα κειμένου εξετάστηκε σε σχέση με το αντίκτυπο κάθε συστατικού σε σχέση με το δόμημα. Η συσχέτιση αυτή πραγματοποιήθηκε με τη χρήση τριών μεταβλητών. Η απόδειξη της παραπάνω υπόθεσης πραγματοποιήθηκε μέσω του δείκτη συμφωνίας w του Kendall. Τα συμπεράσματα που προέκυψαν από αυτήν τη μεθοδολογία, εισάγουν μια καινοτομία για το πεδίο της ΠΓ διότι οι παραπάνω πειραματικές γλωσσολογικές παρατηρήσεις σχετικά με τη σταθερότητα των μικρών κειμένων στα περίπου 15 συστατικά επιβεβαιώνονται, γεγονός που καταδεικνύει τη στατιστική συσχέτιση των μικρών κειμένων μέσω αυτών των τριών μεταβλητών.

Το άρθρο αυτό χωρίζεται στις παρακάτω ενότητες:

- Μεθοδολογία, όπου παρουσιάζεται εκτενώς ο αλγόριθμος και η στατιστική αξιολόγηση του.

- Πειραματικό μέρος, όπου ο αλγόριθμος εφαρμόζεται σε ευρύ δείγμα μικρών κειμένων.

ΙΙ. ΜΕΘΟΔΟΛΟΓΙΑ

Α. VSM

Σύμφωνα με τους Salton, Wong and Yang (1975) τα έγγραφα μπορούν να αναπαρασταθούν ως διανύσματα, ώστε να ευρετηριαστούν και να υπολογιστεί ένας βαθμός ομοιότητας μεταξύ τους (βλέπε ενότητες της διατριβής 2.2.1 και 3.2.3.2). Όπως αναφέρουν οι Raghavan και Wong (1986) τα διανύσματα είναι ιδιαιτέρως χρήσιμα καθώς υπακούν σε βασικά αξιώματα και κανόνες της άλγεβρας. Το VSM χρησιμοποιείται σε διάφορα επιστημονικά πεδία όπως συστήματα φιλτραρίσματος πληροφορίας και αλγόριθμοι κατάταξης με βάση τη συνάφεια (βλέπε ενότητα της διατριβής 1.1). Οι Turney και Patel (2010) αναφέρουν πως με την επέκταση της χρήσης του VSM σε σημασιολογικό πλαίσιο στην επεξεργασία της γλώσσας, μπορούν να επέλθουν λαμπρά αποτελέσματα. Το VSM είναι ένα αλγεβρικό μοντέλο το οποίο καθιστά δυνατή την αναπαράσταση οποιουδήποτε αντικειμένου εγγράφου (βλέπε ενότητα της διατριβής 2.1), όπως ένα κείμενο, μια πρόταση, μια φράση, μια λέξη ή ένα μόρφημα. Η αναπαράσταση μέσω VSM αναλύεται σε τρία βήματα.

Στο πρώτο βήμα, οι όροι που περιέχονται σε ένα κείμενο (τυπικά λέξεις ή μικρές φράσεις) εξάγονται δημιουργώντας ένα ευρετήριο εγγράφου. Η ευρετηρίαση αυτή εκτελείται μέσω δύο εναλλακτικών μεθόδων. Η γλωσσολογική μέθοδος βασίζεται στην συγκέντρωση λειτουργικών λέξεων, οι οποίες έχουν υψηλή συχνότητα και χαμηλή συχνότητα, που αντικατοπτρίζονται σημασιολογικά στο κείμενο. Από την άλλη ή μη γλωσσολογική μέθοδος βασίζεται σε διαφορετικές διαδικασίες ευρετηρίασης όπως πιθανοτική και αυτόματη ευρετηρίαση.

Στο δεύτερο βήμα, τα βάρη των ευρετηριαζόμενων όρων καθορίζονται σύμφωνα με τη συνάφεια τους με το ερώτημα του χρήστη για μια πιθανή διαδικασία ανάκτησης πληροφορίας. Τα βάρη όρων εφαρμόστηκαν ελέγχοντας την ευαισθησία – ειδικότητα της αναζήτησης, όπου η ειδικότητα σχετίζεται με την ακρίβεια και η ευαισθησία με την ανάκληση. Το πιο ευρέως χρησιμοποιούμενο σχήμα απόδοσης βάρους στην ΑΠ είναι το *tf - idf*, το οποίο περιέχει τρεις κυρίαρχες μεταβλητές για τον καθορισμό του βάρους, που συνδέονται με τη συχνότητα όρου στο έγγραφο, τη συχνότητα όρου στη συλλογή εγγράφων και την κανονικοποίηση μεγέθους (βλέπε ενότητα της διατριβής 2.2.2). Αυτές οι τρείς μεταβλητές συνδυάζονται μέσω της πράξης του πολλαπλασιασμού, ώστε να υπολογίζουν το αποτέλεσμα για το βάρος του όρου.

Ως παράδειγμα, ο ορισμός της αλγεβρικής έκφρασης ενός εγγράφου μέσω VSM πραγματοποιείται χρησιμοποιώντας την ακόλουθη εξίσωση:

$$\vec{V}_j = \left[w_{1,j}, w_{2,j} ..., w_{t,j} \right] \quad (1)$$

Όπου το j αντιστοιχεί στον αριθμό συστατικών ενός εγγράφου και το t αναπαριστά τον αριθμό των μεταβλητών που καθορίζουν το βάρος στο προτεινόμενο μοντέλο.

Τέλος, στο τρίτο βήμα το κείμενο κατατάσσεται σύμφωνα με το μέτρο ομοιότητας (βλέπε ενότητα της διατριβής 3.2.3.2) ως προς ένα αντίστοιχο ερώτημα. Η ομοιότητα στο VSM μοντέλο υπολογίζεται μέσω συνδετικών μεταβλητών και βασίζεται στο κανονικοποιημένο εσωτερικό γινόμενο ανάμεσα στο διάνυσμα εγγράφου και διάνυσμα ερωτήματος, όπου η επικάλυψη μεταξύ διανυσμάτων υποδεικνύει ομοιότητα. Το εσωτερικό γινόμενο συνήθως κανονικοποιείται. Το πιο συνηθισμένο μέτρο ομοιότητας είναι μέσο της μεταβλητής συνημίτονου, όπου μετράται η γωνία απόκλισης ανάμεσα στο διάνυσμα εγγράφου και διάνυσμα ερωτήματος.

Β. Η βάση του αλγόριθμου

Η εξαγωγή μεταβλητών βασίζεται στη θεωρία του VSM. Πιο συγκεκριμένα, στο πρώτο βήμα (βλέπε ενότητα ΙΙ.Α) επιλέγεται ένα μικρό κείμενο ως δόμημα, το οποίο αποτελείται από μια απλή ή σύνθετη πρόταση και η λέξη θεωρείται ως συστατικό του δομήματος. Κατά συνέπεια, σύμφωνα με το δεύτερο βήμα (βλέπε ενότητα ΙΙ.Α) μια μη γλωσσική προσέγγιση υιοθετείται γιατί όλα τα συστατικά του δομήματος χρησιμοποιούνται σε μια διαδικασία ευρετηρίασης που εξαρτάται από τη σειρά της θέσης τους. Πιο λεπτομερειακά, κάθε συστατικό λαμβάνει ένα αντίστοιχο βάρος ανεξάρτητα από τον αριθμό εμφάνισης του στο δόμημα.

Στη παρούσα περίπτωση, οι τρεις κυρίαρχες μεταβλητές απόδοσης βάρους όρου αντικαθίστανται από τις μεταβλητές του παρακάτω διανύσματος, όπως φαίνεται τύπο (2):

$$\vec{w}_j = \left[i_j \quad s_j \quad k_j \right] \tag{2}$$

Το *i* αντιστοιχεί στη σειρά του συστατικού στο μικρό κείμενο, το *s* αποτελεί έναν αριθμό κωδικοποίησης, όπου για λόγους απλοποίησης ορίζεται μέσω της διαδικασίας κωδικοποίησης ASCII (Poulos, Papavlasopoulos, Chrissikopoulos 2006) και το *k* αντιστοιχεί στον αριθμό χαρακτήρων του όρου.

Έπειτα η εξίσωση (1) μετατρέπεται στην εξίσωση (3):

$$\vec{V}_j = \left[w_1, w_2 ..., w_j \right] \tag{3}$$

Για λόγους κανονικοποίησης, το διάνυσμα $\vec{V}_j$ υπολογίζεται από το ισοδύναμο διάνυσμα της εξίσωσης (4):

$$\vec{F}_j = \frac{\vec{V}_j}{\left\| \vec{V}_j \right\|} \tag{4}$$

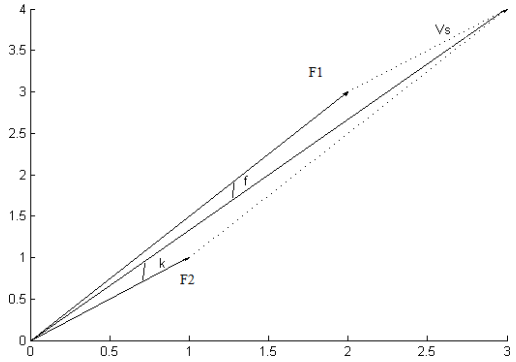
Επιπλέον, ορίζεται το διάνυσμα $\vec{V}_s$, το οποίο απεικονίζεται ως η συνισταμένη διανύσματος και από εδώ και στο εξής θα αναφέρεται ως διάνυσμα μικρού κειμένου (βλέπε εικ. 1).

$$\vec{V}_s = \sum_j \vec{F}_j \tag{5}$$

C. Κριτήριο ομοιότητας

Ο βαθμός συσχέτισης ανάμεσα σε $\vec{F}_j$ και $\vec{V}_s$ εξάγεται μέσω της εξίσωσης 6 και ειδικότερα από το συντελεστή *r* ο οποίος αναπαριστά το εσωτερικό γινόμενο ανάμεσα στο διάνυσμα κειμένου $\vec{V}_s$ και στο διάνυσμα συστατικού $\vec{F}_j$ σύμφωνα με το τρίτο βήμα (βλέπε ενότητα II.A). Ακόμη, η διαδικασία αυτή εκφράζεται από την γενική θεώρηση σχετικά με τη θεωρία Ομοιότητας Εγγράφων (Harispe et al. 2013) (βλέπε ενότητα της διατριβής 3.2.3.2):

$$r^{(j)} = \cos\theta^{(j)} = \frac{\vec{F}_j \vec{V}_s}{\left\|\vec{F}_j\right\| \left\|\vec{V}_s\right\|} \tag{6}$$



Εικ.1 Απεικόνιση διανυσμάτων συστατικών $\vec{F}_1, \vec{F}_2$ και διανύσματος συνισταμένης $\vec{V}_s$ στο Ευκλείδειο επίπεδο

D. Μετατροπή μεταβλητών

Υιοθετείται το διάνυσμα $\vec{U}_j$ αντί για το $\vec{V}_j$, όπου η μεταβλητή *i* του $\vec{w}_j$ αντικαθίσταται από τη μεταβλητή *r* του $\vec{U}_j$ (βλέπε εξίσωση 7). Ο λόγος για αυτήν την αντικατάσταση είναι διότι η μεταβλητή *r* θεωρείται ένας πολύ σημαντικός συντελεστής στη θεωρία Σημασιολογίας σύμφωνα με (Harispe et al. 2013). Η αντικατάσταση πραγματοποιήθηκε σκοπίμως αφού το διάνυσμα $\vec{U}_j$ αναπαριστά το βαθμό επηρεασμού μέσω της μεταβλητής *r* γιατί αυτό θα οδηγήσει σε μια νέα διαδικασία κατάταξης ανάλογα τον επηρεασμό της.

$$\vec{U}_j = \begin{bmatrix} r_j & s_j & k_j \end{bmatrix} \tag{7}$$

Ε. Στατιστική θεμελίωση

Η στατιστική θεμελίωση της παραπάνω μελέτης ολοκληρώνεται σε τρία στάδια. Στο πρώτο στάδιο a υπολογίζεται ο βαθμός επηρεασμού της μεταβλητής *r* στις άλλες δύο μεταβλητές που ορίζεται στην εξίσωση (7), σε σχέση με τον αριθμό λέξεων του κάθε παραδείγματος. Όλη αυτή η διαδικασία έχει περιγραφεί στη θεωρία (βλέπε ενότητες II.C και II.D) και υλοποιείται παρακάτω στο στάδιο a. Στη συνέχεια, στο δεύτερο στάδιο b χρησιμοποιείται το κριτήριο χ^2 συσχετισμένο με το δείκτη συμφωνίας w του Kendall, ώστε στο τρίτο στάδιο c να υπολογιστεί ο βαθμός πιθανότητας να υπάρξει συμφωνία ή μη συμφωνία με βάση το κριτήριο χ^2. Αυτά περιγράφονται αναλυτικά παρακάτω, στο στάδιο b και c αντίστοιχα και συνδέονται με τη θεωρία (βλέπε ενότητες της διατριβής 4.2.1 και 4.2.3):

a. Η στατιστική επιβεβαίωση του παραπάνω μετασχηματισμού επιτυγχάνεται ελέγχοντας εάν οι τρεις μεταβλητές r, s, k συσχετίζονται. Αυτό πραγματοποιείται εφαρμόζοντας το δείκτη συμφωνίας w του Kendall (Zar 1999), στον παρακάτω τύπο 8 (βλέπε ενότητα της διατριβής 4.2.3):

$$W = \frac{\sum_{j=1}^{n} R_j^2 - \frac{\sum_{j=1}^{n} R_j}{n}}{M^2\left(n^2 - 1\right)} \tag{8}$$

Το M αναπαριστά τον αριθμό μεταβλητών που συσχετίζονται, δηλαδή σε αυτή την περίπτωση το M=3 και το n αναπαριστά τον αριθμό συστατικών στο μικρό κείμενο.

$$\text{Όπου} \qquad \left[d\right] = rank(U_j) \tag{9}$$

$$\left[R\right]_{j=1}^{n} = \sum_{j=1}^{n}\left[d\right] \tag{10}$$

Το d αντιστοιχεί στην κατάταξη του κειμένου του διανύσματος $\vec{U}_j$ (εξίσωση 9) και το R ορίζει την κατάταξη ως το άθροισμα των μεταβλητών κάθε εγγράφου (εξίσωση 10).

b. Σε αυτή την περίπτωση ο στατιστικός έλεγχος βασίζεται στη μηδενική υπόθεση (βλέπε ενότητα της διατριβής 4.1.3) ότι αυτές οι τρείς μεταβλητές δεν συσχετίζονται μεταξύ τους, ώστε να οριστεί ο αριθμός συστατικών που επηρεάζουν το συμμετέχων μικρό κείμενο. Η μηδενική υπόθεση H_o υποδεικνύει ότι οι τρείς μεταβλητές δε συσχετίζονται και η εναλλακτική υπόθεση H_A υποδεικνύει το ακριβώς αντίθετο. Κατά συνέπεια, το κριτήριο χ^2 ορίζεται ως εξής:

$$\chi_r^2 = 3*(n-1)*W(r,s,k) \tag{11}$$

c. Η τιμή της αθροιστικής κατανομής του χ^2 εξάγεται χρησιμοποιώντας το βαθμό ελευθερίας (βλέπε ενότητα της διατριβής 4.2.1) $\nu = n - 1$ και υπολογίζεται από την εξίσωση 12.

$$P = 1 - \int_0^x \frac{t^{\frac{\nu-2}{2}} e^{-\frac{t}{2}}}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)} \tag{12}$$

Ο έλεγχος είναι μονόπλευρης κατανομής (βλέπε ενότητα της διατριβής 4.1.3) διότι αναζητάται ο κατάλληλος αριθμός συστατικών για ένα μικρό κείμενο. Έπειτα η μηδενική υπόθεση H_o γίνεται δεκτή/ισχύει όταν $P < 0.001$ και κατά συνέπεια το H_A ικανοποιείται με το $P \geq 0.001$.

ΙΙΙ. ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ

Αυτό το πείραμα αποτελείται από τα παρακάτω βήματα:

1. Αρχικά συλλέχθηκε ένα δείγμα ώστε να εφαρμοστεί ο αλγόριθμος. Το σύνολο δεδομένων αυτής της έρευνας αποτελείται από 100 μικρά κείμενα το οποία εξάχθηκαν από περιλήψεις (abstracts) επιστημονικών άρθρων. Τα επιστημονικά αυτά άρθρα ανακτήθηκαν σε μορφή *pdf* από το Directory of Open Access Journals (DOAJ) και σε σχέση με το θέμα τους όλα προέρχονται από το επιστημονικό πεδίο της βιοιατρικής. Με την χρήση ενός φυλλομετρητή, από τον ιστότοπο του Directory of Open Access Journals (DOAJ) αποθηκεύτηκαν σε τοπικό φάκελο τα 100 άρθρα.
2. Εν συνεχεία, επιλέχθηκαν 100 μικρά κείμενα, όπου κάθε ένα από αυτά είχε μέσο μήκος πάνω από 25 συστατικών. Κάθε μικρό κείμενο επεξεργάστηκε μέχρι το 25^{ο} συστατικό και όχι πιο πάνω για τους παρακάτω λόγους:
    - Καθώς ο αλγόριθμος αυτός βασίζεται στο VSM, υπάρχει περιορισμός σχετικά με τα μεγάλο κείμενα. Η αναπαράσταση τους δεν μπορεί να είναι επιτυχής καθώς σε τέτοια μεγέθη δεν μπορούν να βρεθούν μεταβλητές ομοιότητας σύμφωνα με τον Salton (1975).
    - Ένας ακόμη λόγος για το όριο των 25 συστατικών προέκυψε μέσα από την πειραματική διαδικασία, η οποία έδειξε ότι για μήκος μεγαλύτερο των 25 συστατικών δεν υπήρχε ουσιαστική διαφορά στα αποτελέσματα.
    - Τέλος, ο περιορισμός αυτός ισχύει για όλο το δείγμα δεδομένων, καθώς πρέπει να υπάρχει ομοιογένεια ώστε να μπορούν να διεξαχθούν έγκυρα και αντικειμενικά συμπεράσματα σε ένα γενικό επίπεδο.

Εν τέλει, η σύνταξη του αλγόριθμου και η διαδικασία εφαρμογής του διεξάχθηκαν μέσω του υπολογιστικού προγράμματος MATLAB.

Α. Εφαρμογή του αλγόριθμου και στατιστική επεξεργασία

Σε αυτήν την ενότητα θα παρουσιαστεί η εφαρμογή του προτεινόμενου αλγόριθμου χρησιμοποιώντας ως παράδειγμα το παρακάτω μικρό κείμενο (βλέπε πίνακα 1):

"Many theorists have suggested that working memory capacity plays a crucial role in reading comprehension however, traditional measures of short-term memory, like digit span and word span, are either not correlated or only weakly correlated with reading ability"

Αρχικά μέσω των εξισώσεων (2) – (6), εξάγονται τα δεδομένα των μεταβλητών (r,s,k) (βλέπε εικ. 7). Στη διαδικασία μέσω του δείκτη συμφωνίας w του Kendall τα δεδομένα των μεταβλητών (r, s, k) κατατάσσονται και υπολογίζονται σε σειρά κατάταξης αθροισμάτων (Rj), για κάθε συστατικό μέσω των εξισώσεων (8) – (10).

Για παράδειγμα το συστατικό *many*, το οποίο αποτελεί την πρώτη λέξη του παραπάνω παραδείγματος, αντιστοιχεί σε ένα διάνυσμα με γωνία απόκλισης ***r*** ίση με 1.6435 σε σχέση με το διάνυσμα – συνισταμένη του μικρού κειμένου. Το ***s*** σύμφωνα με την κωδικοποίηση ASCII αντιστοιχεί στο 405 και το ***k*** ισούται με την τιμή 4 (αριθμός χαρακτήρων). Με τον ίδιο τρόπο, εξάγονται οι τιμές για όλα τα 37 συστατικά του μικρού κειμένου, όπως μπορεί να δει κανείς στον Πίνακα 1.

Κατόπιν της διαδικασίας που αναλύθηκε παραπάνω, η τιμή του χ^2 με (38-1=37) βαθμούς ελευθερίας υπολογίζεται μέσω της εξίσωσης (11) και από την τιμή αυτή λαμβάνεται η αθροιστική πιθανότητα $P = 3.5055e - 06$, μέσω της εξίσωσης (12).

Τέλος, χρησιμοποιώντας το έλεγχο μονόπλευρης κατανομής για πιθανότητα $H_0|P < 0.001$, η μηδενική υπόθεση γίνεται δεκτή, υποδεικνύοντας ότι δεν υπάρχει συσχέτιση μεταξύ των τριών μεταβλητών (*r, s, k*).

Πίνακας 1. Algorithm Implementation-Kendall's Coefficient of Concordance-Decision H_0 (an example short text)

Words	r		s		k		Sums of R_j
j	Data	rank	Data	rank	data	rank	
1	1.6435	28	4	12.0	405	8.0	48.0000
2	1.6699	29	9	32.5	997	33.0	94.5000
3	1.3755	23	4	12.0	420	9.0	44.0000
4	1.5487	26	9	32.5	971	32.0	90.5000
5	1.1232	16	4	12.0	433	11.0	39.0000
6	1.3378	22	7	25.5	769	28.0	75.5000
7	1.1818	18	6	21.0	665	22.0	61.0000
8	1.2430	21	8	30.0	846	30.0	81.0000
9	0.8526	11	5	18.0	553	19.0	48.0000
10	4.1013	37	1	1.0	97	1.0	39.0000
11	0.9320	12	7	25.5	739	26.0	63.5000
12	0.2012	4	4	12.0	434	12.5	28.5000
13	1.6753	30	2	3.0	215	3.0	36.0000
14	0.6862	7	7	25.5	730	24.5	57.0000
15	1.1718	17	13	38.0	1402	38.0	93.0000
16	0.6565	6	8	30.0	812	29.0	65.0000
17	0.9587	13	11	37.0	1179	37.0	87.0000
18	0.5982	5	8	30.0	869	31.0	66.0000
19	3.3125	35	2	3.0	213	2.0	40.0000
20	0.6884	8	10	35.0	1045	34.0	77.0000

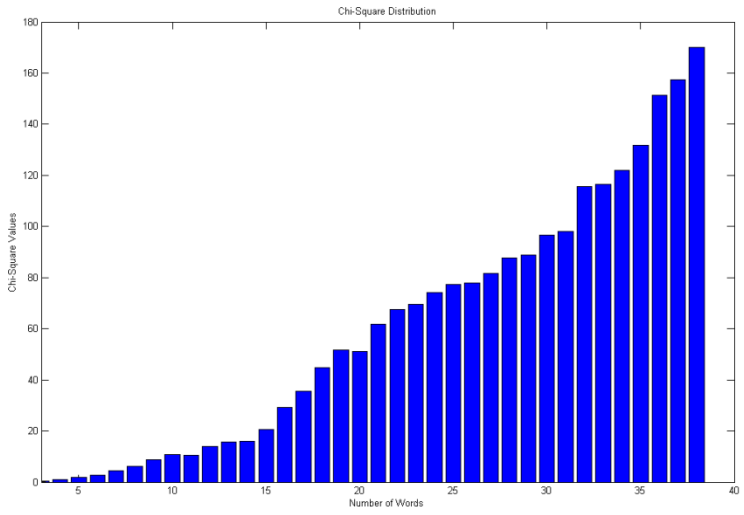
21	0.0927	1	7	25.5	709	23.0	49.5000
22	1.2065	20	4	12.0	421	10.0	42.0000
23	0.7047	9	5	18.0	529	18.0	45.0000
24	1.3804	24	4	12.0	434	12.5	48.5000
25	2.8707	33	3	6.0	307	5.0	44.0000
26	1.5666	27	4	12.0	444	14.5	53.5000
27	1.4494	25	5	18.0	478	17.0	60.0000
28	3.3433	36	3	6.0	312	6.0	48.0000
29	0.8056	10	6	21.0	641	20.0	51.0000
31	3.3023	34	3	6.0	337	7.0	47.0000
32	0.1112	3	10	35.0	1061	35.5	73.5000
33	6.3095	38	2	3.0	225	4.0	45.0000
34	2.4095	31	4	12.0	450	16.0	59.0000
35	1.1958	19	6	21.0	653	21.0	61.0000
36	0.1046	2	10	35.0	1061	35.5	72.5000
37	2.8506	32	4	12.0	444	14.5	58.5000
38	1.1168	15	7	25.5	730	24.5	65.0000
$W = 0.8023\ x_r^2 = 89.06\ P = 3.5055\text{e-}06$ accept for $P < 0.001$							

B. Επαναληπτική διαδικασία

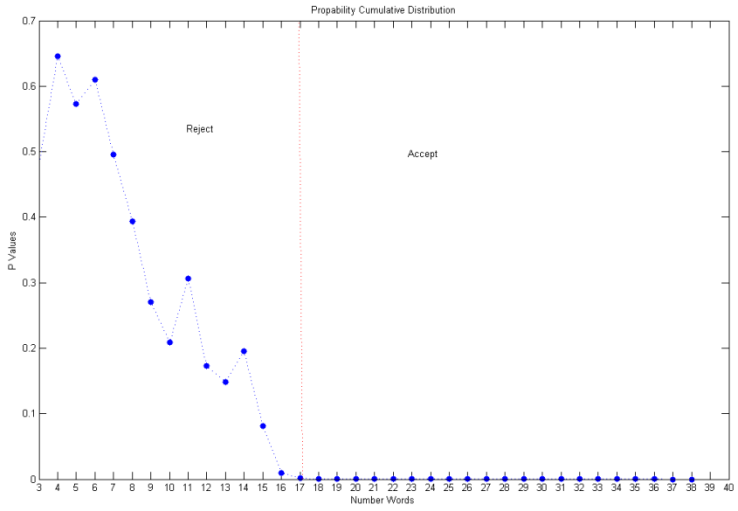
Χρησιμοποιώντας το ίδιο παράδειγμα μικρού κειμένου εφαρμόζεται ο αλγόριθμος $\left[\vec{U}_J\right]_3^{38}$. Η επαναληπτική διαδικασία διεξάγεται με j=3:1:38 λέξεις.

Ως επαναληπτική διαδικασία ορίζεται μια υπολογιστική/μαθηματική διαδικασία συνεχούς επανάληψης ενός κύκλου εργασιών/δραστηριοτήτων, ώστε να προσεγγιστεί ένα επιθυμητό αποτέλεσμα (Everitt 2006, Downing et. al 2009).

Έτσι το αρχικό μικρό κείμενο τμηματοποιείται σε μικρά κείμενα με ποικίλα μήκη ξεκινώντας από μήκος 3 συστατικών έως 38 και αυτή η διαδικασία εκτελείται για κάθε ένα από τα παραπάνω μικρά κείμενα (βλέπε ενότητα του άρθρου III.A). Έπειτα, ο έλεγχος χ^2 επαναλαμβάνεται (35) τριάντα πέντε φορές (βλέπε εικ. 2) και υπολογίζονται οι αθροιστικές πιθανότητες και ο έλεγχος υπόθεσης (βλέπε εικ. 3).



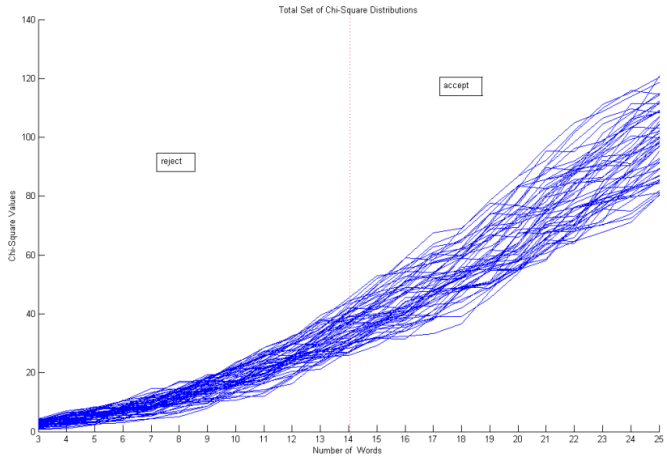
Εικ. 2 Η κατανομή της συνάρτησης χ^2 σύμφωνα με την επαναληπτική διαδικασία του τμηματοποιημένου μικρού κειμένου



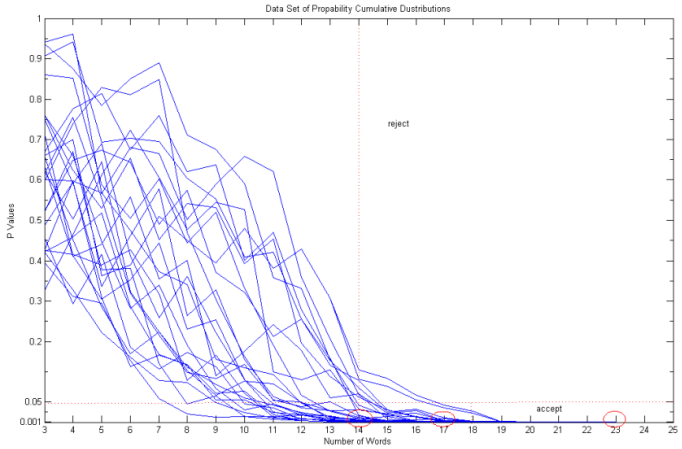
Εικ. 3 Οι αθροιστικές πιθανότητες και ο έλεγχος υπόθεσης

C. Επαναληπτική διαδικασία στο σύνολο δεδομένων

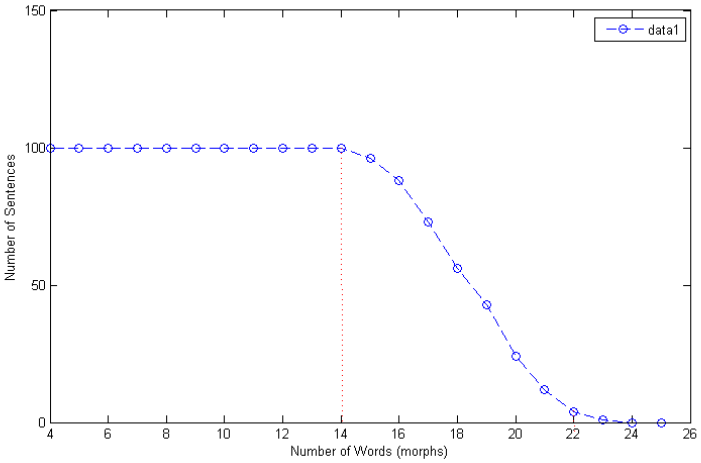
Η επαναληπτική διαδικασία εκτελείται για κάθε μικρό κείμενο σε ένα εύρος από 3 έως 25 συστατικά, συνολικά (22) είκοσι δύο φορές. Η διαδικασία εκτελείται για ένα δείγμα 100 συνόλων δεδομένων, δηλαδή συνολικά 2200 υπολογισμοί. Έπειτα ο έλεγχος χ^2 του συνόλου δεδομένων αναπαρίσταται (βλέπε εικ. 4) και οι αθροιστικές πιθανότητες καθώς και ο έλεγχος της υπόθεσης παρουσιάζονται στις εικ. 5 και 6 αντίστοιχα.



Εικ. 4 Αθροιστικές πιθανότητες και έλεγχος υπόθεσης για το σύνολο των δεδομένων



Εικ. 5 Αθροιστική κατανομή πιθανοτήτων για το σύνολο των δεδομένων



Εικ. 6 Αθροιστικές πιθανότητες και έλεγχος υπόθεσης

Διεξάγοντας αξιολόγηση των παραπάνω στατιστικών αποτελεσμάτων, τα χωρίζουμε σε τρία μέρη και κάνουμε τις παρακάτω παρατηρήσεις:

1. Όπως μπορεί κανείς να δει στην κατανομή στην εικ. 4 «*Αθροιστικές πιθανότητες και έλεγχος υπόθεσης για το σύνολο των δεδομένων*» μέσω της πειραματικής συνάρτησης, η συνεκτικότητα παρατηρείται για όλο συνολικά το δείγμα των 100 μικρών κειμένων σε μήκος πρότασης αποτελούμενο από 14 συστατικά. Από τα 14 και πάνω μπορεί κανείς να παρατηρήσει έλλειψη συνεκτικότητας στην πειραματική συνάρτηση.
2. Σύμφωνα με το πείραμα, όπως απεικονίζεται στην εικ. 5 «*Αθροιστική κατανομή πιθανοτήτων για το σύνολο των δεδομένων*», όλες οι υπολογισμένες πιθανότητες με βάση την εξίσωση (12) απορρίπτουν τη μηδενική υπόθεση για μικρά κείμενα με μήκος ίσο με 14 συστατικά, με πιθανότητα, $p<0.01$ ή p περίπου 99%. Συνεπώς αυτό δείχνει ότι στο επιλεγμένο δείγμα ο συσχετισμός των τριών μεταβλητών ισχύει μέχρι το μήκος των $14^{ων}$ συστατικών. Συνεχίζοντας για μήκος από 15 έως 23 συστατικά οι συσχετίσεις ασθενούν.

Τέλος, για μικρά κείμενα με μήκος 25 συστατικών και πάνω οι συσχετισμοί πλέον δεν υφίστανται/καταρρέουν.

3. Μέσω της εικ. 6 «*Αθροιστικές πιθανότητες και έλεγχος υπόθεσης*», η προαναφερθήσα διαδικασία απεικονίζει συγκεντρωτικά τη μηδενική υπόθεση, η οποία είτε γίνεται δεκτή είτε απορρίπτεται. Όπως φαίνεται στην εικ. 6, όλες οι μηδενικές υποθέσεις απορρίπτονται για μήκος μικρών κειμένων έως 14. Στο διάστημα από 15 έως 22 συστατικά μπορεί κανείς να παρατηρήσει αστάθεια σε σχέση με την απόρριψη της μηδενικής υπόθεσης. Εν τέλει, για μήκος μικρού κειμένου 23 συστατικών και πάνω, η μηδενική υπόθεση ικανοποιείται πλήρως, συνεπώς γίνεται δεκτή.

   Συνοψίζοντας τα παραπάνω αποδεικνύουν ότι:

   - Κάθε μικρό κείμενο εξαρτάται από τρείς προκαθορισμένες μεταβλητές, όταν το μικρό αυτό κείμενο αποτελείται από τουλάχιστον 14 συστατικά.
   - Κάθε μικρό κείμενο, με μήκος περίπου 14 συστατικά, εξαρτάται από τρείς προκαθορισμένες μεταβλητές.

Αυτές οι δύο αρχές ικανοποιούνται πλήρως με πιθανότητα πάνω από 99% και με βαθμό ελευθερίας που κυμαίνεται από 3 – 13.

D. Πολυγλωσσική εφαρμογή

Η προτεινόμενη μέθοδος βασίζεται στην υπόθεση της ποσοτικοποίησης λέξεων (μοναδική ταύτιση) σύμφωνα με τον αριθμό κωδικοποίησης χαρακτήρων ASCII. Έτσι, η ποσοτικοποίηση επιτρέπει την εφαρμογή της μεθόδου να είναι δυνατή για κάθε γλώσσα χωρίς να προκαλεί κάποια διατάραξη στα όρια απόρριψης ή αποδοχής της μηδενικής υπόθεσης.

Προκειμένου να επιβεβαιωθεί η παραπάνω υπόθεση, διεξάχθηκε ένα πείραμα μικρότερης κλίμακας, με μικρά κείμενα τα οποία προέρχονται από τη γερμανική γλώσσα. Για αυτό το μικρότερης κλίμακας πείραμα επαναλήφθηκαν τα ίδια βήματα, όπως έχουν περιγραφεί στο πειραματικό μέρος.

Πιο συγκεκριμένα: Αρχικά συγκεντρώθηκε ένα δείγμα από 10 μικρά κείμενα, με ποικίλα μήκη (λέξεις που αποτελείται), στη γερμανική γλώσσα. Το δείγμα προήλθε κατόπιν αναζήτησης μέσω της μηχανής αναζήτησης Google Scholar και όλα τα μικρά κείμενα ανακτήθηκαν σε μορφή *pdf* και αποθηκεύτηκαν σε τοπικό φάκελο. Δεύτερον, κάθε μικρό κείμενο προ-επεξεργάστηκε όπως έχει αναφερθεί παραπάνω στην ενότητα του άρθρου III.C. Τα αποτελέσματα αυτής της διαδικασίας παρουσιάζονται στον πίνακα 2:

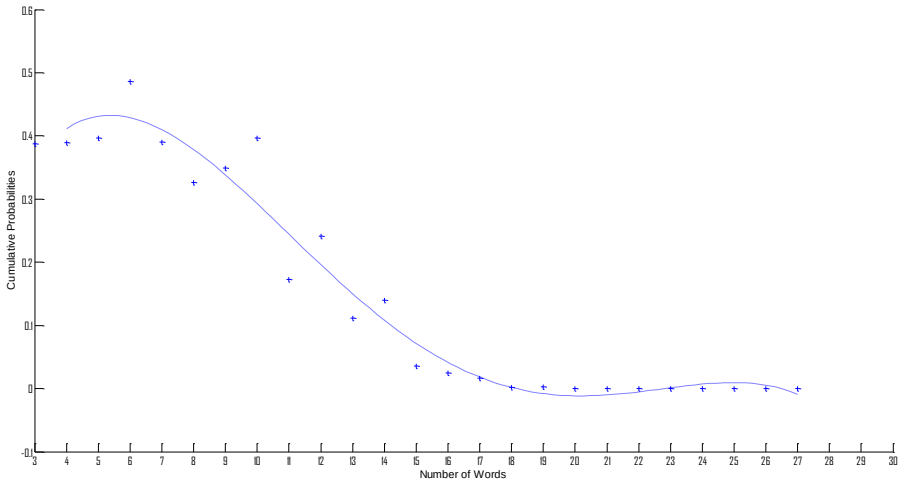
Πίνακας 2. Hypothesis testing of German Sentences

NUMBER_SENTENCES	NUMBER_WORDS	REJECT	ACCEPT
10	3	10	0
10	4	10	0
10	5	10	0
10	6	10	0
10	7	10	0
10	8	10	0
10	9	10	0
10	10	10	0
10	11	10	0

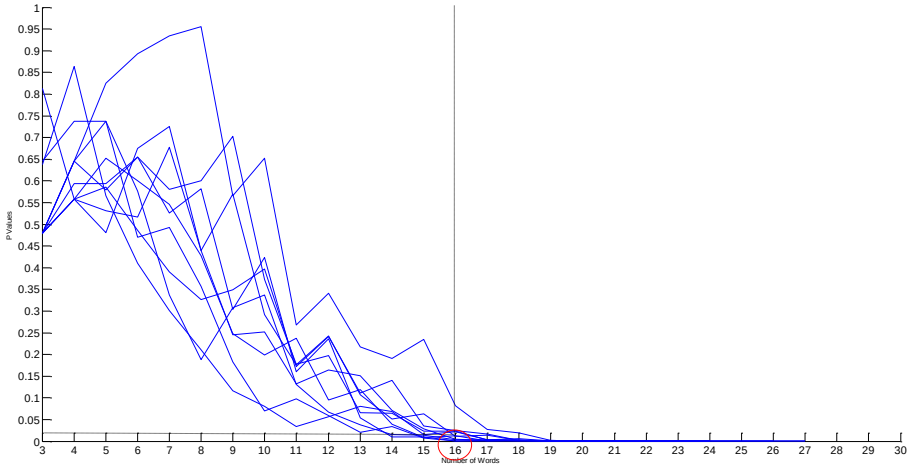
10	12	10	0
10	13	10	0
10	14	10	0
10	15	10	0
10	16	10	0
10	17	9	1
10	18	8	2
10	19	2	8
9	20	0	9
7	21	0	7
7	22	0	7
6	23	0	6
6	24	0	6
3	25	0	3
3	26	0	3
2	27	0	2

Παρουσιάζεται ένα παράδειγμα, το οποίο προέρχεται από τον πίνακα 2 για την παρακάτω πρόταση:

"Die Computerindustrie hätte nach Michael Levitt einen Teil des Nobelpreises für Chemie 2013 vedient denn ihre Forschungs und Entwicklungsleistung hatte zu drastisch höheren Rechengeschwindigkeiten geführt (siehe Tabelle)"



Εικ. 7 Αθροιστικές πιθανότητες και έλεγχος υπόθεσης για την γερμανική περίπτωση



Εικ. 8 Πιθανότητες αθροιστικής κατανομής για σύνολο δεδομένων από δέκα (10) γερμανικά κείμενα

Όπως φαίνεται στον πίνακα 2 και στις εικ. 7 και 8, τα αποτελέσματα του πειράματος που διεξάχθηκε με μικρά κείμενα στη γερμανική γλώσσα βρίσκονται σε συμφωνία με το μεγάλης κλίμακας πείραμα που διεξάχθηκε στην αρχή του πειραματικού μέρους του άρθρου. Αυτό αποδεικνύει ότι τα συστατικά ενός μικρού κειμένου, στο προτεινόμενο μοντέλο μπορούν να θεωρηθούν ως μορφήματα (βλέπε ενότητα της διατριβής 3.2), στα οποία η γλωσσολογική προέλευση του μικρού κειμένου αγνοείται και έτσι κατά συνέπεια η γλώσσα του κειμένου δεν έχει σημασία. Πιο συγκεκριμένα, σύμφωνα με τις εικ. 7 και 8 η αθροιστική πιθανότητα ξεκινά να μειώνεται δραστικά ανάμεσα στην $15^{η}$ και $16^{η}$ λέξη, όπου $p<0.01$.

IV. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΜΕΛΛΟΝΤΙΚΑ ΣΧΕΔΙΑ

Το βασικό ερώτημα αυτού του άρθρου είναι η μέτρηση της λεξιλογικής συνεκτικότητας στα μικρά κείμενα, ώστε να επιτυγχάνεται η αναγνωσιμότητα και η κατανόηση. Παρότι αυτό το ζήτημα έχει μελετηθεί και στο παρελθόν, για πρώτη φορά με αυτήν την προσέγγιση εφαρμόζεται η εξαγωγή τριών μεταβλητών (r, s, k) που προσδιορίζουν μοναδικά κάθε συστατικό μικρού κειμένου. Αυτές οι μεταβλητές θεωρούνται σημαντικοί παράγοντες σε σχέση με τη συνεκτικότητα του μικρού κειμένου. Διεξάχθηκε η διαδικασία υπολογισμού του δείκτη συμφωνίας w του Kendall ώστε να επιβεβαιωθεί ο συσχετισμός μεταξύ των τριών μεταβλητών, σε σχέση με το μήκος μικρών κειμένων. Όλα τα μικρά κείμενα για το παραπάνω πείραμα, είναι στην αγγλική γλώσσα, αλλά προκειμένου να επιβεβαιωθεί η υπόθεση ότι η γλώσσα του κειμένου δεν παίζει σημαντικό ρόλο, διεξάχθηκε και ένα μικρότερης κλίμακας πείραμα, με μικρά κείμενα στη γερμανική γλώσσα, στο οποίο προέκυψε ελάχιστη διαφοροποίηση, η οποία δε διαταράσσει τη μηδενική υπόθεση.

Σημαντικά στατιστικά αποτελέσματα προέκυψαν σε σχέση με τις τρεις μεταβλητές που επιλέχθηκαν, τα οποία έδειξαν ότι ο συσχετισμός είναι σημαντικός

με πιθανότητα που αγγίζει το 95%, για μικρά κείμενα μήκους 17 συστατικών και το 99% για μικρά κείμενα μήκους 14 συστατικών. Συνεπώς, ένα μικρό κείμενο με μήκος που κυμαίνεται από 14-17 συστατικά μπορεί να θεωρηθεί ως ένας δείκτης επιτυχίας του κειμένου σχετικά με την επιτυχημένη μεταφορά του νοήματος του, το οποίο είναι ο λόγος δημιουργίας του (βλέπε ενότητες της διατριβής 3.1.1 και 3.1.2).

Εξάλλου, τα αποτελέσματα τα οποία εξάχθηκαν βρίσκονται σε συμφωνία με τα τρέχοντα επιστημονικά ευρήματα του τομέα της Γλωσσολογίας, όπου ο Kornai (2008) αναφέρει ότι το μέσο μήκος πρότασης αποτελείται από πάνω από 5 συστατικά, ο Cutts (2009) ότι το μέσο μήκος πρότασης κυμαίνεται ανάμεσα σε 15-20 συστατικά, ώστε να διατηρείται η αναγνωσιμότητα και οι Taskar et al. (2004) διεξάγουν τα πειράματα τους θέτοντας περιορισμούς στο μήκος προτάσεων ίσο ή μικρότερο από 15 συστατικά. Το αποτέλεσμα του άρθρου αυτού βρίσκονται ακόμη σε συμφωνία με αποτελέσματα προερχόμενα από τις επιστήμες της Γνωστικής Ψυχολογίας και των Νευροεπιστημών, όπου ο Baddeley (2003) αναφέρει πάλι τον περιορισμό των 15 συστατικών ανά πρόταση. Με βάση τα πειράματα των Davemans και Carpenters (1980), των Anderson et al. (2001), του Zimny (1987), μπορεί κανείς να υπολογίσει ότι η μέση πρόταση θα πρέπει να περιέχει περίπου 16 συστατικά, ώστε να: είναι κατανοητή, να επικοινωνεί το μήνυμα της με επιτυχία και να διατηρεί την αναγνωσιμότητα του κειμένου. Σε αυτήν την έρευνα παρατηρείται η συνάφεια μεταξύ των επιστημονικών τομέων της Γλωσσολογίας, των Νευροεπιστημών και της Γνωστικής Ψυχολογίας.

Μελλοντικά σχέδια:

- Η δημιουργία μιας οντολογίας, η οποία θα βασίζεται στις μεταβλητές και τη στατιστική ανάλυση σε σχέση με το μήκος μικρών κειμένων που προέκυψαν από αυτήν την έρευνα.
- Η χρήση των τριών μεταβλητών (r, s, k), ώστε να διερευνηθεί πειραματικά πως αυτές επηρεάζουν τον τρόπο σύλληψης μικρών κειμένων από τον ανθρώπινο εγκέφαλο.
- Να εκφραστεί η σχέση μεταξύ των τριών μεταβλητών με μια μαθηματική εξίσωση, ώστε να είναι δυνατή η περιγραφή και ο υπολογισμός της λεξιλογικής συνεκτικότητας των μικρών κειμένων με βάση το μήκος τους.
- Ένα σημαντικό χάσμα μεταξύ των Νευροεπιστημών και της Γλωσσολογίας, αναφέρεται ως "Ontological Incommensurability Problem" (OIP) (βλέπε ενότητα της διατριβής 1.2) όπου "*τα βασικά στοιχεία της γλωσσολογικής θεωρίας δεν μπορούν να αντιστοιχηθούν με τις βασικές βιολογικές μονάδες που αναγνωρίζονται από τις Νευροεπιστήμες*" όπως αναφέρουν οι Poeppel και Embick (2005). Αυτή η έρευνα μπορεί να αποτελέσει το θεμέλιο λίθο ώστε να μειωθεί το χάσμα μεταξύ των δύο επιστημονικών πεδίων, Νευροεπιστημών και Γλωσσολογίας, εστιάζοντας στο ζήτημα του μήκους των μικρών κειμένων, καθώς τα αποτελέσματα που προέρχονται και από τις δύο επιστήμες βρίσκονται σε συμφωνία.

- Τα αποτελέσματα αυτής της έρευνας σε συνδυασμό με τα συμπεράσματα από τα επιστημονικά πεδία Νευροεπιστημών και Γλωσσολογίας, σε σχέση με το μήκος μικρών κειμένων μπορούν να αποτελέσουν μια νέα αρχή για τη θεωρία ελέγχου και τη σταθερότητα συστημάτων.
- Περαιτέρω έρευνα που θα αφορά τις εφαρμογές της υπόθεσης του άρθρου σε κείμενα διαφόρων γλωσσών. Πιο συγκεκριμένα, συγκριτική μελέτη ανάμεσα στις γλώσσες, των οποίων οι λέξεις αποτελούνται από μικρότερου μήκους μορφήματα και μεγαλύτερου μήκους μορφήματα, ώστε να οριστεί ένα σταθερό εύρος.
- Τέλος, λαμβάνοντας υπόψη ότι ο νόμος ΜΑ εφαρμόζεται σε γονιδιώματα όπως αναφέρουν και οι Baixeries et al. (2013), Baixeries, Hernández-Fernández, Ferrer-i-Cancho (2012) και οι Forns et al. (2013) θα μπορούσε να διερευνηθεί επιστημονικά η πιθανή σχέση μεταξύ του αριθμού των μεταβλητών και του αριθμού των συστατικών σε άλλα επιστημονικά πεδία (όπως βιολογία), όπου θα μπορούσαν να θεωρηθούν ως ένας άξονας για τη συνεκτικότητα του συστήματος και ως ανερχόμενη επιστημονική προτεραιότητα.

Βιβλιογραφία

Altmann, G. (1980). Prolegomena to Menzerath's Law. Glottometrika 2, 1–10. Bochum: Brock-meyer,

Anderson J. R., Budiu R. and Reder L. M. (2001). A theory of sentence memory as part of a General Theory of Memory. Journal of Memory and Language. 45. 337-336.

Baddeley A. (2003). Working memory: looking back and looking forward. Nature reviews - neuroscience. 5. 829-839.

Baixeries J. et al. (2013). The parameters of the Menzerath-Altmann Law in genomes. Journal of Quantitative Linguistics. 20 (2). 94–104.

Baixeries J., Hernández-Fernández A., Ferrer-i-Cancho R. (2012), Random models of Menzerath–Altmann law in genomes. Biosystems. 107 (3), 167–173.

Carstens W. (2001). Text Linguistics: relevant linguistics? Poetics and linguistics; discources of war and conflict Conference. 588-595.

Cutts, M. (2009). Oxford guide to plain English. 3rded. Oxford: Oxford University Press.

Daveman M., Carpenter P. A. (1980). Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior. 19. 450-466.

Downing, D. et al. (2009). *Dictionary of Computer and Internet Terms*. 10th ed. Hauppauge, NY: Barron's Educational Series

Eroglu S. (2013).Menzerath–Altmann law for distinct word distribution analysis in a large text. Physica A. 392 (12). 2775–2780

Everitt, B. S. (2006). *The Cambridge Dictionary of Statistics*. 3rd ed. Cambridge: Cambridge University Press.

Fahnestock J. (1983). Semantic and Lexical Coherence. College Composition and Communication. 34 (4). 400-416. National Council of Teachers of English.

Forns N. et al. (2013). The challenges of statistical patterns of language: The case of Menzerath's law in genomes. Complexity. 18 (3). 11–17.

Ge Song et al (2014). Short Text Classification: A Survey. Journal of Multimedia. 9 (5). 635-643. Academy Publisher.

Halliday, M.A.K., Hasan R. (1976). Cohesion in English. London: Longman.

Harispe S. et al. (2013). Semantic measures for the comparison of units of language, concepts or entities from text and knowledge base analysis. Arxiv. 1310. 1285. 1-159.

Harispe, S. et al. (2013). The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. Bioinformatics. 30 (5). Oxford: Oxford University Press.

Kornai A. (2008). Mathematical linguistics. [Online]. Advanced Information and Knowledge Processing. London: Springer.

Luděk Hřebíček (2002). Zipf's Law and Text. Glottometrics. 3. 27-38. Ram Verlag.

Poeppel, D., Embick, D. (2005). Defining the relation between linguistics and neuroscience. Twenty-first Century Psycholinguistics: Four Cornerstones. 103–118.

Poulos M., Papavlasopoulos S., Chrissikopoulos V. (2006). A text categorization technique based on a numerical conversion of a symbolic expression and an onion layers algorithm. Journal of Digital Information. 6 (1).

Raghavan V. V., Wong S. K. M. (1986). A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science. 37 (5). 279-287.

Richards, J. C. and Schmidt, R. (2002). Longman dictionary of language teaching and applied linguistics. 3rd ed. London: Longman

Salton G., Wong A., and Yang C. S. (1975). A vector space model for automatic indexing. Communications of the ACM. 18 (11). 613-620.

Sparks J.R. (2012). Language/discourse comprehension and understanding. Encyclopedia of the Learning Sciences. 1713-1717. Springer.

Taskar B. et al. (2004). Max-Margin Parsing. Proceedings of EMNLP 2004. 1-8.

Tesitelová M. (1992). Quantitative linguistics. Linguistics and literary studies in Eastern Europe (LLSEE). 37. Philadelphia: John Benjamins

Turney P. D., Pantel P. (2010). From frequency to meaning: vector space models of semantics. Journal of Artificial Intelligence Research. 37 (1). 141-188.

Xiaojun Quan, Gang Liu et al. (2009). Short text similarity based on probabilistic topics. Knowl Inf Syst. 473-491.

Zar Jerrold H. (1999). Biostatistical Analysis. 4th ed. Upper Saddle River, N.J.: Prentice Hall.

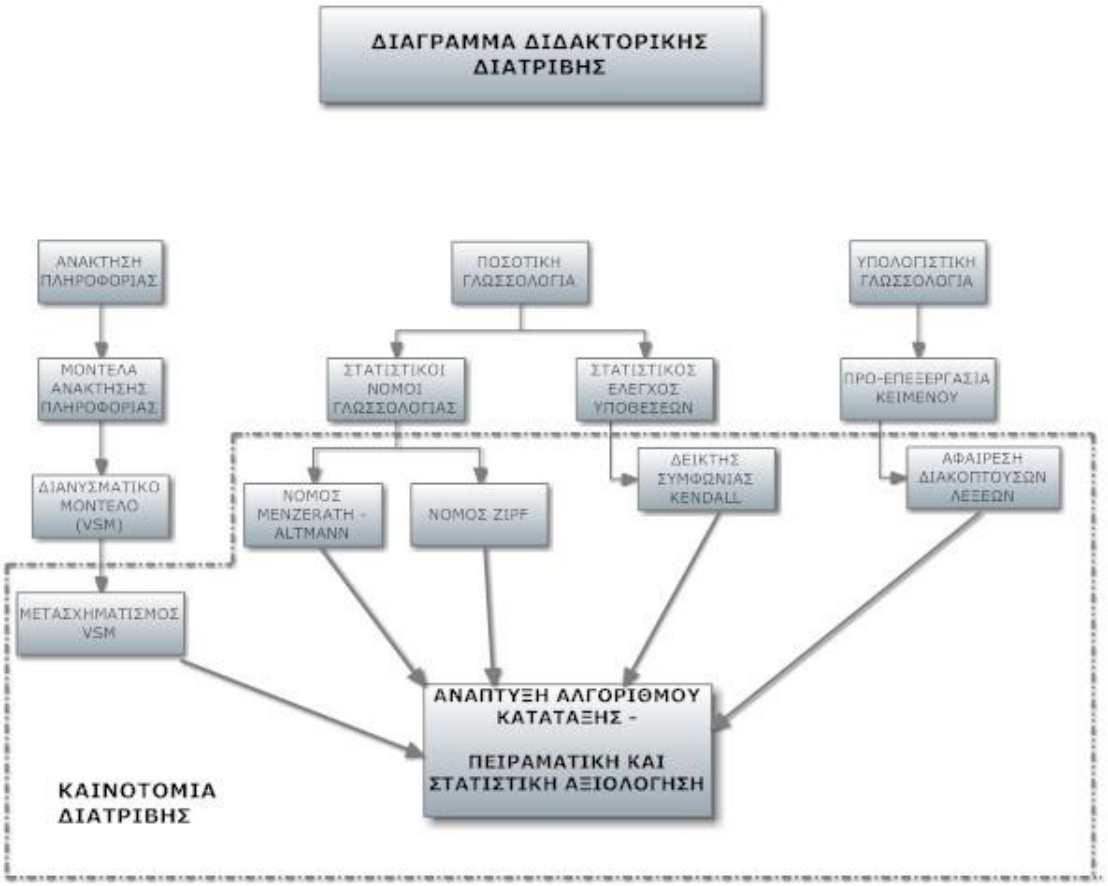
Zimny, S. T. (1987). Recognition memory for sentences from a discourse. Unpublished doctoral dissertation, Boulder: University of Colorado.

ΚΕΦΑΛΑΙΟ 6ο

ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1 Συμπεράσματα

Στο παρακάτω διάγραμμα (εικ. 20) αναπαρίσταται με γραφικό τρόπο η καινοτομία της παρούσας διατριβής και η συμβολή της στους επιστημονικούς κλάδους της ΑΠ, ΥΓ και ΠΓ.



Εικ. 20 Τομή επιστημονικών κλάδων και καινοτομία διατριβής

Σύμφωνα με την εικ. 20, η διδακτορική διατριβή κινήθηκε σε τρεις (3) σημαντικούς επιστημονικούς κλάδους, την ΑΠ, την ΥΓ και την ΠΓ.

Στον 1ο κλάδο, της ΑΠ, η διατριβή επικεντρώθηκε στο διανυσματικό μοντέλο VSM. Όπως προαναφέρθηκε αυτό το μοντέλο ανάκτησης βασίζεται σε μια διανυσματική αναπαράσταση των εγγράφων, σε έναν πολυδιάστατο χώρο όρων. Το πρόβλημα που παρουσιάζεται σε αυτήν τη μοντελοποίηση είναι η πολυδιάστατη προσέγγιση του, λόγω της εμφάνισης μεγάλου αριθμού όρων σε κάθε κείμενο. Στην παρούσα διατριβή επιχειρήθηκε, μέσω της φιλοσοφίας του μοντέλου VSM, η

διανυσματική απεικόνιση των όρων ενός κειμένου, η οποία πραγματοποιείται σε έναν *a priori* τρισδιάστατο χώρο. Ο τρισδιάστατος αυτός χώρος βασίστηκε ουσιαστικά σε τρεις μεταβλητές των συστατικών - λέξεων ενός κειμένου. Οι μεταβλητές αυτές επελέγησαν έχοντας ως κοινό γεωμετρικό τόπο την εκάστοτε μοναδικότητα τους σε ένα συγκεκριμένο κείμενο. Αυτή η μοναδικότητα ορίστηκε από τη θέση που βρίσκεται η κάθε λέξη στο κείμενο (*i*), από τον αριθμό χαρακτήρων της κάθε λέξης (*k*) και από την ένταση της λέξης που αντιστοιχεί σε έναν αριθμό κωδικοποίησης (*s*). Στη συνέχεια, επιχειρήθηκε να αποδοθεί μία ξεχωριστή οντότητα της κάθε λέξης σε σχέση με την πρόταση που εντάσσεται. Αυτό πραγματοποιήθηκε με την αντικατάσταση της (*i*) με την (*r*) μεταβλητή (Poulimenou et al. 2014, Poulimenou et al. *to appear in* 2016) που υποδηλώνει τη διανυσματική σχέση που έχει η κάθε λέξη με τη συνολική πρόταση.

Στον 2ο κλάδο, της ΥΓ, επιχειρήθηκε να γίνει μια προ-επεξεργασία των κειμένων σύμφωνα με γραμματικούς και συντακτικούς κανόνες, οι οποίοι απέφεραν ένα κείμενο προς τελική επεξεργασία που δε διέθετε πλέον λέξεις που θεωρούνται διακόπτουσες (Poulimenou et al. 2014). Η υλοποίηση αυτού πραγματοποιήθηκε μέσω της χρήσης προγράμματος επισημειωτή κειμένου.

Στον 3ο κλάδο, της ΠΓ, επιχειρήθηκε να γίνει μια στατιστική συσχέτιση των τριών μεταβλητών, οι οποίες εξήχθησαν από την πρώτη επιστημονική ενότητα, αυτήν της ΑΠ, ώστε να διαπιστωθεί εάν και κατά πόσον αυτές συσχετίζονται μεταξύ τους, σε σχέση με τον αριθμό των λέξεων του εξεταζόμενου κειμένου. Για το σκοπό αυτό χρησιμοποιήθηκε ο έλεγχος συσχέτισης μέσω του δείκτη συμφωνίας w του Kendall. Το αποτέλεσμα αυτού του ελέγχου έδειξε ότι αυτές οι τρεις μεταβλητές συσχετίζονται για όλο το πλήθος των κειμένων της συλλογής εφόσον το κάθε ένα από αυτά αποτελείται από ένα πλήθος (μήκος κειμένου) ως 14 λέξεις (Poulimenou et al. *to appear in* 2016). Από 15-22 λέξεις η συσχέτιση αρχίζει προοδευτικά να ατονεί. Από 23 λέξεις και πάνω διαπιστώθηκε ότι οι τρεις μεταβλητές δεν έχουν καμία συσχέτιση και κατά συνέπεια οι συσχετισμοί καταρρέουν.

Το συνολικό συμπέρασμα για την παρούσα διατριβή δείχνει μια σημαντική καινοτομία στον κλάδο της ΠΓ. Όπως προαναφέρθηκε σε παραπάνω ενότητα (βλέπε ενότητα 3.3.1) η ΠΓ μπορεί να ερμηνεύσει ζητήματα κατανομής λέξεων σε κείμενα και σύνδεσης ιδιοτήτων τους (βλέπε ενότητες 3.3.1.1, 3.3.1.2), όταν τα κείμενα αναφέρονται σε ένα μεγάλο αριθμό όρων – λέξεων. Μέχρι τούδε, η ερμηνεία κατανομής και δόμησης λέξεων σε μικρής έκτασης κείμενα ήταν αδιερεύνητη.

Η παρούσα διατριβή δημιουργεί ένα πρώτο σοβαρό στάδιο ερμηνείας δόμησης μικρών κειμένων και ιδιαιτέρως προτάσεων. Ο σημαίνον αριθμός 14 επιβεβαιώνει την εμπειρική θεώρηση των γλωσσολόγων που έχουν σχέση με τη συνεκτικότητα της πρότασης (Poulimenou et al. *to appear in* 2016) και ορίζει ένα ιδανικό μήκος μικρού κειμένου, προκειμένου αυτό να μπορεί να μεταδώσει το νόημα του με επιτυχία χωρίς να διαταράσσει τη συνεκτικότητα του. Η επίτευξη του σκοπού αυτού έγινε μέσω μιας άλλης καινοτόμου εξέλιξης του μοντέλου VSM.

Ειδικότερα όσον αφορά τους περιορισμούς του μοντέλου ΑΠ VSM, αποδεικνύεται με μαθηματικό τρόπο (Poulimenou et al. to appear in 2016) ο περιορισμός εφαρμογής του σε κείμενα μεγάλου μήκους, καθώς από 15 συστατικά και πάνω ξεκινά μια εξασθένηση συσχετίσεων, ενώ από τα 23 συστατικά και πάνω οι συσχετίσεις αυτές πλέον παύουν να υφίστανται.

Ακόμη, μέσω της σύγκρισης του αλγόριθμου που αναπτύχθηκε στην διατριβή με το σχήμα απόδοσης βαρών *tf-idf*, αποδεικνύεται η δυνατότητα του αλγόριθμου να αυτό-προσδιορίζεται μέσα σε κάθε μικρό κείμενο και να εφαρμόζεται αποτελεσματικά, χωρίς να εξαρτάται από το σύνολο των αποτελεσμάτων του δείγματος ενός πληθυσμού για την απόδοση βάρους.

6.2 Ανοικτά ερευνητικά ζητήματα

Τα προηγούμενα συμπεράσματα δημιουργούν νέα μεγάλα ερευνητικά ζητήματα. Η διαθεματική συσχέτιση των αποτελεσμάτων που προαναφέρθηκαν μπορούν να συσχετιστούν και να διερευνηθούν περαιτέρω με τους επιστημονικούς κλάδους των Νευροεπιστημών, της Γνωστικής Ψυχολογίας και της Γενετικής. Επίσης ανοίγονται νέοι ορίζοντες σε σχέση με το ζήτημα της ευστάθειας ή της συνεκτικότητας συνόλων που εκφράζονται με ποσοτικό τρόπο. Τέλος, ανοίγεται ένας νέος διεπιστημονικός ορίζοντας των τριών παραπάνω επιστημονικών κλάδων ΑΠ, ΥΓ και ΠΓ.

Ειδικότερα, η επιστημονική συμβολή της παρούσας διατριβής και τα συμπεράσματα της μπορούν να εκφραστούν με μια μαθηματική σχέση, η οποία θα μπορούσε να αποτελέσει έναν επιστημονικό νόμο στο πλαίσιο της ΠΓ (βλέπε ενότητα 3.3.1), που θα αφορά στο μέγεθος μικρών κειμένων, ο οποίος μέσω της παρούσας εργασίας θεμελιώνεται τόσο λόγω σύνδεσής του με τα εμπειρικά επιστημονικά αποτελέσματα του κλάδου όσο και με πειραματικό τρόπο σε πραγματικά δεδομένα.

Ακόμη, τα συμπεράσματα της παρούσας διδακτορικής διατριβής μπορούν να αποτελέσουν ένα **βασικό κριτήριο ποιότητας και αξιολόγησης κειμένου σε σχέση με τη συνεκτικότητα του** (βλέπε ενότητα 3.1.3) και την κατανόηση του από τον αναγνώστη, έναν σταθμό στον οποίο στα 14 – 17 συστατικά ένα δόμημα φέρει ένα πλήρες και ολοκληρωμένο μήνυμα προς μετάδοση. Η δε έννοια της συνεκτικότητας, μπορεί να τεθεί σε ευρύτερο πλαίσιο, λόγω της διεπιστημονικής εφαρμογής των ποσοτικών νόμων στη Γλωσσολογία και να αποτελέσει γενικότερα έναν παράγοντα σταθερότητας ανάλογα το σύστημα υπό εξέταση (π.χ. σχέση μεταξύ χρωμοσωμάτων - γονιδιωμάτων). Ακόμη τα συμπεράσματα της παρούσας εργασίας θα μπορούσαν να **ενσωματωθούν ως χαρακτηριστικά προς εξέταση σχετικά με την κατανόηση κειμένων, στο πλαίσιο των readability formulae** και στο πλαίσιο της μέτρησης αναγνωσιμότητας και ποιότητας κειμένου. Το θέμα αυτό προς διερεύνηση συνδέεται με τον απώτερο στόχο της διατριβής.

Στα ανοικτά επιστημονικά ζητήματα της παρούσας διατριβής ανήκει και η πιθανή συμβολή των συμπερασμάτων για την αντιμετώπιση του χάσματος μεταξύ των επιστημών Γλωσσολογίας και Νευροεπιστημών, υπό το πρίσμα του φαινομένου ασυμμετρίας που παρουσιάζεται μεταξύ τους. Ειδικότερα, το πρόβλημα οντολογικής ασυμμετρίας (Poeppel και Embick 2005) παρατηρείται ως ένα γενικό πρόβλημα αδυναμίας ενοποίησης επιστημών είτε μέσω της μείωσης είτε μέσω της αντιστοίχησης εννοιολογικών γενικεύσεων που περιλαμβάνουν οι επιστήμες αυτές, καθώς αποτελούν γενικεύσεις διαφορετικού είδους. Πρόκειται για ένα φαινόμενο που παρατηρείται γενικότερα στις Γνωσιακές Νευροεπιστήμες, αλλά και ειδικότερα μεταξύ των πεδίων της Γλωσσολογίας και της Νευροεπιστήμης, όπως παρατηρείται με πιο μεγάλη λεπτομέρεια στην εικ. 21. Ειδικότερα, οι μονάδες των διάφορων επιπέδων εννοιών των επιστημών αυτών, με την έννοια της αρχιτεκτονικής δομής τους, δεν δύναται να μετρηθούν και έτσι παρατηρείται το φαινόμενο της ανεξάρτητης ανάπτυξης οντολογιών τους, καθώς δεν υπάρχουν σημαντικοί δεσμοί για την απόπειρα ένωσης τους.

Linguistics	Neuroscience
Fundamental elements of representation (at a given analytic level)	
distinctive feature	dendrites, spines
syllable	neuron
morpheme	cell-assembly/ensemble
noun phrase	population
clause	cortical column
Fundamental operations on primitives (at a given analytic level)	
concatenation	long-term potentiation (LTP)
linearization	receptive field
phrase-structure generation	oscillation
semantic composition	synchronization

Εικ. 21 Εννοιολογική αναντιστοιχία κατηγοριών Γλωσσολογίας και Νευροεπιστημών από Poeppel και Embick (2005)

Στα ανοικτά επιστημονικά ζητήματα σχετικά με τα μικρά κείμενα εντάσσεται και η **περαιτέρω έρευνα όσον αφορά τις διαφορετικές γλωσσικές προελεύσεις** και κατά πόσον αυτές επηρεάζουν τον αριθμό συστατικών που απαρτίζουν ένα δόμημα σε σχέση με τη συνεκτικότητα του. Η διεξαγωγή συμπερασμάτων σχετικά με το ζήτημα αυτό μπορεί να εστιάσει σε ευρύτερες ομάδες γλωσσικών προελεύσεων προερχόμενες από διαφορετικούς τύπους αλφαβήτων. Με την πραγματοποίηση συγκριτικής μελέτης της τυπολογίας των μορφημάτων κάθε γλώσσας και ελέγχοντας την επιρροή τους σε σχέση με ζητήματα αναγνωσιμότητας και κατανόησης κειμένου μπορεί να γίνει περαιτέρω ανάλυση ως προς δεδομένο μήκος κειμένου για τα μικρά κείμενα.

ΟΡΟΛΟΓΙΑ

Α

Αδόμητα δεδομένα = unstructured data

Αθροιστικές συχνότητες = cumulative frequencies

Αθροιστική κατανομή = cumulative distribution function

Αιτιοκρατικά = deterministic

Ακολουθία = stream

Ακρίβεια = precision

Αλληλεπίδραση χρήστη = user interaction

Αμφίπλευρους / μη-κατευθυνόμενου = two-tailed test

Αμφισημία = ambiguity

Ανάγκη πληροφόρησης = information need

Αναγνωριστικό εγγράφου = document identifier/docID

Αναγνωσιμότητα = readability

Ανάδραση = feedback

Ανάκληση = recall

Ανάκτηση Πληροφοριών (ΑΠ) = Information Retrieval (IR)

Αναλογική = ratio

Ανάλυση αρχείων καταγραφής πλοήγησης = clickthrough log analysis

Αναλυτής = parser

Αναλυτής γλώσσας ερωτημάτων = query language parser

Αναπαράσταση κειμένου = text representation

Αναφορές = anaphors

Ανεπιθύμητη αλληλογραφία = spam filter

Ανοικτές κλάσεις = open class

Αντεστραμμένο ευρετήριο/αρχείο = inverted index

Αντωνυμίες = pronouns

Αξιολόγηση = evaluation

Απλή = simple

Αποβλεπτικότητα = intentionality

Αποδοχή = acceptability

Απόδοση = efficiency / performance

Αποσαφήνιση = disambiguation

Αποτελεσματικότητα = effectiveness

Αριθμοί = numerals

Αρχεία καταγραφής = log files ή logs

Αρχειοθήκες καταγραφής ερωτημάτων = query logs

Ασάφεια = ambiguity

Ασαφή σύνολα = fuzzy–set

Αυτόματη επεξεργασία φυσικής γλώσσας = automatic processing of NL

Β

Βαθμοί ελευθερίας = degrees of freedom

Βαθμολογία = score

Βασισμένα σε γνώση = knowledge-based

Βελτιστοποίηση ερωτήματος = query optimization

Βοηθητικά ρήματα = auxiliaries

Γ

Γεωμετρικές = geometric

Γλώσσα = language

Γνωσιακό = cognitive

Γραμμική σάρωση = linear scanning

Γραμματικό έλεγχο = spell checking

Δ

Δειγματικός χώρος = event space

Δηλώσεις = statements

Δημιουργία ευρετηρίου = index creation

Διάγραμμα διασποράς = scatter diagram

Διαδικασία επισημείωσης μερών του λόγου = part of speech tagging/tagging

Διαδοχής στοιχείων/ακολουθίας = string

Διαίρεση σε σύμβολα = tokenization

Διακειμενικότητα = intertextuality

Διακόπτουσες λέξεις = stop words

Διακύμανση = variance

Διαστημική = interval

Διαχωρισμό λέξεων = hyphenation

Δομημένα δεδομένα = structured data

Δομικές ενδείξεις = structural cues

Δύναμη = power

Δυαδική = binary

Ε

Έγγραφο = document

Εγγύτητα = proximity

Είδος κειμένου = genre-based cues

Εισαγωγή δεδομένων = input

Εισαγωγή δεδομένων ερωτήματος = query input

Έλεγχος = test

Έλεγχοι ανεξαρτησίας/διαφορών = independence/differences

Έλεγχος καλής προσαρμογή = goodness of fit test

Έλεγχος στατιστικών υποθέσεων = statistical hypothesis testing

Εμφάνισης λέξης = word occurrence

Εναλλακτική υπόθεση = alternative hypothesis

Ενδείξεις = cues

Ενδιάμεσο = intermediate

Ενθήματα = infixes

Έννοια = concept

Εννοιολογικές αναπαραστάσεις = meaning representations

Εντεθειμένη σε ομοειδή δομή = nested

Εξαγωγή δεδομένων = output

Εξαγωγή αποτελεσμάτων = results output

Επαγωγική στατιστική / στατιστική συμπερασματολογία = inferential statistics

Επίθετα = adjectives

Επιθήματα = suffixes

Επικάλυψη = overlap

Επίπεδο σημαντικότητας = level of significance

Επιρρήματα = adverbs

Επισημειωτές = taggers

Επιφανειακό = surface

Ερώτημα = query

Ετικέτες = tag

Εύρος = range

Ευφυής επεξεργασία φυσικής γλώσσας = intelligent NLP

Εφαρμοσμένη γλωσσολογία = applied linguistics

Ζ

Η

Ήδη υπάρχουσα γνώση = prior knowledge

Θ

Θέματα = stems

Θεώρημα κεντρικού ορίου = central limit theorem

Ι

Κ

Κακόβουλα = spam

Κανόνες αντιστοίχησης = mapping rules

Κανόνες παραγωγής = production rules

Κανονική απόκλιση = normal deviate

Κανονική / γκαουσιανή κατανομή = normal/ gaussian distribution

Κανονική μορφή = canonical form

Κανονικοποίηση = normalize

Κανονιστικές αρχές = regulative principles

Κριτήριο χ^2 / κριτήριο ελέγχου ανεξαρτησίας / κριτήριο ελέγχου πινάκων συνάφειας = chi square test of independence / contingency tables

Κατανεμημένα = distributional

Κατανομή πιθανότητας = probability distribution

Κατανομή αθροιστικών συχνοτήτων = cumulative frequency distribution

Κατανόηση = comprehension

Κατανόηση λόγου = discourse comprehension

Καταταγμένη ανάκτηση = ranked retrieval

Κατάταξη = ranking

Καταχώρηση = posting

Καταστασιακότητα = situationality

Κείμενο = text

Κομμάτια = chunks

Κλειστές κλάσεις = closed class

Κυρίαρχες = dominant

Λ

Λεξήματα = lexemes

Λεξικό = dictionary / lexicon

Λεξιλογικές ενδείξεις = lexical cues

Λεξιλογικές ετικέτες = lexical tags

Λημματοποίηση = lemmatization

Λίστα καταχωρήσεων/ αντεστραμμένη λίστα = posting list

Λόγος = discourse

Λογικές δηλώσεις = Boolean statements

Λογικές προτάσεις = propositions

Μ

Μέρη του λόγου = part of speech/POS

Μετατροπή = conversion

Μετατροπή εγγράφων = text transformation

Μετατροπή ερωτήματος = query transformation

Μέτρα ασυμμετρίας = measures of skewness

Μέτρα διασποράς ή διασκόρπισης = measures of variability / dispersion

Μέτρα θέσης = measures of location

Μέτρα κεντρικής τάσης = measures of central tendency

Μέτρα κύρτωσης = kyrtosis measures

Μέτρα ομοιότητας = similarity measures

Μηδενική υπόθεση = null hypothesis

Μικρού μήκους έγγραφα = short documents

Μονόπλευρη κατανομή = one-tailed distribution

Μονόπλευρος / κατευθυνόμενους ελέγχους = one-tail test

Μοντέλο ανάκτησης = retrieval model

Μόρια = particles

Μορφή λέξης = word form

Μορφήματα = morphs

Μορφολογία = morphology

Μορφολογικός συντακτικός αναλυτής = morphological parser

Μορφοτακτικοί περιορισμοί = morphotactics

Μορφότυπο = format

Ν

Ξ

Ο

Ορθογραφικοί κανόνες = orthographic rules

Οριοθέτης = dilimiter

Ομωνυμία = homonymy

Ονομαστική = nominal

Ονοματική φράση = noun phrase

Όροι = terms

Όροι ευρετηρίου = index terms

Ουρές = tails

Ουσιαστικά = nouns

Π

Παλινδρόμησης = regression

Παράγοντας κανονικοποίησης = normalization factor

Παράγοντας συχνότητας εμφάνισης όρου = term frequency factor

Παράγοντας συχνότητας εμφάνισης όρου στην συλλογή = collection frequency factor

Περιγραφική στατιστική = descriptive statistics

Περιεχόμενο κειμένου = text content

Περιήγηση = browsing

Περιθήματα = circumfixes

Περικείμενο = context

Περιστασιακή ανάκτηση = ad hoc retrieval

Περιστολή = stemming

Πιθανοτικό = probabilistic

Πίνακες αναζήτησης = look-up tables

Πληροφορητικότητα = informativity

Πολλαπλή = multiple

Πολυωνυμική παλινδρόμηση = polynomial regression

Προθέσεις = prepositions

Προθήματα = prefixes

Προσδιοριστής = determinant

Προσδιορισμός συντελεστών στάθμισης/ στάθμιση όρων = weighting

Προσφύματα = affixes

Πρόταση = sentence

Πυκνότητα πιθανότητας = probability density

Ρ

Ρήματα = verbs

Ρηματική φράση = verb phrase

Ροές δεδομένων = feeds

Ρυθμός διαμεταγωγής = throughput

Σ

Σειρα κατάταξης αθροισμάτων = rank sums

Σημαντική διαφορά = significant difference

Σημασία = meaning

Σημασιολογία = semantics

Σημασιολογική ανάλυση = semantic analysis

Σημασιολογική πηγή = semantic proxy

Στοχαστικά = stochastic / probabilistic

Συλλογή εγγράφων = text acquisition

Συλλογή / σώμα εγγράφων = corpus

Σύμβολα = tokens

Συνάφεια = relevance

Σύνδεσμοι = conjunctions

Συνδετικά = connectives

Συνδετικότητα = connectivity

Συνεκτικότητα = coherence

Σύνολο ετικετών = tagset

Συνοχή = cohesion

Συντελεστές στάθμισης = weight

Συντακτική ανάλυση = parsing

Συσταδοποίηση = clustering

Συστατικό = constituent

Συστατικότητας = constituency

Συσχέτιση = correlation

Συχνότητες = frequencies

Σχεδιασμός πειραμάτων = experimental design

Σχέσεις = relations

Τ

Τακτική = ordinal

Ταξινόμηση = classification

Ταξινομητής = classifier

Τελεστής εγγύτητας = proximity operator

Τελεστές ερωτημάτων = query operator

Τυπική απόκλιση = standard deviation

Τύπος = type

Τυποποιημένη τιμή = standard score

Τυποποίηση = standardizing

Υ

Υπόθεση = hypothesis

Φ

Χ

Ψ

Ψήγματα αποτελεσμάτων = result snippets

Ω

ΞΕΝΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Ahlsen, E. (2006). *Introduction to Neurolinguistics*. Amsterdam/Philadelphia: John Benjamins Publishing

Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatr*. 96 (3). P. 338-41.

Altmann, G. (2002). Zipfian linguistics. *Glottometrics*. 3. P. 19-26.

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*. 2. P. 1–10.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval*. 2nd Ed. New York: Addison Wesley

Baixeries, J. et al. (2013). The parameters of the Menzerath-Altmann Law in genomes. *Journal of Quantitative Linguistics*. 20 (2). P. 94–104.

Bendersky, M. and Croft, B. W. (2012). Modeling Higher-Order Term Dependencies in Information Retrieval using Query Hypergraphs. In *SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. P. 941-950.

Blair, D. C. and Kimbrough, S. O. (2002). Exemplary Documents: A Foundation for Information Retrieval Design. *Information Processing and Management*. 38 (3). P. 363-379

Bolshakov, I. A. and Gelbukh, A. (2004). *Computational linguistics: Models, Resources, Applications*. Mexico: Universidad Nacional Autonoma.

Brants, T. (2004). Natural Language Processing in Information Retrieval. In *Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands*. P. 1-13

Buettcher, S., Clarke, C. L. A. and Cormack G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. Cambridge, Mass.: MIT Press

Buk, S., Rovenchak, A. (2008). Menzerath – Altmann law for syntactic structures in Ukrainian. *Glottotheory*. 1(1). P. 10-17

Canfora, G. and Cerulo, L. (2004). A Taxonomy of Information Retrieval Models and Tools. *Journal of Computing and Information Technology (CIT)*. 12 (3). p. 175–194

Carstens, W. (2001). Text linguistics: relevant linguistics? *Poetics, Linguistics and History: Discourses of War and Conflict. PALA Conference Papers 1999*. Potchefstroom University, South Africa, p. 588-595.

Ceri, S. et al. (2013). *Web Information Retrieval*. New York: Springer

Crestani, F. and Wu, S. (2006). Testing the cluster hypothesis in distributed information retrieval. *Information Processing and Management: an International Journal archive*. 42 (5). P. 1137 – 1150

Croft, B. W., Metzler D. and Strohman, T. (2010). *Search Engines: information retrieval in practice*. Boston: Pearson.

Dale, E. and Chall, J. S. (1949). The concept of readability. *Elementary English*. 26 (23). P. 19-26

De Beaugrande, R. A. and Dressler, W. U. (1981). *Introduction to text linguistics*. New York: Longman.

De Saussure, F. (1966). *Course in General Linguistics*. 3rd ed. New York: McGraw-Hill Book Company

Diggle, P. J. and Chetwynd A. G. (2011). *Statistics and Scientific Method: an Introduction for Students and Researchers*. Oxford: Oxford University Press

Dubay, W. H. (2004). *The Principles of Readability*. Costa Mesa, CA: Impact Information.

Dubin, D. (2004). The most influential paper Gerard Salton never wrote. *Library Trends*. 52 (4). P. 748-764.

Eroglu, S. (2013). Menzerath – Altmann law for distinct word distribution analysis in a large text. *Physica A*. 392 (12). P. 2775–2780

Espunya i Prat, A. (1994). Computational Linguistics: a Brief Introduction. *Links and Letters*. 1. P. 9-23.

Everitt, B. S. (2006). *The Cambridge Dictionary of Statistics*. 3rd ed. Cambridge: Cambridge University Press.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*. 27. P. 861–874

Fisher, R. A. (1950). *Statistical methods for research workers*. 11th ed. Rev. London: Oliver and Boyd.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*. 32 (200). P. 675 - 701.

Galvez, C., de Moya-Anegon, F. and Solana, V. H. (2005). Term conflation methods in information retrieval: Non-linguistic and linguistic approaches. *Journal of Documentation information (JDOC)*. 61 (4). P. 520-547

Gries, S. Th. (2013). *Statistics for linguistics with R*. 2nd rev. and ext. ed. Berlin: De Gruyter Mouton

Gries, S. (to appear in). Quantitative methods in linguistics. In Wright J. D. (ed.) *International Encyclopedia of the Social and Behavioral Sciences*. Amsterdam:

Elsevier. Retrieved 10 November 2014 from: http://www.linguistics.ucsb.edu/faculty/stgries/research/ToApp_STG_QuantMethInLing_IESBS2.pdf

Grossman, D.A. and Frieder, O. (2000). *Information Retrieval: Algorithms and Heuristics*. The Kluwer International Series in Engineering and Computer Science (SECS). 461. Boston: Kluwer

Haitao, L., Wei, H. (2012). Quantitative Linguistics : State of the Art, Theories and Methods. *Journal of Zhejiang University (Humanities and Social Science)*. 43(2). P. 178-192.

Harispe, S. et al. (2013). Semantic measures for the comparison of units of language, concepts or entities from text and knowledge base analysis. *Arxiv*. 1310 (1285). P. 1-159

Hauser, M.D., Chomsky, N. and Fitch, W.T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*. 22 (298 (5598)). Pp.1569-79

Hayes, B. (2010). *Introductory linguistics*. Los Angeles: Department of Linguistics University of California. Retrieved 13/12/2015 from: http://www.linguistics.ucla.edu/people/hayes/20/Text/HayesIntroductoryLinguistics2010.pdf

Hempel, C. G. (1964). *Fundamentals of Concept Formation in Empirical Science*. International Encyclopedia of Unified Science. 2 (7). Chicago: University of Chicago Press

Hoey, M. (2003). *Lexical priming and the properties of text*. University of Liverpool. Retrieved 14/09/2014 from: http://www.monabaker.com/tsresources/LexicalPrimingandthePropertiesof Text.hm

Hogan, P. C. (ed.) (2014). *The Cambridge Encyclopedia of the Language Sciences*. Cambridge: Cambridge University Press

Hrebicek, L. (2002). Zipf's Law and Text. *Glottometrics*. 3. P. 27-38.

Ingwersen, P. (1992). *Information Retrieval Interaction*. London: Taylor Graham Publishing

Johnson, K. (2008). *Quantitative Methods in Linguistics*. Malden (USA): Blackwell.

Joho, H. and Jose, J. M. (2006). A Comparative Study of the Effectiveness of Search Result Presentation on the Web. *Advances in Information Retrieval: Lecture Notes in Computer Science.* 3936. P. 302-313

Jurafsky, D. and Martin, H. J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Upper Saddle River, NJ: Prentice Hall PTR

Kamps, J. (2009). Presenting Structured Text Retrieval Results. *Encyclopedia of Database Systems*. P. 2130-2134

Karaman, B. I. (2003). *Polysemy in Natural Language: Case Studies on the Structural Description of Polysemous Lexemes in English, German and Turkish*. PhD thesis. University of Surrey. Retrieved 22/09/2014 from http://epubs.surrey.ac.uk/2816/1/412061.pdf

Kennedy, C. (2009). Ambiguity and Vagueness: an Overview. In: Maienborn, C., Heusinger, K. and Portner, P. (Eds.). *Semantics: an International Handbook of Natural Language Meaning*. Handbooks of Linguistics and Communication Science (HSK). 33 (1). Berlin: De Gruyter Mouton

Kovacs, E. (2011). Polysemy in Traditional vs Cognitive Linguistics. *Eger Journal of English Studies.* XI. P. 3–19

Kowalski, G. (2011). *Information Retrieval Architecture and Algorithms*. New York: Springer.

Kracht, M. (2007). *Introduction to linguistics*. Los Angeles: Department of Linguistics, UCLA. Retrieved 08/12/2014 from: http://wwwhomes.uni-bielefeld.de/mkracht/html/ling-intro.pdf

Kulacka, A. and Macutek, J. (2007). A discrete formula for the Menzerath - Altmann law. *Journal of Quantitative Linguistics*. 14 (1). P. 23-32

Kuropka, D. (2004). *Modelle zur Repräsentation natürlichsprachlicher Dokumente: Ontologie-basiertes, Information-Filtering und -Retrieval mit relationalen Datenbanken*. Advances in Information Systems and Management Science. 10. Berlin: Logos.

Lalkhen, A.G. and McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care and Pain*. 8. P. 221–223.

Lan, M., Tan, C. L. and Su, J. (2009). Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 31(4). P. 721-735

Legendre, P. (2005). Species Associations: The Kendall Coefficient of Concordance Revisited. *Journal of Agricultural, Biological, and Environmental Statistics*. 10 (2). P. 226–245.

Legendre, P. (2010). Coefficient of concordance. *Encyclopedia of Research Design*. 1. P. 164-169.

Lewis, D. D. and Sparck Jones, K. (1996). Natural language processing for information retrieval. *Communications of the ACM*. 39 (1). P. 92-101.

Manning, C., Raghavan, P. and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Manning, C. D. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Metzler, D. and Croft, B. W. (2005). Modeling Query Term Dependencies in Information Retrieval with Markov Random Fields. *SIGIR '05 Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. P. 472-479.

Mikk, J. (2005). Text comprehensibility (Textverständlichkeit). In Köhler, R., Altmann, G. and Piotrowski, R. G. (eds.) *Quantitative Linguistik - Quantitative Linguistics. Ein Internationales Handbuch*. Berlin: Walter de Gruyter

Mikros, G. and Milicka, J. (2014). Distribution of the Menzerath's law on the syllable level in Greek texts. In Altmann, G. et al. (eds.) *Empirical Approaches to Text and Language Analysis*. Lüdenscheid: RAM-Verlag.

Miller, G.A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Science*. 7 (3). P. 141–144

Moens, M. F. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context.* The Information Retrieval Series. Dordrecht: Springer.

Nadel L. (2005). Encyclopedia of Cognitive Science. Wiley

Neideen, T. and Brasel K. (2007). Understanding Statistical Tests. *Journal of Surgical Education*. 64 (2). P. 93-96

Onwuchekwa, E. O. (2011). Information Retrieval Methods in Libraries and Information Centers. *An International Multidisciplinary Journal, Ethiopia*. 5(6), P. 108-120.

Panik M. J. (2012). *Statistical Inference: a short course*. Hoboken, NJ: Wiley

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review*. 21 (5). P. 1112-1130

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*. 2 (1). P. 37-63

Power law: definitions of power law in Oxford Dictionary (British and World English) (2015). Retrieved 08/01/2015 from: http://www.oxforddictionaries.com/definition/english/power-law

Poulos, M. et al. (2007). Specific Selection of FFT Amplitudes from Audio Sports and News Broadcasting for Classification Purposes. *Journal of Graph Algorithms and Applications*. 11 (1). P. 277–307

Poulimenou S. et al. (2014). Keywords Extraction from Articles' Title for Ontological Purposes. In *Proceedings of the 2014 International Conference on Pure Mathematics, Applied Mathematics, Computational Methods (PMAMCM 2014)*. P. 120-125.

Poulimenou, S. et al. (to appear in). Short text coherence hypothesis. *Journal of Quantitative linguistics*. 22 (3).

Pulvermuller, F. (2003). *The Neuroscience of Language: On Brain Circuits of Words and Serial Order*. Cambridge: Cambridge University Press.

Raghavan, V. V. and Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*. 37 (5). P. 279-287.

Richards, J. C. and Schmidt, R. (2002). *Longman dictionary of language teaching and applied linguistics*. 3rd ed. London: Longman

Robertson, S. (2004). Understanding Inverse Document Frequency: On theoretical arguments for IDF. *Journal of Documentation*. 60 (5). P. 503–520

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*. 24 (5). P. 513-523. Ithaca: Department of Computer Science, Cornell University.

SAS Institute (1999). *SAS/STAT User's Guide*, Version 8. Cary, NC: SAS Institute, 1999. Retrieved 10/12/2014 from http://www.okstate.edu/sas/v8/saspdf/stat/chap13.pdf

Singhal, A. (2001). Modern Information Retrieval: a Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*. 24 (4). P. 35–43

Sorell, J. C. (2012). Zipf's Law and Vocabulary. In Chapelle, C. A. (ed.). *The Encyclopedia of Applied Linguistics*. Chisester, UK: Wiley-Blackwell

Sparck Jones, K. and Kay, M. (1977). Linguistics and information science: a postscript. In Walker, D., Karlgren, H., Kay, M. (Eds) *Natural Language in Information Science - Perspectives and Directions for Research*. Stockholm: Skriptor

Spoerri A. (1995). *InfoCrystal: a Visual Tool for Information Retrieval*. PhD Thesis. Massachusetts: MIT. Retrieved 10/12/2014 from http://comminfo.rutgers.edu/~aspoerri/InfoCrystal/InfoCrystal.htm

Squire, L. et al (eds.) (2008). *Fundamental Neuroscience*. 3rd ed. Boston: Elsevier / Academic Press.

Strzalkowski, T. et al. (1999). Evaluating natural language processing techniques in information retrieval: a TREC perspective. In Strzalkowski, T. (Ed.) *Natural Language Information Retrieval*. Dordrecht: Kluwer Academic Publishers.

Tanskanen, S. K. (2006). *Collaborating towards coherence*. Amsterdam: John Benjamins Publishing

Tesitelova, M. (1992). *Quantitative linguistics*. Linguistic and Literary Studies in Eastern Europe, 37. Amsterdam: John Benjamins.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*. 37 (1). P. 141-188.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. 2nd ed. London: Butterworth.

Woods, W. A. et al. (2000). Linguistic knowledge can improve information retrieval. In *ANLC '00 Proceedings of the sixth conference on applied natural language processing*. P. 262-267

Wong, H. B., Gek, Lim G. H. (2011). Measures of Diagnostic Accuracy: Sensitivity, Specificity, PPV and NPV. In *Proceedings of Singapore Healthcare*. 20:4. P. 316-318

Wu, S., Bi, Y. and Zeng, X. (2010). Retrieval Result Presentation and Evaluation. In *Proceeding KSEM'10 of the 4th international conference on Knowledge science, engineering and management*. P. 125-136.

Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law. *Library Trends*. 30 (1). P. 53-64.

Zamanian, M. and Heydari, P. (2012). Readability of Texts: State of the Art. *Theory and Practice in Language Studies*. 2 (1). P. 43-53.

Zamir, O. and Etzioni O. (1999). Grouper: a dynamic clustering interface to Web search results. *Proceeding WWW '99 Proceedings of the eighth international conference on World Wide Web archive*. P. 1361-1374. New York, NY: Elsevier

Zar, J. H. (2010). *Biostatistical Analysis*. 4th ed. Upper Saddle River, NJ: Prentice Hall

Zhu, W., Zeng N. and Wang N. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations. In *Proceedings of the NESUG Health Care and Life Sciences*. P. 1-9

ΕΛΛΗΝΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Αγγελής, Β. και Δημάκη, Κ. (2011). *Στατιστική*. Τόμος Ά: Στατιστική, Πιθανότητες, Στατιστική Συμπερασματολογία. Θεσσαλονίκη: Εκδόσεις Σοφία.

Αδαμόπουλος, Λ., Δαμιανός, Χ. και Σβέρκος Α. (2014). *Μαθηματικά και στοιχεία στατιστικής*. Αθήνα: Οργανισμός Εκδόσεως διδακτικών βιβλίων

Δαμιανού Χ., Παπαδάτος Ν. και Χαραλαμπίδης Χ. Α. (2003). *Εισαγωγή στις πιθανότητες και τη στατιστική: διδακτικές σημειώσεις*. Τμήμα Μαθηματικών Πανεπιστημίου Αθηνών, Αθήνα.

Εμβαλωτής, Α., Κατσής, Α. και Σιδερίδης Γ. (2006). *Στατιστική μεθοδολογία εκπαιδευτικής έρευνας*. Ιωάννινα: Πανεπιστήμιο Ιωαννίνων

Κατσάνος, Χ. και Αβούρης, Ν. (2008). Στατιστικές μέθοδοι ανάλυσης πειραματικών δεδομένων συνεργασίας . Στο Ν. Αβούρης, Χ. Καραγιαννίδης, Β. Κόμης (επιμ.) *Συνεργατική τεχνολογία, συστήματα, και μοντέλα συνεργασίας για εργασία, μάθηση, κοινότητες πρακτικής και δημιουργία γνώσης* (σελ. 483-516). Αθήνα: Εκδόσεις Κλειδάριθμος.

Μπούτσικας Ν. Β. (2003). *Σημειώσεις στατιστικής ΙΙ*. Τμήμα Οικονομικής Επιστήμης. Πανεπιστήμιο Πειραιώς. Retrieved 21/04/2015 from http://www.fme.aegean.gr/sites/default/files/cn/statii_ch1a_v3.pdf

Τσίμπος, Κ. και Γεωργιακώδης, Φ. (1999). *Περιγραφική και διερευνητική στατιστική ανάλυση δεδομένων*. 1. Αθήνα: Εκδόσεις Αθαν. Σταμούλης.

ΠΑΡΑΡΤΗΜΑ Α

Κώδικας εξαγωγής μεταβλητών V_w

```
num1=double('Advances knowledge discovery data mining');
 k1=find (num1==32);
 k2=length(k1);
 sol1=[];
 for i=2:1:k2;
    j=i-1;
    t=num1(k1(j)+1:k1(j+1)-1);
    sol=[i,length(t),sum(t)];
   sol1=[sol1;sol];
 end
        t=num1(1:k1(1)-1);
       solbeg=[1,length(t),sum(t)];
                    t=num1(k1(k2)+1:length(num1));
       sollast=[k2+1,length(t),sum(t)];
       sol1=[solbeg;sol1;sollast];
  % sol1=Vw word vector
  % word normalization vector Vwe
    d1=sol1/norm(sol1);
```

Κώδικας εξαγωγής μεταβλητών συνισταμένης (V_s)

```
" u=sum(d1)"
```

Κώδικας βαθμού προσαρμογής πολυώνυμου

```
c = polyfit(x,y,5);
xfit = linspace(min(x),max(x),length(x));
yfit = polyval(c,xfit);
plot(x,y,'o',xfit,yfit,'--')
```

Κώδικας *tf-idf*

```
D = {'recent','advances','chemical','synthesis','self','assembly','applications','FePt','nanoparticles';'advances','biotechnology','new','tools','future','pig','production','agriculture','biomedicine';'preimplantation','genetic','diagnosis','technological','advances','improve','accuracy','range','applications';'aliphatic','polyester','polymer','stars','synthesis','properties','applications','biomedicine','nanotechnology';'politics','life','itself','biomedicine','power','subjectivity','twenty','first','century';'recent','advances','nanobiotechnology','high','throughput','molecular','techniques','systems','biomedicine';'aging','stem','cells','regenerative','biomedicine','concepts','opportunities','technological','advances';'admixed','human','embryos','stem','cells','legislative','ethical','scientific','advances';'computer','aided','detection','diagnosis','breast','cancer','mammography','recent','advances';'advances','methods','assessing','tumor','hypoxia','vivo','implications','treatment','planning'}
A = unique(D)
m=length(A);
f=size(D);
n=f(1);
tel1=[];
for k = 1:n
   tel=[];
      for j=1:m
        seq_sum = sum(ismember( D(k,:), A(j)));
        tel=[tel, seq_sum];
      end
      tel1=[tel1;tel];
end
tel1=tel1'
Y = tfidf( tel1)
```

ΠΑΡΑΡΤΗΜΑ Β

Κώδικας consistency:

```
function [m,n,h,w1]=consistency(str);
% m the reject (1) of accept (0) of null hypothesis which is indicated that the three variables are not associated
% h are the propabilities for a number of words
% stri is the investigated sentence for example str1='returns a column vector containing the rows where matches were found'
% n is the number of the investigated words
a1=findstr(str, ' ');
k=length(a1);
m1=[]; xs1=[]; n1=[]; h=[]; w1=[];
for i=2:1:k
    str1=str(1:a1(i));
  [v1,v2,v3,n]=triplen(str1);
[m,w,xs,p]=kendall1(v1,v2,v3);
m1=[m1,m];
w1=[w1,w];
xs1=[xs1,xs];
h=[h,p];
end
% words
% v1 angles, v2 number of character, v3 sum of Asci , number of words
end
```

<u>Παράδειγμα εκτέλεσης αλγόριθμου consistency</u>

Χρησιμοποιώντας ως παράδειγμα το μικρό κείμενο:

"*Many theorists have suggested that working memory capacity plays a crucial role in reading comprehension however, traditional measures of short-term memory, like*

digit span and word span, are either not correlated or only weakly correlated with reading ability”.

Ο αλγόριθμος εφαρμόζεται αφού πρώτα πραγματοποιηθεί η γλωσσική προ-επεξεργασία (διακόπτουσες λέξεις κλπ.). Στη συνέχεια εκτελείται ο αλγόριθμος όπως φαίνεται στο παράδειγμα:

[m,n,h,w1]=consistency('Many theorists have suggested working memory capacity plays crucial role reading comprehension traditional measures short term memory like digit span word correlated weakly correlated reading ability')

Το *m* αντιστοιχεί στην απόρριψη (τιμή 1) ή αποδοχή (τιμή 0) της μηδενικής υπόθεσης, το *n* αντιστοιχεί στον αριθμό συστατικών του μικρού κειμένου και το *h* αντιστοιχεί στις υπολογισμένες πιθανότητες του ελέγχου υπόθεσης.

Αποτελέσματα εκτέλεσης αλγόριθμου για το παράδειγμα:

m =

0

n =

26

h =

Columns 1 through 6

0.4795 0.6456 0.5724 0.5125 0.4548 0.3487

Columns 7 through 12

0.3103 0.2306 0.2392 0.1531 0.0338 0.0061

Columns 13 through 18

0.0020 0.0017 0.0020 0.0011 0.0014 0.0009

Columns 19 through 24

0.0010 0.0009 0.0000 0.0000 0.0000 0.0000

w1 =

Columns 1 through 6

0.1667 0.1458 0.2222 0.2731 0.3127 0.3727

Columns 7 through 12

0.3933 0.4382 0.4282 0.4819 0.6352 0.7702

Columns 13 through 18

```
    0.8350    0.8242    0.7906    0.8096    0.7795    0.7870

  Columns 19 through 24

    0.7686    0.7595    0.9135    0.9495    1.0929    1.1489

>>
```

Κώδικας triplen:

```
function [v1,v2,v3,n] = triplen(str1)
num1=double(str1);
 k1=find (num1==32);
 k2=length(k1);
 he=k2;
 %n=he;
 sol1=[];
 for i=2:1:k2;
     j=i-1;
     t=num1(k1(j)+1:k1(j+1)-1);
     sol=[i,length(t),sum(t)];
    sol1=[sol1;sol];
 end
          t=num1(1:k1(1)-1);
         solbeg=[1,length(t),sum(t)];
                        t=num1(k1(k2)+1:length(num1));
         sollast=[k2+1,length(t),sum(t)];
         sol1=[solbeg;sol1;sollast];

         %%%%%%%%%%%%%%%%%%%%

       d1=sol1/norm(sol1);
```

```
        u=sum(d1);

%%%%%%%%%%%%%%%%%%%%%% complete the angle of vector
f1new=[];
t=size(sol1);
t1=t(1);
for n=1:1:t1;
CosTheta = dot(u, sol1(n,:))/(norm(u)*norm( sol1(n,:)));
f1 = acos(CosTheta)*180/pi;
f1new=[f1new,f1];
end
%%%%%%%%%%%%%%%%%%%%%%

h1=1:length(f1new);
h=[f1new',h1'];
g=sortrows(h);
angle=g(1:he);
order=g(:,2);
word1=sol1(order(1,:));
%%%%%%%%%%%%%%%%%%%%%%

 k1=find (num1==32);
 k1=[1,k1,length(num1)];
 k2=length(k1);
r=[];
 for i=1:1:he;
    l=k1(order(i));
    n=k1(order(i)+1);
    t=num1(l:n);
    r=[r,char(32),t];
 end
k2=length(sol1(:,1));
v1=angle';
v2=sol1(:,2);
```

```
v3=sol1(:,3);
n=length(v2);
return
```

Παράδειγμα εκτέλεσης αλγόριθμου triplen

Χρησιμοποιώντας ως παράδειγμα το μικρό κείμενο:

"*Many theorists have suggested that working memory capacity plays a crucial role in reading comprehension however, traditional measures of short-term memory, like digit span and word span, are either not correlated or only weakly correlated with reading ability"*.

Ο αλγόριθμος εφαρμόζεται αφού πρώτα πραγματοποιηθεί η γλωσσική προ-επεξεργασία (διακόπτουσες λέξεις κλπ.). Στη συνέχεια εκτελείται ο αλγόριθμος όπως φαίνεται στο παράδειγμα:

```
[v1,v2,v3,n] = triplen('Many theorists have suggested working memory capacity plays
crucial role reading comprehension traditional measures short term memory like digit
span word correlated weakly correlated reading ability')
```

Τα *v1, v2, v3* αντιστοιχούν στις τρείς μεταβλητές και το *n* αντιστοιχεί στον αριθμό συστατικών του μικρού κειμένου.

Αποτελέσματα εκτέλεσης αλγόριθμου για το παράδειγμα:

```
v1 =

    0.1143
    0.1509
    0.2112
    0.2222
```

0.2454

0.2462

0.3761

0.3909

0.4420

0.4610

0.5570

0.5834

0.5998

0.6646

0.7013

0.8377

0.8878

0.9117

0.9329

0.9435

0.9589

0.9832

1.0088

1.3744

1.5647

v2 =

4

9

4

9

7

6

8

5

7

4

7

13

11

8

5

4

6

4

5

4

4

10

6

10

7

7

v3 =

405

997

420

971

769

665

846

553

739

434

730

1402

1179

869

560

440

665

421

529

434

444

1061

653

1061

730

750

```
n =

    26
```

Κώδικας Kendall:

```
function [m,w,xs,p]=kendall1(v1,v2,v3)
%v1=[10.4,10.8,11.1,10.2,10.3,10.2,10.7,10.5,10.8,11.2,10.6,11.4]
k=length(v1);
%v2=[7.4,7.6,7.9,7.2,7.4,7.1,7.4,7.2,7.8,7.7,7.8,8.3]
%v3=[17,17,20,14.5,15.5,13,19.5,16,21,20,18,22]
%%%%%%%%%%%%%%%%%%%%%%%%% (first variable)
g=sort(v1);
g2=[];
for j=1:1:k
 g3=find(v1(j)==g);
 if length(g3)>1;
     g3=mean(g3);
 end
 g2=[g2,g3];
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
g=sort(v2);
g4=[];
for j=1:1:k;
 g5=find(v2(j)==g);
 if length(g5)>1
     g5=mean(g5);
 end
 g4=[g4,g5];
```

```
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

g=sort(v3);
g6=[];
for j=1:1:k
 g7=find(v3(j)==g);
 if length(g7)>1
     g7=mean(g7);
 end
 g6=[g6,g7];
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
e1=[];
e=[g2;g4;g6];
e=e';

for h=1:1:k;

e3=e(h,:);
e1=[e1,sum(e3)];
end
a=e1;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

 %function [r,t,p]=kendall(a,m)
m=3;
%a=[14.5,21,30.5,6,11,3.5,20,11.5,29,28.5,22.5,36]
 k=length(a);
 R2=[];
 for i=1:1:k;

     R1=a(i)^2;
```

```
   R2=[R2,R1];
end

R=sum(R2);

F=sum(a)^2;

w=(R-F/k)/(m^2*(k^3-k)/k);

xs=3*(k-1)*w;
%%%% degrees of freedom
df=k-1;

 p = 1 - chi2cdf(xs,df);

 if  0.001<p<0.005;
   m= 0;
 else
   m= 1;
 end
end
```

<u>Παράδειγμα εκτέλεσης αλγόριθμου Kendall</u>

Η εκτέλεση του αλγόριθμου αυτού, βασίζεται στις τιμές που εξάγονται από τον προηγούμενο αλγόριθμο, triplen, πάλι χρησιμοποιώντας για παράδειγμα το μικρό κείμενο:

"*Many theorists have suggested that working memory capacity plays a crucial role in reading comprehension however, traditional measures of short-term memory, like digit span and word span, are either not correlated or only weakly correlated with reading ability"*.

```
[m,w,xs,p]=kendall1(v1,v2,v3)
```

Το *m* συμβολίζει την απόρριψη ή αποδοχή της μηδενικής υπόθεσης, το *w* αντιστοιχεί στο βαθμό συσχέτισης Kendall, το *xs* στην τιμή του x^2 και το *p* στην αντίστοιχη πιθανότητα. Τα *v1, v2, v3* είναι ήδη γνωστά από τον παραπάνω αλγόριθμο, αντιστοιχούν στις τρείς μεταβλητές και το *n* αντιστοιχεί στον αριθμό συστατικών του μικρού κειμένου.

Αποτελέσματα εκτέλεσης αλγόριθμου για το παράδειγμα:

```
m =

     0


w =

    1.2479


xs =

   89.8469


p =

   1.5281e-09

>>
```