3. Ποσοτική & Υπολογιστική Γλωσσολογία

Περίληψη

Στο κεφάλαιο αυτό γίνεται εκτενής αναφορά στις επιμέρους Γλωσσικές Τεχνολογίες όπως ο η Ποσοτική & η υπολογιστική Γλωσσολογία. Η συνεκτικότητα ενός κειμένου οι αρχές κειμενικότητας ως επίσης και ην μετρική που έχει αναπτυχτεί σχετικά με την ποσοτική κατανόηση του κειμένου αποτελούν ζητήματα που θα εξετασθούν σε αυτό κεφάλαιο.

Προαπαιτούμενες γνώσεις

Ευρετηρίαση, Ποσοτική & Υπολογιστή Γλωσσική Τεχνολογία

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 σε προηγούμενο κεφάλαιο..

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 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» (nestedstructure) 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 DeBeaugrande 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.

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 DeBeaugrande και Dressler (1981), οι αρχές κειμενικότητας είναι οι ακόλουθες:

1. **Συνοχή** (cohesion): αφορά στους τρόπους σύνδεσης των συστατικών του επιφανειακού κειμένου (surfacetext), 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 DeBeaugrande και Dressler (1981) προσθέτουν πως εκτός των αρχών κειμενικότητας υπάρχουν ακόμη κανονιστικές αρχές (regulativeprinciples), προκειμένου να υπάρχει έλεγχος της επικοινωνίας του κειμένου και αναφέρουν τις δύο βασικότερες: αποδοτικότητα (efficiency), όταν απαιτείται η ελάχιστη προσπάθεια 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, 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 (InformationExtraction) 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.

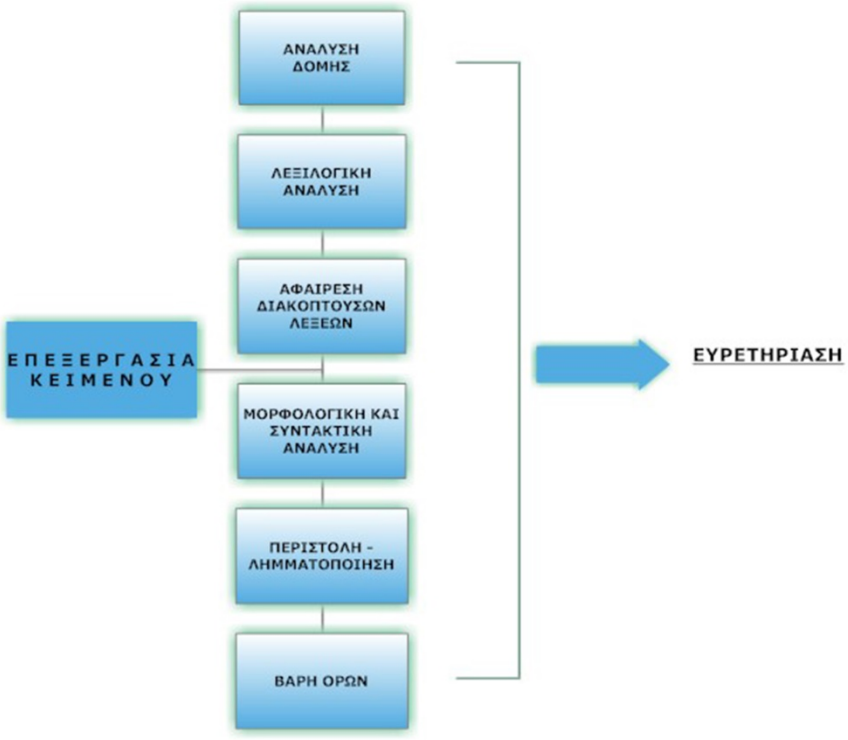
Με βάση λοιπόν τα παραπάνω, θα αναφερθούν επιγραμματικά οι συνηθέστερες διαδικασίες επεξεργασίας κειμένου, οι οποίες έχουν ενσωματωθεί στις μηχανές αναζήτησης για την ΑΠ σύμφωνα με τη βιβλιογραφία (Cerietal. 2013), καθώς έχουν επεξηγηθεί αναλυτικά στο προηγούμενο κεφάλαιο . Πρόκειται για τις διαδικασίες: ανάλυσης δομής εγγράφου, λεξιλογικής ανάλυσης, αφαίρεσης διακοπτουσών λέξεων, ανίχνευσης φράσεων (μέσω μορφολογικής και συντακτικής ανάλυσης), περιστολής - λημματοποίησης, απόδοσης βαρών όρων. Στο σχήμα 3.1 φαίνονται σχηματικά όλες οι παραπάνω 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 ZipfG. 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



Σχήμα 3.1. *Διαδικασίες επεξεργασίας κειμένου*

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. Ευφυΐα.
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 readabilityformulae, 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 .

Αναλυτικότερα σε σχέση με την εφαρμογή των readabilityformulae, 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 readabilityformulae, 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 (predictorvariables). 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 “readabilityformulae”:

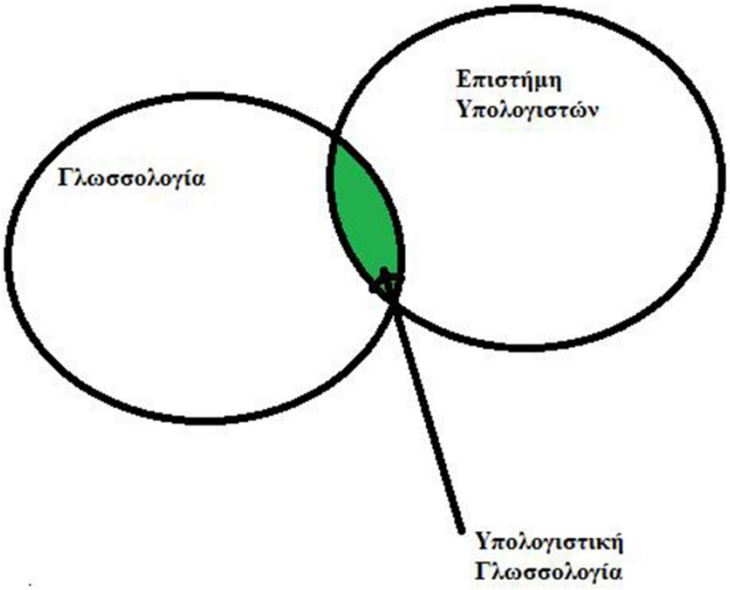
1. Μέσω των μετρήσεων ο συγγραφέας έχει στη διάθεσή του πληροφορίες ώστε να έρθει κοντά με το κοινό στο οποίο απευθύνεται και να μετατρέψει το κείμενο του σε ένα απλό κείμενο.
2. Η εφαρμογή τους πραγματοποιείται πριν το κείμενο φτάσει στον αναγνώστη μέσω υπολογιστών.

Ενώ παρουσιάζουν αντίστοιχα και τα ακόλουθα μειονεκτήματα:

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

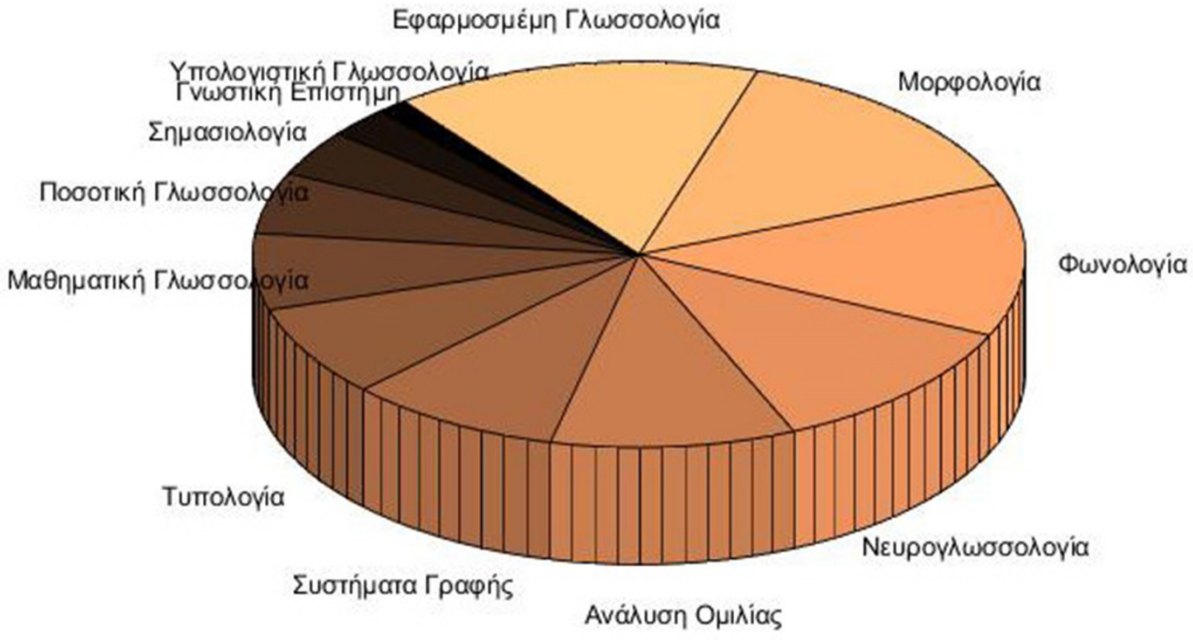
3.2 Εισαγωγή στην ΥΓ και τις βασικές της έννοιες

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 (spellchecking), 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 στο σχήμα 3.2.



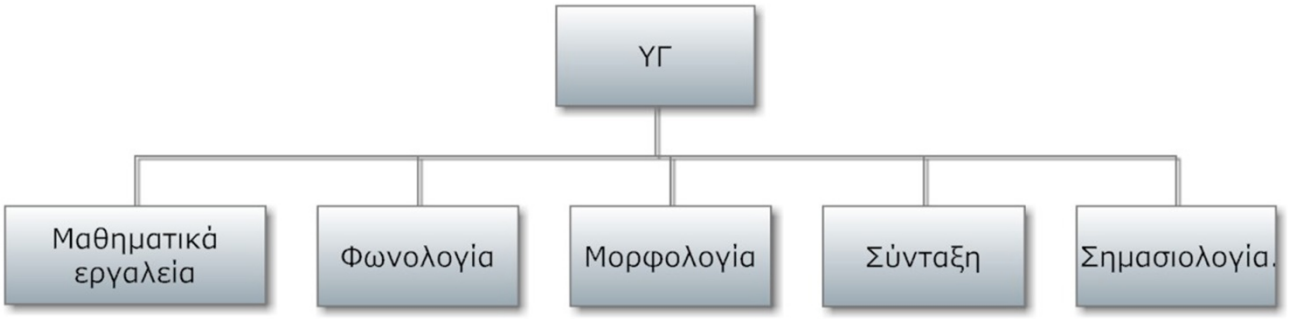
Σχήμα 3. 2. *Ο υπο-κλάδος της ΥΓ*

Γενικότερα, η ΥΓ σύμφωνα με την LINGUIST List αποτελεί έναν από τους πολλούς υπο-κλάδους της Γλωσσολογίας, όπως φαίνεται σχηματικά στο σχήμα 3.3. Στο διάγραμμα παρατηρούνται και άλλοι επιστημονικοί και διεπιστημονικοί κλάδοι στους οποίους έχει ήδη γίνει αναφορά όπως η Νευρογλωσσολογία και η Μορφολογία. Ο κλάδος της Μορφολογίας θα αναλυθεί παρακάτω.



Σχήμα 3. 3. *Υπο-κλάδοι Γλωσσολογίας*

Οι Bolshakov και Gelbukh (2004) αναφέρουν τα πεδία που μελετά η ΥΓ, που αφορούν στον προφορικό και γραπτό λόγο. Στο σχήμα 3.4 συνοψίζονται σε διαγραμματική μορφή:



Σχήμα 3. 4. *Πεδία έρευνας ΥΓ*

Αρχικά θα αναφερθούν κάποια βασικοί ορισμοί που χρησιμοποιούνται στην ΥΓ. Χρησιμοποιώντας 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* (wordoccurrence), 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* (wordform). 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 1 καθώς και ορισμένοι κανόνες παραγωγής:

Πίνακας 3.1. 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 → NPVP), είναι κατανοητό ότι η ονοματική και ρηματική φράση αποτελούν συστατικά της πρότασης.

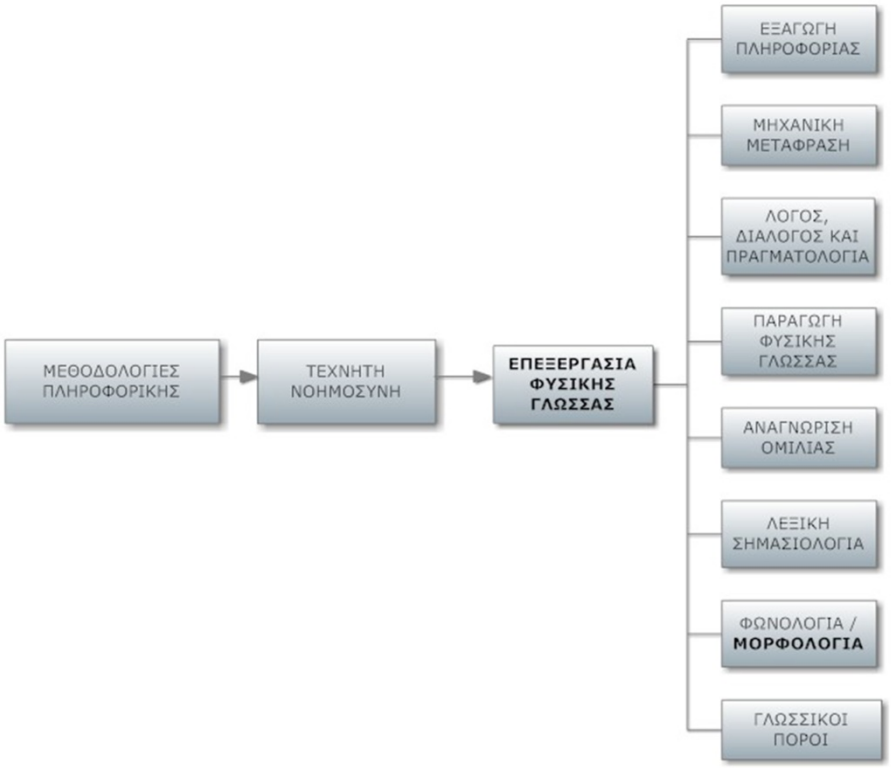
Στις παρακάτω ενότητες θα αναλυθούν τα πεδία της ΥΓ που αφορούν το γραπτό λόγο και τα κείμενα, δηλαδή η Μορφολογία, η Σύνταξη και η Σημασιολογία.

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 θεμελιώδη νοηματική μονάδα στη γλώσσα. Αποτελεί από μόνη της έναν ξεχωριστό υπο-κλάδο της Γλωσσολογίας και αποτελεί αντικείμενο μελέτης και για την ΥΓ ειδικότερα (βλέπε σχήμα 3.4).

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 όπως φαίνεται και στο σχήμα 3.5.

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.



Σχήμα 3. 5. *Επεξεργασία Φυσικής Γλώσσας και Μορφολογία*

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 (morphologicalparser) 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** (orthographicrules).

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 finitestateautomata (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 (mappingrules), οι οποίοι αντιστοιχίζουν μια λέξη σε μια ακολουθία μορφημάτων συνοδευμένη με επιμέρους μορφολογικά χαρακτηριστικά. Έτσι ένας μετατροπέας αντιστοιχίζει μεταξύ του finite-statetransducer (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.

3.2.3 Μέρη του λόγου και επισημειωτές

Σύμφωνα με τους Jurafsky και Martin (2000) ως **μέρη του λόγου** (partofspeech/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 (lexicaltags). 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 (automaticword-sensedisambiguatingalgorithms), 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** (closedclass), όπου τα μέλη τους είναι σταθερά σε αντίθεση με τις ανοικτές κλάσεις (openclass), 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 (partofspeechtagging/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. 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-BasedTagging ή Brilltagging, 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-freegrammar (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.

3.2.4 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*».

3.2.4.1 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. 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.

3.2.4.2 Μέτρα σημασιολογικής εγγύτητας (ομοιότητας)

Τα μέτρα σημασιολογικής εγγύτητας, όπως αναφέρουν οι Harispeetal. (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 (Harispeetal. 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 (Harispeetal. 2013) σύμφωνα με το είδος των στοιχείων προς σύγκριση, τη σημασιολογική πηγή (semanticproxy) από όπου προέρχονται οι πληροφορίες για τα στοιχεία και τέλος την κανονική μορφή (canonicalform) 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 (Harispeetal. 2013) η κλασική κανονική μορφή υλοποιείται μέσω της αναπαράστασης με διανύσματα, η οποία για την ΑΠ αντιστοιχεί στο γνωστό μοντέλο VSM .

Έτσι σύμφωνα με τη βιβλιογραφία (Harispeetal. 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.

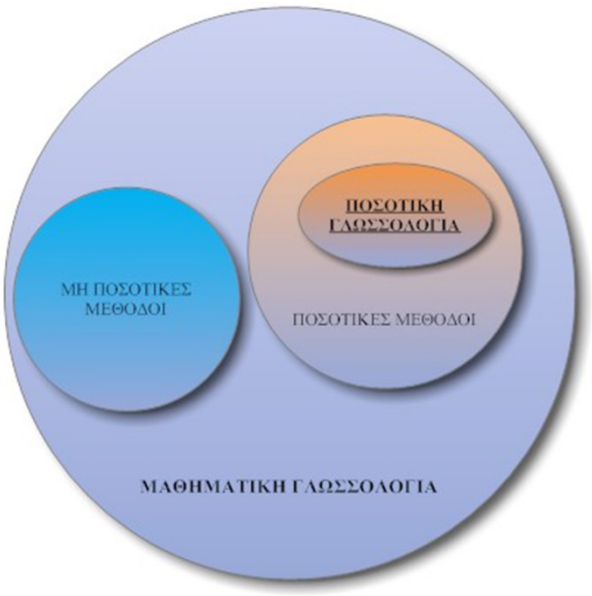
Ειδικότερη σημασία έχουν τα **κατανεμημένα** καθώς αφορούν σε κείμενα και χωρίζονται στις εξής προσεγγίσεις:

- **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 – setbased), 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 (QuantitativeLinguistics) ως υπο-κλάδο της Μαθηματικής Γλωσσολογίας (MathematicalLinguistics). Σύμφωνα με το Oxforddictionaries 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 (textprocessing), 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 στο σχήμα 3.6. 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.



Σχήμα 3. 6. *ΠΓ ως υπο-κλάδος της Μαθηματικής Γλωσσολογίας*

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. Ακόμη θα αναλυθεί η μεταξύ τους σχέση.

3.3.2 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 PaulMenzerath, 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 GabrielAltmann 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 ΜΑ 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 Baixeriesetal. (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 (3.1):

$$\frac{y'}{y} = -c \quad (3.1)$$

Μέσω της μαθηματικής ολοκλήρωσης (integration), ο παραπάνω μαθηματικός τύπος παίρνει τη μορφή του (3.2):

$$y = Ae^{-cx} \quad (3.2)$$

Όπου το x αντιστοιχεί στο δόμημα και το c σε μια μεταβλητή. Έτσι μέσω της εξίσωσης (3.3), 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 (monotonicdecreasingfunction).

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 (3.3) το συστατικό y'είναι ανάλογο του μέσου μήκους y και αντιστρόφως ανάλογο του μήκους x.

$$y' = \frac{cy}{x} \quad (3.3)$$

Σύμφωνα με τον Eroglu (2013), *«ο νόμος του ΜΑ αποτελεί ένα συνεχόμενο μοντέλο κατανομής πιθανοτήτων (probabilitydistributionmodel), το οποίο χρησιμοποιείται για να περιγράψει την πιθανοτική σχέση μεταξύ των διακριτών αποτελεσμάτων των ποσοτήτων»* όπως φαίνεται στο μαθηματικό τύπο (3.4):

$$y(x \mid A, b, c) = Ax^b e^{-cx} \quad (3.4)$$

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

3.3.3 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 GeorgeKingsleyZipf, καθηγητή φιλολογίας και γλωσσολόγος στο HarvardUniversity. Το 1935 στο βιβλίο του "*ThePsycho-BiologyofLanguage*", 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 (powerlaws). Με βάση το OxfordDictionary (2015) ως powerlaw 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 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 (3.5):

$$r \cdot f = c \quad (3.5)$$

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:

$$\log r \cdot \log f = \log c \quad (3.6)$$

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'scurves).

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 υπο-περίπτωσης 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 (3.7):

$$y(x \mid A, b) = Ax^{-b} \quad (3.7)$$

Συνεπώς κάτω από συγκεκριμένες συνθήκες ο νόμος του Zipf, αποτελεί υπο-περίπτωση του νόμου ΜΑ.

3.3.4 Θεωρία της Πληροφορίας και ΠΓ

Όπως ήδη έχει αναφερθεί παραπάνω 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 (ShannonC. E. 1998), καθώς θα επιχειρηθεί η σύνδεσή της με το σχήμα απόδοσης βάρους (tf – idf) και τη θεωρία πιθανοτήτων.

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 ηθεωρία του Shannon μπορεί να συνδεθεί με το συστατικό *idf*, το οποίο χρησιμοποιείται κυρίως στο σχήμα απόδοσης βαρών *tf – idf* . Το *idf* συστατικό ορίζεται από τον παρακάτω μαθηματικό τύπο (3.8):

$$idf(t_i) = \log \frac{N}{n_i} \quad (3.8)$$

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,αλλά αντεστραμμένο. Έτσι θα μπορούσε να υπολογιστεί η πιθανότητα για ένα τυχαίο έγγραφο να περιέχει έναν όρο t_i, από τον μαθηματικό τύπο (3.9):

$$P(t_i) = \frac{n_i}{N} \quad (3.9)$$

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 (3.10):

$$idf(t_i) = -\log P(t_i) \quad (3.10)$$

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.

Βιβλιογραφία

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.

Baeza-Yates, R., Ribeiro-Neto, B., & others. (1999). Modern information retrieval (T. 463). ACM press New York. Ανακτήθηκε από ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap10_user_interfaces_and_visualization-modern_ir.pdf

Baixeries, J., Hernandez-Fernandez, A., Forns, N., & Ferrer-i-Cancho, R. (2013). The parameters of the Menzerath-Altmann Law in genomes. Journal of Quantitative Linguistics, 20(2), 94–104.

Bendersky, M., & Croft, W. B. (2012). Modeling higher-order term dependencies in information retrieval using query hypergraphs. Στο Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (σσ 941–950). ACM. Ανακτήθηκε από http://dl.acm.org/citation.cfm?id=2348408

Blair, D. C., & Kimbrough, S. O. (2002). Exemplary documents: a foundation for information retrieval design. Information processing & management, 38(3), 363–379.

Bokos, M. P. G., Papavlasopoulos, N. K. S., & Avlonitis, M. (2007a). Specific selection of FFT amplitudes from audio sports and news broadcasting for classification purposes. Ανακτήθηκε από http://www.emis.ams.org/journals/JGAA/accepted/2007/Poulos+2007.11.1.pdf

Bokos, M. P. G., Papavlasopoulos, N. K. S., & Avlonitis, M. (2007b). Specific selection of FFT amplitudes from audio sports and news broadcasting for classification purposes. Ανακτήθηκε από http://www.emis.ams.org/journals/JGAA/accepted/2007/Poulos+2007.11.1.pdf

Bolshakov, I. A., & Gelbukh, A. (2004). Computational Linguistics Models, Resources, Applications.

Brants, T. (2003). Natural Language Processing in Information Retrieval. Στο CLIN. Ανακτήθηκε από http://www.clips.ua.ac.be/clin2003/proc/03Brants.pdf

Buk, S., & Rovenchak, A. (2008). Menzerath–Altmann Law for Syntactic Structures in Ukrainian. Glottotheory, 1(1), 10–17.

Büttcher, S., Clarke, C. L., & Cormack, G. V. (2010). Information retrieval: Implementing and evaluating search engines. Mit Press. Ανακτήθηκε από https://www.google.com/books

Carstens, W. (1999). Text-linguistics: Relevant linguistics. School of Languages and Arts, Potchefstroom University for CHE. Ανακτήθηκε από http://www.pala.ac.uk/uploads/2/5/1/0/25105678/carstens.pdf

Cerulo, L., & Canfora, G. (2004). A taxonomy of information retrieval models and tools. CIT. Journal of computing and information technology, 12(3), 175–194.

Crestani, F., & Wu, S. (2006). Testing the cluster hypothesis in distributed information retrieval. Information Processing & Management, 42(5), 1137–1150.

Croft, W. B., Metzler, D., & Strohman, T. (2010). Search engines: Information retrieval in practice. Addison-Wesley Reading. Ανακτήθηκε από http://library.mpib-berlin.mpg.de/toc/z2009_2465.pdf

Dale, E., & Chall, J. S. (1949). The concept of readability. Elementary English, 26(1), 19–26.

De Beaugrande, R. (1981). Introduction to text linguistics. Ανακτήθηκε από http://www.citeulike.org/group/236/article/202549

Diggle, P. J., & Chetwynd, A. G. (2011). Statistics and scientific method: an introduction for students and researchers. Oxford University Press.

Dubay, W. H. (2004). The Principles of Readability A brief introduction to readability research. Impact Information, Costa Mesa, CA.

Dubin, D. (2004). The most influential paper Gerard Salton never wrote. Ανακτήθηκε από https://www.ideals.illinois.edu/handle/2142/1697

Eroglu, S. (2013). Menzerath–Altmann law for distinct word distribution analysis in a large text. Physica A: Statistical Mechanics and its Applications, 392(12), 2775–2780.

Espunya i Prat, A. (1994). Computational linguistics: a brief introduction. Στο Links & Letters (σσ 009–23). Ανακτήθηκε από http://ddd.uab.cat/pub/lal/11337397n1/11337397n1p9.pdf

Everitt, B. S., & Skrondal, A. (2002). The Cambridge dictionary of statistics. Cambridge: Cambridge. Ανακτήθηκε από http://www.maa.org/publications/maa-reviews/the-cambridge-dictionary-of-statistics

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861–874.

Fisher, R. A. (1925). Statistical methods for research workers. Genesis Publishing Pvt Ltd.

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), 675–701.

Galvez, C., de Moya-Anegón, F., & Solana, V. H. (2005). Term conflation methods in information retrieval: non-linguistic and linguistic approaches. Journal of Documentation, 61(4), 520–547.

Gries, S. T. (2013). Statistics for linguistics with R: a practical introduction. Walter de Gruyter. Ανακτήθηκε από https://books.google.com/books?hl=en&lr=&id=-SFmpJ5N-98C&oi=fnd&pg=PP5&dq=Statistics+for+linguistics+with+R.+&ots=EJtROz94e-&sig=_7bDzRBLUXoXRBjUDqH1vqHTXFU

Grossman, D. A., & Frieder, O. (2012). Information retrieval: Algorithms and heuristics (T. 15). Springer Science & Business Media. Ανακτήθηκε από https://books.google.com/books

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2013). Semantic Measures for the Comparison of Units of Language, Concepts or Instances from Text and Knowledge Base Analysis. arXiv preprint arXiv:1310.1285. Ανακτήθηκε από http://arxiv.org/abs/1310.1285

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? science, 298(5598), 1569–1579.

Hayes, B., Curtiss, S., Szabolcsi, A., Stowell, T., Stabler, E., Sportiche, D., … others. (2001). Linguistics: An introduction to linguistic theory. Wiley-Blackwell.

Hempel, C. G. (1952). Fundamentals of concept formation in empirical science.(Int. Encyclopedia Unified Science, Vol. II. No. 7). Ανακτήθηκε από http://psycnet.apa.org/psycinfo/1953-03119-000

Hoey, M. (2004). Lexical priming and the properties of text. Corpora and discourse, 385–412.

Hogan, P. C. (2011). The Cambridge encyclopedia of the language sciences. Cambridge University Press Cambridge, NY. Ανακτήθηκε από http://www.langtoninfo.com/web_content

Hřebíček, L. (2002). Zipf's Law and Text. Glottometrics, 27.

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 relationalenDatenbanken. 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., Stamou, S., Papavlasopoulos, S., & Poulos, M. (χ.χ.). Short Text Coherence Hypothesis. Ανακτήθηκε από http://www.researchgate.net/profile/Marios_Poulos/publication/271167995_Short_Text_Coherence_Hypothesis/links/54bf5d570cf28ce68e6b4c71.pdf

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Ανακτήθηκε από http://dspace2.flinders.edu.au/xmlui/handle/2328/27165

Pulvermüller, F. (2002). The neuroscience of language: on brain circuits of words and serial order. Cambridge University Press. Ανακτήθηκε από https://www.google.com/books

Raghavan, V. V., & Wong, S. 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., & Schmidt, R. W. (2013). Longman dictionary of language teaching and applied linguistics. Routledge. Ανακτήθηκε από https://www.google.com/books

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. Journal of documentation, 60(5), 503–520.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513–523.

Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24(4), 35–43.

Spoerri, A. (1995). InfoCrystal, a visual tool for information retrieval. Massachusetts Institute of Technology. Ανακτήθηκε από http://dspace.mit.edu/handle/1721.1/36946

Squire, L., Berg, D., Bloom, F. E., du Lac, S., Ghosh, A., & Spitzer, N. C. (2012). Fundamental neuroscience. Academic Press. Ανακτήθηκε από https://www.google.com

Strzalkowski, T., Lin, F., Wang, J., & Perez-Carballo, J. (1999). Evaluating natural language processing techniques in information retrieval. Στο Natural language information retrieval (σσ 113–145). Springer. Ανακτήθηκε από http://link.springer.com/chapter/10.1007/978-94-017-2388-6_5

Tanskanen, S.-K. (2006). Collaborating towards coherence: Lexical cohesion in English discourse (T. 146). John Benjamins Publishing. Ανακτήθηκε από https://www.google.com/books

Teˇsˇitelova, M. (1983). On the state of Quantitative Linguistics in studies of Czech. Prague bulletin of mathematical linguistics, (40), 15–30.

Turney, P. D., Pantel, P., & others. (2010). From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37(1), 141–188.

Wei, L. H. H. (2012). Quantitative Linguistics: State of the Art, Theories and Methods. Journal of Zhejiang University (Humanities and Social Sciences), 2, 017.

Wong, H. B., & Lim, G. H. (2011). Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV. Proceedings of Singapore Healthcare▪ Volume, 20(4). Ανακτήθηκε από http://www.singhealthacademy.edu.sg/documents/publications/vol20no42011/11_stats_032-0911wonghb.pdf

Woods, W. A., Bookman, L. A., Houston, A., Kuhns, R. J., Martin, P., & Green, S. (2000). Linguistic knowledge can improve information retrieval. Στο Proceedings of the sixth conference on Applied natural language processing (σσ 262–267). Association for Computational Linguistics. Ανακτήθηκε από http://dl.acm.org/citation.cfm?id=974183

Wu, S., Bi, Y., & Zeng, X. (2010). Retrieval result presentation and evaluation. Στο Knowledge Science, Engineering and Management (σσ 125–136). Springer. Ανακτήθηκε από http://link.springer.com/chapter/10.1007/978-3-642-15280-1_14

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

Zamanian, M., & Heydari, P. (2012). Readability of texts: State of the art. Theory and Practice in Language Studies, 2(1), 43–53.

Zamir, O., & Etzioni, O. (1999). Grouper: a dynamic clustering interface to Web search results. Computer Networks, 31(11), 1361–1374.

Zar, J. H., & others. (1999). Biostatistical analysis. Pearson Education India.

Zhu, W., Zeng, N., Wang, N., & others. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland, 1–9.