Προβλήματα και μέθοδοι αυτόματης αναγνώρισης όρων σε υπολογιστικά συστήματα

ΣΟΦΙΑ ΑΝΑΝΙΑΔΟΥ & ΚΑΛΛΙΟΠΗ ΖΕΡΒΑΝΟΥ

Εισαγωγή

Οι ταχείς ρυθμοί τεχνολογικής και επιστημονικής εξέλιξης έχουν ως συνέπεια τη συνεχή δημιουργία νέων όρων. Οι όροι είναι σημαντικοί για την εξόρυξη γνώσης καθώς πραγματώνουν γλωσσικά τις έννοιες ενός επιστημονικού τομέα, εκφράζουν το σημαντικό περιεχόμενο ενός επιστημονικού κειμένου και το χαρακτηρίζουν σημασιολογικά. Υπάρχοντα λεξικά, αναγνωρισμένη ορολογία, ονοματολογίες, θησαυροί, κατηγοριοποιήσεις και βάσεις δεδομένων προσφέρουν μία μερική λύση στο πρόβλημα της συστηματικής καταγραφής, πρόσβασης και ανάκτησης αυτής της γνώσης. Η αυτόματη αναγνώριση επιστημονικής και τεχνικής ορολογίας είναι πρωταρχικής σημασίας για την ηλεκτρονική επεξεργασία γλώσσας γιατί μπορεί να βελτιώσει αισθητά την απόδοση συστημάτων μηχανικής μετάφρασης, εξαγωγής πληροφοριών, αυτόματης κατηγοριοποίησης και δεικτοδότησης (*indexing*) κειμένων και άλλων γλωσσικών εφαρμογών.

Σε αυτό το κεφάλαιο θα αναφερθούμε αρχικά στο πεδίο έρευνας της αυτόματης αναγνώρισης όρων, καθώς και στα προβλήματα που καλείται να δώσει λύσεις. Θα μιλήσουμε για τους τρόπους σχηματισμού όρων και για τις ιδιαιτερότητες που παρουσιάζει η ορολογία σε σχέση με τη γενική γλώσσα. Στο δεύτερο μέρος αυτού του κεφαλαίου, θα αναφερθούμε στις διάφορες μεθόδους και υπολογιστικά συστήματα για αυτόματη αναγνώριση όρων.

1 Ερευνητικό πεδίο και προβλήματα της αυτόματης αναγνώρισης όρων

Πέρα από τη συμβολή της σε γλωσσικές εφαρμογές και στην ανανέωση και εμπλουτισμό πόρων λεξικολογικών και πόρων γνώσης (*knowledge resources*), η αυτόματη αναγνώριση όρων μπορεί να συμβάλει καθοριστικά και στη διατήρηση *ορολογικής συνέπειας*. Δηλαδή μπορεί να συνεισφέρει στον περιορισμό φαινομένων, όπου ένας όρος αναφέρεται σε διαφορετικές έννοιες, ή αντιστρόφως, φαινομένων, όπου πολλοί όροι αντιστοιχούν σε μία έννοια, εντοπίζοντας και καταγράφοντας τέτοια φαινόμενα *ασάφειας*.

Η αυτόματη αναγνώριση όρων δεν αποτελεί εύκολο έργο καθώς η διαδικασία εντοπισμού όρων είναι συχνά δύσκολη ακόμη και για τους ειδικούς. Ο Διεθνής Οργανισμός Τυποποίησης (ISO) ορίζει ως *όρο* «*την απόδοση μιας συγκεκριμένης έννοιας μιας εξειδικευμένης γλώσσας σε μια γλωσσική έκφραση*» και ως *έννοια* «*...το νοητικό κατασκεύασμα που στοχεύει στην κατηγοριοποίηση μεμονωμένων αντικειμένων του εσωτερικού και του εξωτερικού κόσμου...*» (ISO704, 1986). Θεωρητικά οι όροι πρέπει να είναι *μονοσήμαντοι*, δηλαδή ένας και μόνο όρος πρέπει να αντιστοιχεί σε μία και μόνο έννοια. Στην πράξη όμως, πολλές φορές εμφανίζονται φαινόμενα *αμφισημίας* και *πολυσημίας*, όταν ένας όρος αντιστοιχεί σε πολλές έννοιες, και *συνωνυμίας*, όπου πολλοί όροι εκφράζουν την ίδια έννοια.

Ένα άλλο πρόβλημα για τον εντοπισμό όρων είναι ότι, πέρα από το ειδικό σημασιολογικό περιεχόμενό τους, δεν υπάρχουν γενικά και καθορισμένα κριτήρια για τη διάκρισή τους. Μια λεξικολογική μονάδα μπορεί να είναι όρος σε κάποιο ειδικό γνωστικό πεδίο, αλλά όχι σε κάποιο άλλο. Επίσης όροι, μπορεί να είναι και λέξεις προερχόμενες από τη γενική γλώσσα, οι οποίες όμως στο συγκεκριμένο τομέα χρησιμοποιούνται με ειδική έννοια.

Η απλή στατιστική μέτρηση της *συχνότητας εμφάνισης* χρησιμοποιείται στην λεξικογραφία για τον εντοπισμό μιας λέξης, γιατί μπορούμε να θεωρήσουμε ότι μια λέξη υπάρχει επειδή

την συναντούμε συχνά σε κείμενα. Ωστόσο, για τον εντοπισμό όρων δεν αποτελεί πάντα επαρκή και καθοριστική ένδειξη, καθώς πολλές λέξεις μπορεί να εμφανίζονται συχνά σε εξειδικευμένα κείμενα χωρίς να είναι όροι.

1.1 Τυπολογία σχηματισμού ορολογίας

Παρόλο που δεν υπάρχουν καθορισμένα κριτήρια διάκρισης μεταξύ όρων και άλλων λέξεων, γεγονός που αποτελεί και μια από τις βασικές δυσκολίες στην αυτόματη αναγνώριση όρων, υπάρχουν κάποιες διαφορές στους τρόπους σχηματισμού όρων σε σχέση με τις άλλες λέξεις.

Σε αντίθεση με τις λέξεις της γενικής γλώσσας, οι όροι σχηματίζονται με στόχο, όχι μόνο την ονομασία, αλλά και την συνειδητή κατηγοριοποίηση και ταξινόμηση εννοιών. Ο σχηματισμός όρων υπόκειται σε μια μελετημένη, συνειδητή προσπάθεια, που δρα καθοριστικά στις διεργασίες δημιουργίας τους. Όταν οι επιστήμονες ανακαλύπτουν νέες έννοιες και προσπαθούν τις ονομάσουν, συχνά χρησιμοποιούν συγκεκριμένα πρότυπα για την ονομασία τους. Αυτό βέβαια δεν αποκλείει και ad hoc ονομασίες. Για το σχηματισμό όρων διακρίνουμε τρεις τάσεις: την χρήση υπαρκτών πόρων, την τροποποίηση υπαρκτών πόρων και την χρήση νέων πόρων (νεολογισμοί).

Στην πρώτη περίπτωση, μπορεί να χρησιμοποιηθούν λέξεις που προέρχονται για παράδειγμα από την γενική γλώσσα, των οποίων το σημασιολογικό περιεχόμενο συρρικνώνεται σε μια συγκεκριμένη έννοια, ή επεκτείνεται σε μια νέα, όπως π.χ. για τις λέξεις «ποντίκι» «παράθυρο» και «ιστός» στην Πληροφορική.

Στην δεύτερη περίπτωση, μπορεί να σχηματιστούν νέοι όροι βασισμένοι σε υπάρχοντες όρους, είτε με χρήση μορφολογικών προσφυμάτων, π.χ. *methyltransfer-ase, act-in,* είτε με σύνθεση σε μορφή πολυλεκτικών όρων, π.χ. *protein kinase C*, είτε με σύντμηση σε μορφή ακρωνύμων, π.χ. *EGFR (epidermal growth factor receptor)*.

Τέλος, μπορεί να δημιουργούνται νεολογισμοί, εντελώς καινούριοι όροι, που είτε αποτελούν ad hoc ονομασίες, είτε σύνθετες λέξεις ή πολυλεκτικούς όρους, *νεοκλασικού τύπου*, δηλαδή όρους, που σχηματίζονται από λατινικά ή ελληνικά, θέματα ή μεμονωμένα γράμματα, π.χ. *protein kinase alpha, annexin II mRNA*. Επίσης μπορεί να χρησιμοποιούνται γράμματα και αριθμοί μαζί, π.χ. *NIH 3T3 fibroblasts, CYP1A1 promoter, Ca2+-calmodulin-dependent protein kinase IV*, ή κύρια ονόματα, επώνυμα ενδεχομένως των επιστημόνων που ανακάλυψαν τη νέα έννοια, π.χ. *Jurkat T cells*.

Ανάλογα με τον επιστημονικό χώρο κάποιες από τις παραπάνω τάσεις υπερισχύουν. Στην Ιατρική ή στη Βιολογία, για παράδειγμα, δημιουργούνται όροι βασισμένοι κυρίως σε προθήματα ή επιθήματα, ακρώνυμα, ή σύνθετες λέξεις νεοκλασικού τύπου, πιο συχνά από ότι σε άλλες επιστήμες. Κάποιες φορές η σημασία και η λειτουργία των προσφυμάτων είναι εξαιρετικά περιορισμένη, κάτι που δεν συμβαίνει στην γενική γλώσσα, κι εδώ ως χαρακτηριστικό παράδειγμα θα μπορούσε κανείς να αναφέρει όρους όπως *«γλυκ-όζη»* και *«μεθ-άν-ιο»* από το χώρο της Οργανικής Χημείας, όπου ο σχηματισμός νέων όρων είναι ιδιαίτερα ελεγχόμενος.

1.2 Προβλήματα στον εντοπισμό όρων: *ένθετοι όροι, ασάφεια* και *ποικιλότητα*

Όπως είδαμε, η δημιουργία πολυλεκτικών και σύνθετων όρων είναι ένας αρκετά συνηθισμένος τρόπος δημιουργίας όρων. Θα μπορούσε κανείς να πει ότι η πλειονότητα των όρων είναι σύνθετες πολυλεκτικές μονάδες. Οι περισσότεροι πολυλεκτικοί όροι είναι ουσιαστικά, που αποτελούνται, είτε από παράταξη δύο ή περισσοτέρων ουσιαστικών, π.χ. *bel-2 protein level,* είτε από κάποιο ουσιαστικό συνοδευόμενο από επίθετο, π.χ. *RA mediated tumor cell invasion*. Επίσης η ύπαρξη *ένθετων όρων* είναι αρκετά συνηθισμένη σε πολυλεκτικούς όρους, π.χ. [*Leukaemic* [*T cell*]] *line Kit225*] – όπου οι αγκύλες δηλώνουν ένθετους όρους μέσα στον πολυλεκτικό όρο.

Τα συστατικά πολυλεκτικών όρων μπορεί να συνδέονται με διάφορους τρόπους και ο εντοπισμός ένθετων όρων, ενώ μπορεί να είναι προφανής στον ειδικό επιστήμονα, δεν είναι εξίσου απλός για ένα αυτόματο σύστημα αναγνώρισης όρων. Η ύπαρξη πολυλεκτικών και ένθετων όρων αποτελεί μια από τις κύριες ερευνητικές προκλήσεις στο χώρο της αυτόματης

αναγνώρισης όρων, γιατί είναι στενά συνυφασμένη με τα προβλήματα *ασάφειας* και *ποικιλότητας* των όρων.

Η *ασάφεια* (*ambiguity*) είναι ένα γενικό φαινόμενο των φυσικών γλωσσών κι έχει εκτεταμένα απασχολήσει τον ερευνητικό χώρο της επεξεργασίας γλώσσας. Ειδικότερα στον τομέα της αυτόματης αναγνώρισης όρων η ασάφεια μπορεί να είναι *μορφοσυντακτική, συντακτική σημασιολογική,* και *ταξινόμησης*.
Μορφοσυντακτική ασάφεια έχουμε όταν μια λέξη μπορεί να ερμηνευθεί, για παράδειγμα, είτε ως ουσιαστικό, είτε ως ρήμα. Με την σημερινή εξέλιξη στον τομέα των ηλεκτρονικών μορφοσυντακτικών αναλυτών, αυτό το πρόβλημα μπορεί να περιοριστεί σημαντικά, καθώς οι αναλυτές μπορούν να εκπαιδευτούν σε κείμενα συγκεκριμένου τύπου και αντικειμένου. Επίσης, η πλειονότητα των όρων είναι συνήθως ουσιαστικά κι όχι ρήματα.
Καθαρά *συντακτική ασάφεια* έχουμε σε περιπτώσεις όπου είναι δύσκολο να διαχωριστεί ο προσδιορισμός από το προσδιοριζόμενο, ή είναι δύσκολος ο εντοπισμός του προσδιοριζομένου (*attachment ambiguity*) σε έναν πολυλεκτικό όρο που αποτελείται, για παράδειγμα, από πολλά ουσιαστικά. Η επίλυση ενός τέτοιου προβλήματος είναι συχνά συνδεδεμένη με την αναγνώριση ένθετων όρων.
Η *σημασιολογική ασάφεια* είναι ένα πιο δύσκολο πρόβλημα και οφείλεται σε φαινόμενα *πολυσημίας* και *ομωνυμίας*. *Πολυσήμαντοι όροι* μπορεί να έχουν την ίδια μορφή επιφανείας αλλά να αναφέρονται σε διαφορετικές συγγενείς έννοιες, είτε σε διαφορετικούς χώρους, είτε και στο ίδιο γνωστικό αντικείμενο. Στην πρώτη περίπτωση, για παράδειγμα, ο όρος «φακός / lens», μπορεί να πάρει τρεις διαφορετικές έννοιες στη Φωτογραφία, τη Φυσική και την Ανατομία αντίστοιχα. Στο χώρο της Φωτογραφίας αναφέρεται ως εξάρτημα φωτογραφικής μηχανής (π.χ. *«ευρυγώνιος φακός, φακός μεταβλητής εστιακής απόστασης»*) στο χώρο της Φυσικής αναφέρεται ως όργανο εστίασης ή τροποποίησης της κατεύθυνσης των ακτίνων φωτός, ήχου, ηλεκτρονίων, κτλ., ενώ στην Ανατομία, αναφέρεται στο φακό του ματιού (*«crystalline lens»*). Στην δεύτερη περίπτωση, ο όρος είναι πολυσήμαντος μέσα στο ίδιο γνωστικό αντικείμενο, όταν, για παράδειγμα, ο όρος «*γονίδιο*» έχει την έννοια *α) «του καταγεγραμμένου τμήματος του DNA που μεταφράζεται σε πρωτεΐνη» β) «της βιολογικού ενδιαφέροντος περιοχής του DNA που κατονομάζεται και φέρει ένα γενετικό φαινότυπο»* (Schulze-Kremer, 1998). Οι *ομώνυμοι όροι* από την άλλη πλευρά, έχουν την ίδια μορφή επιφανείας αλλά αναφέρονται σε εντελώς διαφορετικές έννοιες, όπως για παράδειγμα οι όροι «*plant*» (φυτό αλλά και βιομηχανική μονάδα) και «*cell*» (κύτταρο, κελί).
Τέλος, *ασάφεια ταξινόμησης* δημιουργείται όταν μία έννοια, και αντίστοιχα ο όρος που την εκφράζει, μπορεί να κατηγοριοποιηθεί με πολλούς τρόπους. Αυτού του είδους η ασάφεια ονομάζεται και *πολυ-διαστατικότητα* (*multi-dimensionality*) του όρου και εμφανίζεται ως πολλαπλή κληρονομικότητα σε μια ιεραρχία όρων, δηλαδή ένας κόμβος σε μια δενδρική ιεραρχία εμφανίζεται να έχει περισσότερους από έναν γονικούς κόμβους. Για παράδειγμα, ο όρος «*pulmonary tuberculosis*» μπορεί να ταξινομηθεί και ως ασθένεια του αναπνευστικού συστήματος και ως μεταδοτική ασθένεια.
Η άρση της ασάφειας ενός όρου είναι σημαντική για την αυτόματη αναγνώριση όρων γιατί επιτρέπει την κατηγοριοποίηση των όρων ανάλογα με τη σημασία τους. Ο βαθμός λεπτομέρειας, που επιθυμούμε να έχουμε, στην ταξινόμηση και διάκριση μεταξύ των ορολογικών εννοιών εξαρτάται από την εφαρμογή (π.χ. εξαγωγής πληροφοριών ή ανάπτυξης οντολογίας). Η σημασιολογική αποσαφήνιση των όρων μπορεί να επιτευχθεί με στατιστικές *μετρήσεις ομοιότητας* μεταξύ των εξαχθέντων όρων και του γλωσσικού τους περιβάλλοντος.

Ποικιλότητα (*term variation*) παρουσιάζεται όταν μία έννοια εκφράζεται με πολλούς όρους, συνώνυμους ή παραλλαγές του ίδιου όρου (*term variants*). Η ποικιλότητα είναι ένα αρκετά διαδεδομένο φαινόμενο στην ορολογία. Υπολογίζεται ότι το 37% των όρων που εμφανίζονται σε ένα κείμενο αποτελούν παραλλαγές όρων (Jaquemin, 1999). Η ποικιλότητα των όρων μπορεί να κυμαίνεται από απλή ορθογραφική διαφοροποίηση μέχρι και σημασιολογική. Ο εντοπισμός και η αναγνώριση παραλλαγών όρων είναι κεφαλαιώδους σημασίας για εφαρμογές όπως η εξαγωγή πληροφοριών, η ανάκτηση πληροφοριών και η

δημιουργία θησαυρών. Είναι σημαντικό να γνωρίζει κανείς κατά πόσον διαφορετικές επιφανειακές μορφές όρων αναφέρονται στην ίδια έννοια ή δεν σχετίζονται μεταξύ τους.
Κατά τον Jaquemin (Jaquemin, 1999) η ποικιλότητα μπορεί να είναι είτε καθαρά *μορφολογική*, *συντακτική*, ή *σημασιολογική*, είτε συνδυασμός αυτών των διαφοροποιήσεων.
Στην περίπτωση της *μορφολογικής ποικιλότητας* η διαφοροποίηση οφείλεται σε χρήση μορφολογικών καταλήξεων, είτε γραμματικών, π.χ. Ενικός/Πληθυντικός, *biochemical study / biochemical studies*, είτε παραγωγικών, π.χ. *enzyme activity / enzymatic activity*. Στην περίπτωση της *συντακτικής ποικιλότητας* μπορεί, για παράδειγμα, να έχουμε παρεμβολή (*insertion*) ενός προσδιορισμού, π.χ. *human clones / human DNA clones* ή ανάπτυξη (*expansion*) και αντιμετάθεση προσδιορισμού (*permutation*), π.χ. *genetic disease / disease is genetic, enzyme activity / activity of enzyme*. Η συντακτική ποικιλότητα μπορεί επίσης να οφείλεται σε διάφορους τύπους παρατακτικής σύνδεσης όρων, με χρήση συμπλεκτικού ή διαζευκτικού συνδέσμου, π.χ. *enzyme and bactericical activity*, ή με χρήση κόμματος, παύλας ή καθέτου, π.χ. *immuno-affinity chromography / immuno-, steroid-, and site specific DNA-affinity chromography*. Η *σημασιολογική ποικιλότητα* παρατηρείται με τη χρήση συνωνύμων, π.χ. *genetic disease / hereditary disease*.
Πέρα από αυτές τις κατηγορίες και τους συνδυασμούς τους, μπορεί να εμφανίζονται παραλλαγές όρων λόγω ορθογραφικών διαφοροποιήσεων, π.χ. *signalling pathway* (Αγγλικά Η.Β.) / *signaling pathway* (Αγγλικά Η.Π.Α.) ή *amino acid / amino-acid* (χρήση παύλας). Επίσης ακρώνυμα και συντομευμένες εκδοχές είναι χαρακτηριστικά είδη παραλλαγών της ανεπτυγμένης μορφής ενός όρου.
Η *ποικιλότητα*, ιδιαίτερα στις περιπτώσεις σύντμησης και περικοπής όρων δημιουργεί σοβαρά προβλήματα ασάφειας. Οι συντετμημένοι όροι και τα ακρώνυμα πολλές φορές αναφέρονται σε διαφορετικούς ανεπτυγμένους όρους. Για παράδειγμα, το ακρώνυμο ATR μπορεί στην ανεπτυγμένη μορφή του να αναφέρεται σε *Above Threshold Reporting, Air Turbo Rocket, Acceptance Test Report, Automatic Target Recognition, Aerial Target Review, Advanced Tactical Radar,…*κ.ά. μέσα στο ίδιο γνωστικό αντικείμενο (Mountain Data Systems).
Ερευνητικές προσπάθειες για τον εντοπισμό παραλλαγών όρων περιλαμβάνουν τη χρήση stemmer, δηλαδή συστημάτων που διαχωρίζουν το θέμα από την κατάληξη μιας λέξης, τη χρήση μορφολογικών αναλυτών, ηλεκτρονικών λεξικών και ορολογικών βάσεων, μορφοσυντακτικών αναλυτών, στατιστικών μετρήσεων και γλωσσολογικών κανόνων. Ένα από τα πιο γνωστά συστήματα αναγνώρισης παραλλαγών όρων είναι το σύστημα FASTR (Jaquemin, 1999).

2 Μέθοδοι ερευνητικής προσέγγισης

Οι ερευνητικές προσεγγίσεις για την αντιμετώπιση των ιδιαιτεροτήτων και των προβλημάτων που παρουσιάζει η αναγνώριση όρων έχουν κατά καιρούς ως αφετηρία τεχνικές και θεωρίες από άλλους ερευνητικούς τομείς. Έτσι, θα μπορούσε κανείς να τις διακρίνει ανάλογα με τον ερευνητικό χώρο από τον οποίο εμπνέονται:

- Μεθόδους που προέρχονται από το χώρο της Επιστήμης της Πληροφόρησης (*Information Science*), όπως στατιστικές μεθόδους και άλλες τεχνικές δεικτοδότησης (indexing), κατηγοριοποίησης και ομαδοποίησης κειμένων προερχόμενες από το ερευνητικό αντικείμενο της Ανάκτησης Πληροφοριών.
- Μεθόδους που βασίζονται σε Γλωσσολογικές θεωρίες, κυρίως των τομέων λεξικολογικής, μορφολογικής και μορφοσυντακτικής έρευνας.
- Υβριδικές μέθοδοι, που συνήθως συνδυάζουν τεχνικές στατιστικής με γλωσσολογικές θεωρίες και μεθόδους.
- Και τέλος, μέθοδοι, που προέρχονται από τον ερευνητικό χώρο της Τεχνητής Νοημοσύνης και που εφαρμόζουν διαφόρους αλγορίθμους μηχανικής μάθησης (*Machine Learning algorithms*).

Στη συνέχεια λοιπόν, θα αναφερθούμε σε γενικές γραμμές στις θεωρητικές βάσεις κάθε μιας από αυτές τις ερευνητικές προσεγγίσεις και σε συστήματα αναγνώρισης όρων που τις

εφαρμόζουν, προσπαθώντας να καταδείξουμε τα πλεονεκτήματα και τα μειονεκτήματα κάθε μεθόδου.

2.1 Στατιστικές μέθοδοι

Τα τελευταία χρόνια παρατηρείται γενικότερα μια αύξηση στη χρήση στατιστικών μεθόδων σε εφαρμογές και συστήματα επεξεργασίας γλώσσας. Στον τομέα της αναγνώρισης όρων, οι στατιστικές μέθοδοι που χρησιμοποιήθηκαν αρκετά σε αρχικό στάδιο προέρχονταν κυρίως από την ερευνητική παράδοση του τομέα της ανάκτησης πληροφοριών, όπως θα δούμε και στη συνέχεια. Έτσι χρησιμοποιήθηκαν αρχικά μετρήσεις *συχνότητας εμφάνισης* και μετρήσεις *κατανομής* (*distribution measures*) και *ομοιότητας* (*similarity measures*) για τον εντοπισμό και την κατηγοριοποίηση όρων αντίστοιχα.

Η *συχνότητα εμφάνισης* (*frequency of occurrence*) είναι από τις πιο απλές και τις πιο δημοφιλείς μεθόδους στην αυτόματη εξαγωγή όρων. Μπορεί να εφαρμοστεί ανεξάρτητα από το αντικείμενο και τη φύση του κειμένου και δεν χρειάζεται άλλες εξωτερικές πηγές. Βασίζεται στη λογική *Termhood(a) = f(a),* όπου η πιθανότητα μια *στοιχειοσειρά* (*string*) *a* να αποτελεί όρο ισούται με την συχνότητα εμφάνισης του *α* στο κειμενικό σώμα. Πολλές μέθοδοι μέχρι σήμερα, χρησιμοποιούν τη συχνότητα εμφάνισης για την κατηγοριοποίηση εξαγόμενων πιθανών όρων (Dagan & Church, 1994, Enguehard & Pantera, 1994, Lauriston, 1996, κ.ά.). Στις μετρήσεις συχνότητας εμφάνισης ο πιθανός πολυλεκτικός ή σύνθετος όρος αντιμετωπίζεται ως μια μονάδα, χωρίς να χρησιμοποιούνται άλλες πληροφορίες για τα συστατικά του. Όπως θα δούμε και στη συνέχεια, τέτοιου είδους μετρήσεις δεν αρκούν, αν δεν συνοδεύονται και από κάποιες γλωσσικές πληροφορίες. Κι αυτό γιατί στην εξαγωγή όρων, μας ενδιαφέρουν τόσο όροι που εμφανίζουν υψηλή συχνότητα, όσο και όροι σπάνιοι ή όροι με μέτριες τιμές εμφάνισης. Ωστόσο, η μέτρηση αυτή παρουσιάζει αρκετά καλά αποτελέσματα στον εντοπισμό αμετάβλητων φράσεων (*fixed phrases*). Σήμερα χρησιμοποιείται σε υβριδικά συστήματα αναγνώρισης όρων σε συνδυασμό με κάποιο γλωσσικό φίλτρο αποδεκτών συντακτικών μοτίβων, όπως στο TERMINO (Lauriston, 1994), TERMIGHT, (Dagan & Church, 1995) και Enguehard & Pantera (1994).

Μια άλλη μέτρηση στατιστική προερχόμενη από το χώρο της λεξικογραφίας και της γλωσσολογικής έρευνας κειμενικών σωμάτων (*Corpus Linguistics*) είναι αυτή της *αμοιβαίας πληροφορίας* (*Mutual Information*). Μετρήσεις αμοιβαίας πληροφορίας χρησιμοποιήθηκαν αρχικά για τον εντοπισμό λέξεων που εμφανίζονται συνήθως μαζί σε ιδιωματικές εκφράσεις (*collocations*), όπως για παράδειγμα «*μέρα μεσημέρι*», «*είναι στα μαχαίρια*», «*παίρνω δρόμο*», «*κόβω βόλτες*», «*έκανα μαύρα μάτια να την δω*», «*κάνω υπομονή*», κ.ά. Σε τέτοιες περιπτώσεις μία λέξη έχει πολύ στενή σχέση με αυτή που τη συνοδεύει στη φράση. Υπάρχουν λοιπόν σαφείς ομοιότητες μεταξύ αυτών των εκφράσεων της γενικής γλώσσας και των όρων, γιατί και στις δύο περιπτώσεις η σχέση των λέξεων που τις απαρτίζουν παρουσιάζει μια σταθερότητα. Δηλαδή δεν θα έλεγε κανείς, για παράδειγμα, «*κόβω περίπατο*» αντί για «*κόβω βόλτες*» ή «*παίρνω τον δρόμο*» αντί για «*παίρνω δρόμο*». Όπως και οι ιδιωματικές εκφράσεις έτσι και οι όροι αναφέρονται σε μια συγκεκριμένη έννοια και επιδέχονται συγκεκριμένους προσδιορισμούς. Η αμοιβαία πληροφορία στατιστικά ορίζεται ως *το ποσό της πληροφορίας που παρέχεται από την εμφάνιση του γεγονότος* y_i *για την εμφάνιση του γεγονότος* x_k *ως:*

$$I(x_k y_i) \equiv \log P(x_k y_i) / P(x_k)\, P(y_i)$$ Fano (1961:27-28)

Η μέτρηση αμοιβαίας πληροφορίας μας δείχνει τι πληροφορία μας δίνει μια λέξη για άλλες που δύνανται να εμφανιστούν μαζί της. Προβλήματα σε τέτοιου είδους μετρήσεις αποτελεί η έλλειψη επαρκών δεδομένων (*data sparseness*) γιατί υπερεκτιμώνται προτεινόμενα ζεύγη λέξεων (*bigrams*), που αποτελούνται από λέξεις με χαμηλή συχνότητα στο κειμενικό σώμα (Kita et al., 1994). Επίσης οι μετρήσεις γίνονται εξαιρετικά περίπλοκες όταν χρησιμοποιούνται για φράσεις άνω των δύο λέξεων (Frantzi & Ananiadou, 1995). Μετρήσεις

αμοιβαίας πληροφορίας έχουν χρησιμοποιηθεί σε μεθόδους εντοπισμού πιθανών όρων που αποτελούνται από δύο λέξεις από τους Damerau (1993) και Daille et al. (1994).
Σε έρευνες που έγιναν από τους Daille et al. (1994) μελετήθηκαν για την εξαγωγή διλεκτικών όρων και άλλες στατιστικές μετρήσεις, όπως μετρήσεις του *συντελεστή* Φ^2 (Gale & Church, 1991) και του *συντελεστή λογαριθμικού τύπου* (*loglike coefficient*) (Dunning, 1993). Όμως τα αποτελέσματα που αναφέρονται κατά Daille et al. (1994), δεν παρουσιάζουν σημαντική βελτίωση. Υποστηρίζεται ότι ο συντελεστής λογαριθμικού τύπου αποτελεί καλή μέτρηση για υποψήφιους όρους μεγάλης συχνότητας. Ο συντελεστής λογαριθμικού τύπου δεν χρησιμοποιείται για όρους άνω των δύο λέξεων.

Για τον εντοπισμό συσχετισμών λέξεων σε πολυλεκτικούς όρους προτείνονται μετρήσεις της C/ NC τιμής (Frantzi & Ananiadou, 1999). Ως τιμή C ορίζεται ο συσχετισμός της συνολικής τιμής συχνότητας εμφάνισης μιας σειράς λέξεων σε ένα σώμα κειμένων, με τη συχνότητα εμφάνισης της σειράς λέξεων ως τμήμα μεγαλύτερων πιθανών όρων, τον αριθμό αυτών των μεγαλύτερων όρων και την έκταση σε αριθμό λέξεων αυτής της σειράς:

C-value (a) $= log_2 | a |$ *f(a)* , για τους μη ένθετους όρους και:

$$\textit{C-value (a)} = log_2 | a | \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right)$$, για ένθετους όρους

όπου *α* ο υποψήφιος πολυλεκτικός όρος, | *α* |, η έκταση του πολυλεκτικού όρου σε λέξεις και *f(a)*, η συχνότητα εμφάνισής του είτε μεμονωμένα, είτε ως ένθετου όρου, μέσα σε άλλους πολυλεκτικούς όρους. Κατά αυτό τον τρόπο εντοπίζονται κι εξάγονται πολυλεκτικοί και ένθετοι όροι της μορφής: *adenoid cystic basal cell carcinoma, cystic basal cell carcinoma, ulcerated basal cell carcinoma, recurrent basal cell carcinoma, basal cell carcinoma*.
Η τιμή *NC* αποδίδει κάποιους συντελεστές βαρύτητας (*weights*) σε *συγκείμενες λέξεις* (*context words*) και συγκεκριμένα σε ρήματα, ουσιαστικά κι επίθετα, που εμφανίζονται στο γλωσσικό περιβάλλον του όρου. Υπολογίζει τον αριθμό των όρων στους οποίους εμφανίζεται μια συγκείμενη λέξη, τη συχνότητά της ως συγκείμενη λέξη και τη συνολική συχνότητά της στο κειμενικό σώμα. Ο συντελεστής βαρύτητας αποδίδει υψηλότερες τιμές σε λέξεις που τείνουν να εμφανίζονται με όρους:

$$weight(w) = \frac{T(w)}{n}$$

Η τιμή NC στη συνέχεια ορίζεται ως εξής:

$$\textit{NC-value(a)} = 0.8 * \textit{C-value(a)} + 0.2 * \textit{CF(a)}$$

όπου *α* είναι ο πιθανός προτεινόμενος όρος, *C-value(α)* είναι η τιμή C για αυτό τον όρο *α*, και *CF(a)* είναι ο συντελεστής γλωσσικού περιβάλλοντος, που υπολογίζεται πολλαπλασιάζοντας το σύνολο των συντελεστών βαρύτητας των συγκειμένων λέξεων με την συχνότητα συνεμφάνισής τους με αυτό τον όρο.
Η αναγνώριση όρων με βάση την τιμή *NC* κατά την αξιολόγηση εμφανίζει απόδοση που κυμαίνεται μεταξύ 91% και 99% (*Nuclear Receptors Corpus*) (Ananiadou et al. 2000).

2.2 Αυτόματη Δεικτοδότηση

Η βασική υπόθεση στην διαδικασία δεικτοδότησης είναι ότι μια λέξη, ή περισσότερες λέξεις ενός κειμένου, που εμφανίζονται συχνότερα σε αυτό το κείμενο σε σχέση με άλλα, μπορούν να είναι ενδεικτικές του περιεχομένου του, ως λέξεις κλειδιά (*key words*). Αυτές οι λέξεις ονομάζονται ενδεικτικοί όροι (*index terms*) και αφού εντοπιστούν μπορεί να αποτελέσουν τη βάση για την ταξινόμηση ενός κειμένου και μια γενική «ετικέτα», κατά κάποιο τρόπο, του κειμένου, που το διαφοροποιεί σε σχέση με άλλα και μπορεί να χρησιμοποιηθεί για την αναζήτηση και την ανάκτησή του μέσα από μια μεγάλη συλλογή κειμένων. Χαρακτηριστική

εφαρμογή μιας τέτοιας διαδικασίας βρίσκουμε στις μηχανές αναζήτησης (*search engines*), όπου αφού υποβάλουμε ένα ερώτημα με τη μορφή λέξης κλειδί, η μηχανή αναζήτησης ανακτά από το σύνολο των ιστοσελίδων που έχει κατηγοριοποιήσει, τις ιστοσελίδες που αντιστοιχούν στο ερώτημά μας.
Βασισμένα σε αυτή τη λογική, τα συστήματα αυτόματης δεικτοδότησης κειμένων προσπαθούν να εντοπίσουν χρήσιμους ενδεικτικούς όρους. Για το σκοπό αυτό χρησιμοποιούνται διάφορες στατιστικές μετρήσεις και κυρίως μετρήσεις, της απλής *συχνότητας εμφάνισης* μιας λέξης, ή φράσης σε ένα κείμενο, και *μετρήσεις κατανομής* των λέξεων σε μια συλλογή κειμένων, που μπορεί να μας δώσουν μια πιο ακριβή εικόνα για τις διαφορές μεταξύ τους.

Το σύστημα FASIT ήταν ένα από τα πρώτα συστήματα αναζήτησης ενδεικτικών όρων με χρήση κάποιων πολύ απλών εργαλείων επεξεργασίας γλώσσας. Επιχειρεί να «*εντοπίσει κειμενικές μονάδες που φέρουν πληροφορίες για το περιεχόμενο του κειμένου χωρίς πλήρη ανάλυση και, χωρίς χρήση σημασιολογικών κριτηρίων, ομαδοποιεί αυτές τις μονάδες σε σύνολα ημι-συνώνυμων μονάδων*» (Dillon & Gray, 1983:99).
Το FASIT επεξεργάζεται κείμενα σε δύο στάδια: εντοπισμού όρων και ομαδοποίησης. Αρχικά γίνεται μορφοσυντακτική ανάλυση του κειμένου με βάση λεξικό που περιέχει πληροφορίες για την δυνατή μορφοσυντακτική κατηγορία των λημμάτων και κυρίως των καταλήξεων. Θα πρέπει να σημειωθεί εδώ ότι ως κατάληξη, ορίζονται πολύ μηχανικά, οι τελευταίοι χαρακτήρες μιας λέξης και όχι η μορφολογική κατάληξη, όπως αυτή ορίζεται στην γλωσσολογία. Με βάση αυτές τις πληροφορίες προτείνονται κάποιες κατηγορίες, ρήμα, ουσιαστικό, επίθετο, κτλ. Για την αντιμετώπιση προβλημάτων μορφοσυντακτικής ασάφειας, χρησιμοποιείται συνδυαστικά μια λίστα εξαιρέσεων, με τα πιθανά ασαφή λήμματα, και κανόνες βασισμένοι στις καταλήξεις. Στις μονάδες που παραμένουν ασαφείς αποδίδεται η προκαθορισμένη επισημείωση: [ADJ|N|V]. Στη συνέχεια εντοπίζονται ενδεικτικοί του περιεχομένου όροι με βάση την αντιστοιχία που παρουσιάζουν (*pattern matching*) με κάποια προκαθορισμένα μορφοσυντακτικά μοτίβα και ένα λεξικό βασικών εννοιών.
Κατά την ομαδοποίηση των όρων χρησιμοποιούνται παραδοσιακές τεχνικές ανάκτησης πληροφοριών. Γίνεται απλοποίηση των κειμενικών μονάδων καταρχήν με την εξάλειψη λέξεων που θεωρείται ότι δεν έχουν σημασιολογικό περιεχόμενο (*function words*), όπως π.χ, «*of*», «*by*», κ.ά., με βάση λίστα καταγεγραμμένων τέτοιων λέξεων (*stoplist*). Επίσης γίνεται αποκοπή της κατάληξης (*stemming*). Στη συνέχεια εφαρμόζονται τεχνικές ανακατάταξης των όρων με βάση τον πιθανό όρο *κεφαλή* και αλφαβητικής ταξινόμησης, τεχνικές εξίσου συνηθισμένες στην αυτόματη δεικτοδότηση και κατηγοριοποίηση για ανάκτηση κειμένων. Η ομαδοποίηση γίνεται με βάση μετρήσεις συχνότητας εμφάνισης και στατιστικές μετρήσεις κατανομής. Μετράται η συχνότητα εμφάνισης του θέματος των όρων στο κείμενο και στο λεξικό εννοιών σε σχέση με τον αριθμό των διαφορετικών εννοιών στις οποίες εμφανίζεται ο όρος. Αν ο όρος εμφανίζεται σε μεγάλο αριθμό διαφορετικών εννοιών τότε δεν γίνεται ομαδοποίηση, εάν πάλι εμφανίζεται σε μικρό αριθμό εννοιών, τότε αυτές ομαδοποιούνται.
Το FASIT ήταν ένα από τα πρώτα ολοκληρωμένα συστήματα εντοπισμού ενδεικτικών όρων που επιχείρησε να ενσωματώσει επιφανειακές τεχνικές επεξεργασίας γλώσσας για ανάκτηση πληροφοριών και να εκμεταλλευτεί απλά γλωσσικά εργαλεία, όπως stemmers και απλά λεξικά. Ήταν επίσης ένα από τα πρώτα που εισηγήθηκαν αυτόματη και μεγάλης κλίμακας δεικτοδότηση κειμένων.
Σήμερα ένα τέτοιο σύστημα θεωρείται μάλλον «πρωτόγονο», καθώς έχει πλέον αναγνωριστεί η ανάγκη χρήσης γλωσσολογικών πληροφοριών στην επεξεργασία κειμένου. Με την ανάπτυξη αποτελεσματικών μορφοσυντακτικών και μορφολογικών αναλυτών μπορεί να έχουμε πολύ καλύτερα και πιο ακριβή αποτελέσματα από αυτά που μπορεί να προσφέρει ένα απλό λεξικό με επιφανειακές πληροφορίες μορφοσυντακτικών μοτίβων κι ένας stemmer.

Ένα άλλο σύστημα που χρησιμοποιεί τεχνικές αυτόματης δεικτοδότησης για των εντοπισμό όρων είναι το σύστημα CLARIT (Evans et al., 1995). Το CLARIT χρησιμοποιεί επεξεργασία κειμένου για τον εντοπισμό πολυλεκτικών όρων. Μετά από μορφολογική ανάλυση εντοπίζονται οι ονοματικές φράσεις και η επιλογή των πιθανών πολυλεκτικών όρων γίνεται

μέσω δύο στατιστικών μετρήσεων που βασίζονται σε παρατήρηση (heuristics). Μετράται στατιστικά σε ποιο βαθμό τα συνθετικά μιας πολυλεκτικής μονάδας εμφανίζονται συχνά μαζί και συμπεριφέρονται σαν μια λέξη (lexicalisation), και κατά πόσον άλλες λέξεις μπορεί να παρεμβάλλονται μεταξύ τους.

Τέτοιου είδους τεχνικές ανάκτησης πληροφοριών όπως η δεικτοδότηση, φαίνεται φυσικό με μια πρώτη ματιά να έχουν κάποια συνάφεια με την αυτόματη αναγνώριση όρων, καθώς και η αυτόματη δεικτοδότηση και η αυτόματη αναγνώριση όρων εστιάζουν στον εντοπισμό λέξεων. Όμως ενώ υπάρχουν σαφείς ομοιότητες μεταξύ της εξαγωγής όρων και της δεικτοδότησης κειμένων, τα προβλήματα και οι στόχοι είναι διαφορετικοί.
Ο στόχος στην αυτόματη δεικτοδότηση κειμένων όπως είδαμε είναι ο εντοπισμός ενδεικτικών όρων που διακρίνουν ένα κείμενο και συνεπώς διευκολύνουν την ταξινόμηση και ανάκτησή του μέσω ενός συστήματος ανάκτησης πληροφοριών. Από την άλλη πλευρά ο στόχος ενός συστήματος αυτόματης αναγνώρισης όρων είναι ο εντοπισμός των όρων που εκφράζουν έννοιες εξειδικευμένων γλωσσών. Οι ενδεικτικοί όροι δεν είναι απαραίτητα τεχνικοί / επιστημονικοί όροι. Συνήθως αποτελούν ένα μείγμα ορολογίας και λέξεων της γενικής γλώσσας. Επίσης οι περισσότερες μέθοδοι αυτόματης δεικτοδότησης εστιάζονται στον εντοπισμό μεμονωμένων λέξεων κλειδιά, ενώ όπως είδαμε οι όροι στην πλειονότητά τους είναι πολυλεκτικές μονάδες. Εξάλλου οι όροι μπορεί να μην είναι ενδεικτικοί όροι για κάποια κείμενα.

2.3 Γλωσσολογικές Μέθοδοι

Αυτές οι μέθοδοι βασίζονται σε κανόνες και χρησιμοποιούν γλωσσική προ-επεξεργασία και επισημείωση κειμένου που γίνεται από μορφοσυντακτικούς αναλυτές ή από επιφανειακούς συντακτικούς αναλυτές. Χρησιμοποιούν ηλεκτρονικά λεξικά, υπάρχουσες βάσεις γνώσης του τομέα και εντοπίζουν επαναλαμβανόμενα μορφολογικά και συντακτικά μοτίβα σχηματισμού όρων στο συγκεκριμένο γνωστικό αντικείμενο. Για παράδειγμα, μορφολογικά, διακρίνουν συνήθη επιθέματα π.χ. *Ουσιαστικό* + *επίθεμα* (*-ase, -in*) και συνήθη συντακτικά μοτίβα όπως, για παράδειγμα, *Ουσιαστικό* + *Ουσιαστικό,* (*Επίθετο* | *Ουσιαστικό*) + *Ουσιαστικό,* κτλ.
Πρώτο στάδιο της γλωσσικής προ-επεξεργασίας αποτελεί ο εντοπισμός στο κείμενο των *ελαχίστων κειμενικών μονάδων* ή *tokens*. Η διαδικασία αυτή ονομάζεται *tokenisation*. Οι *tokenisers,* εκμεταλλεύονται επιφανειακά κειμενικά χαρακτηριστικά όπως σημεία στίξης, διαστήματα μεταξύ των λέξεων και την ύπαρξη κεφαλαίων ή μικρών γραμμάτων, για το χωρισμό του κειμένου σε λέξεις και προτάσεις. Σε αυτό το στάδιο άλλες γλωσσικές πληροφορίες είναι ελάχιστες. Ο διαχωρισμός λέξεων και προτάσεων κατά αυτό τον τρόπο είναι ένα δύσκολο έργο που μπορεί και να δημιουργήσει προβλήματα αν σκεφτεί κανείς το ρόλο κάποιων σημείων στίξης και των κεφαλαίων και μικρών γραμμάτων στην δημιουργία όρων. Για παράδειγμα, η φράση «*text-based medium*» μπορεί κατά τον tokeniser να αποτελείται από δύο ελάχιστες μονάδες / λέξεις. Όμως γλωσσολογικά πρέπει να θεωρηθεί ως δύο λέξεις, τρεις ή μία; Στις περιπτώσεις σύνθετων, πολυλεκτικών όρων και των παραλλαγών τους, ουσιαστικά όπως «*data bases*» μπορεί να εμφανίζονται στο κείμενο ως «*data bases*», «*databases*» , «*data-bases*» , «*DBs*» ή «*D.B.*».
Κατά την *tokenisation* αναγνωρίζονται και κάποιες χρήσιμες επιφανειακές πληροφορίες γραφής, τις οποίες μπορούμε να κρατήσουμε με τη μορφή επισημειώσεων (*tags*), για χρήση σε επόμενο στάδιο επεξεργασίας , π.χ. NF [orth=upper], Cys [orth=mixed] – όπου οι αγκύλες δηλώνουν την επισημείωση, γραφή με κεφαλαία ή μεικτά αντίστοιχα.
Η μορφολογική επεξεργασία είναι ένα άλλο στάδιο γλωσσικής προ-επεξεργασίας. Σε αυτό το στάδιο μπορούμε είτε να εφαρμόσουμε απλό διαχωρισμό θέματος και κατάληξης (stemming), είτε πιο σύνθετη μορφολογική ανάλυση. Οι stemmers διαχωρίζουν το θέμα από την κατάληξη με απλό, μηχανικό τρόπο, με την περικοπή, για παράδειγμα, των τελευταίων 2-3 χαρακτήρων μιας λέξης: *studies* → *stud*. Η επεξεργασία τέτοιου τύπου συνήθως δεν περιλαμβάνει ιδιαίτερες γλωσσικές πληροφορίες, πέρα από κάποιους πολύ γενικούς κανόνες για τη μορφή της κατάληξης και την αναμενόμενη έκταση της σειράς στοιχείων / χαρακτήρων που πρέπει να περικοπούν. Μπορεί λοιπόν κατά το stemming να υπάρξουν σφάλματα και απώλεια πληροφορίας είτε λόγω περικοπής χαρακτήρων που δεν ανήκουν

στην κατάληξη αλλά στο θέμα (*overstemming*), είτε λόγω περικοπής λιγότερων χαρακτήρων από αυτούς που ανήκουν στην κατάληξη (*understemming*). Όμως, από την άλλη πλευρά, το stemming είναι μια επεξεργασία αρκετά απλή και γρήγορη, με σχετικά ικανοποιητικά αποτελέσματα και για το λόγο αυτό χρησιμοποιείται ευρέως σε εφαρμογές ανάκτησης πληροφοριών και αυτόματης αναγνώρισης όρων. Ένας από τους πιο γνωστούς αλγορίθμους για stemming είναι ο αλγόριθμος Porter (Porter, 1980).
Πιο σύνθετοι μορφολογικοί αναλυτές (*morphological analysers*), βασίζονται σε γλωσσικούς κανόνες και λεξικά. Σε αυτή την περίπτωση, εντοπίζονται τα προσφύματα και τα γλωσσικά χαρακτηριστικά κάθε προσφύματος και παράγονται δενδρικές δομές λέξεων που αναλύουν τον τρόπο με τον οποίο τα μορφήματα συνδυάζονται για τη δημιουργία λέξεων.
Η μορφολογική επεξεργασία είναι ιδιαίτερα σημαντική στην αναγνώριση όρων γιατί αποδίδει την κανονική, απλοποιημένη μορφή (*normalised form*) ενός όρου, περιορίζοντας προβλήματα μορφολογικής ποικιλότητας. Για αυτό το λόγο, όταν χρησιμοποιείται μορφολογική επεξεργασία σε συνδυασμό με στατιστικές μεθόδους, τα αποτελέσματα των μετρήσεων μπορούν να βελτιωθούν σημαντικά με τη χρήση του θέματος ενός όρου στις μετρήσεις αντί των επιφανειακών μορφολογικών παραλλαγών του.
Η κυρίως γλωσσική επεξεργασία συνίσταται στη χρήση μορφοσυντακτικών αναλυτών (*part of speech/POS taggers*) και κάποιες φορές και επιφανειακών συντακτικών αναλυτών (*shallow parsers*). Οι μορφοσυντακτικοί αναλυτές μπορεί να είναι διαφόρων τύπων. Μπορεί να βασίζονται σε γλωσσικούς κανόνες, ή σε στατιστικές μεθόδους για την ανάλυση. Τα δεδομένα εξόδου (*output data*) ενός μορφοσυντακτικού αναλυτή είναι κείμενα, όπου οι λέξεις συνοδεύονται από τη *μορφοσυντακτική τους επισημείωση* (*POS tag*). Για παράδειγμα, η πρόταση *«Polymorphs and nuclear debris are present in the keratinous cap.»* μετά από μορφοσυντακτική ανάλυση παίρνει τη μορφή: *polymorphs/NNP and/CC nuclear/JJ debris/NN are/VBP present/JJ in/IN the/DT keratinous/NNP cap/NN./.* όπου η κάθετος δηλώνει μορφοσυντακτική επισημείωση (Ουσ.-NNP, Επίθ.-JJ, κτλ.). Στο παράδειγμα παρατηρούμε ότι το επίθετο *keratinous* δεν έχει αναγνωριστεί σωστά. Πρέπει να λάβουμε υπόψη μας ότι, παρά την εξέλιξη των συστημάτων μορφοσυντακτικής ανάλυσης, υπάρχει πάντα ένα ενδεχόμενο ποσοστό λάθους, που μπορεί να δημιουργήσει προβλήματα στα επόμενα στάδια επεξεργασίας.

Μια από τις πρώτες καθαρά γλωσσολογικές μεθόδους αναγνώρισης όρων είναι αυτή της Ανανιάδου (Ananiadou, 1988, 1994) που αναγνωρίζει μονολεκτικούς όρους από το χώρο της Ανοσολογίας. Η μέθοδος βασίζεται θεωρητικά σε έρευνες λεξικολογικής μορφολογίας για την Αγγλική γλώσσα, που επεκτείνονται και εμπλουτίζονται με στοιχεία για τον εντοπισμό και την κατηγοριοποίηση συνθέτων *νεοκλασικού* τύπου, δηλαδή συνθέτων με θέματα ή καταλήξεις προερχόμενες από την Ελληνική και την Λατινική γλώσσα, ή υβριδικού τύπου, όπως για παράδειγμα, «*alphafetoprotein*», «*immunoosmoelectrophoresis*», «*radioimmuno-assay*». Διακρίνει συνήθη μοτίβα σχηματισμού σύνθετων όρων, όπως για παράδειγμα όρους που αποτελούνται από λέξη της γενικής γλώσσας με ορολογικό επίθημα, ή από όρο με επίθημα της γενικής γλώσσας. Πέρα από εντοπισμό προθημάτων, όπως «*auto-*», «*hypo-*», και επιθημάτων, όπως «*-osis*»,«*-itis*», κ.ά., που χρησιμοποιούνται στο σχηματισμό όρων, εντοπίζεται και διακρίνεται και μια άλλη κατηγορία μορφολογικών μονάδων, αυτών που εμφανίζονται μέσα σε σύνθετους όρους ως *συνθετικό* και μπορεί να συνδυαστούν με άλλα συνθετικά, όρους ή προσφύματα, π.χ. «*electro*», «*immuno*», κ.ά. Τα μοτίβα αυτά σχηματισμού όρων χρησιμοποιούνται για τη μορφολογική ανάλυση και στη συνέχεια την αναγνώριση όρων. Έτσι για την αναγνώριση του όρου «*periimunoglobulinaemia*» μπορεί να χρησιμοποιηθούν πληροφορίες όπως:

peri [cat=prefix, type=term, prefixes=N/V/AdJ, makes=N/V/AdJ],
immuno [cat=comb, type= term],
globulin [cat=comb, type=term],
aemia [cat=suffix, type=term, suffixes=comb, makes=N]

όπου οι αγκύλες δηλώνουν μορφολογική επισημείωση. Ως κατηγορία *comb*, κατατάσσονται μορφολογικές μονάδες που εμφανίζονται ως *συνθετικά όρων*, τα πεδία *prefixes/suffixes,* για προθήματα και επιθήματα αντίστοιχα, περιέχουν πληροφορίες σχετικά με τη μορφοσυντακτική κατηγορία του θέματος και το πεδίο *makes* περιέχει πληροφορίες σχετικά με τη μορφοσυντακτική κατηγορία του σύνθετου όρου, (Ουσ., Ρ., ή Επίθ.), που το πρόσφυμα δύναται να δημιουργήσει.

Έχει παρατηρηθεί ότι οι περισσότεροι πολυλεκτικοί όροι εμφανίζουν κάποια συγκεκριμένα χαρακτηριστικά σε σχέση με άλλες ονοματικές φράσεις. Για παράδειγμα, δεν επιτρέπουν την παρεμβολή άρθρων ή επιρρημάτων μέσα στην ονοματική φράση. Έτσι, για την αναγνώριση πολυλεκτικών όρων ανεξαρτήτως γνωστικού αντικειμένου οι περισσότερες γλωσσολογικές μέθοδοι χρησιμοποιούν *γλωσσικά φίλτρα* (*linguistic filters*), δηλαδή κανόνες που επιτρέπουν μόνο σε συγκεκριμένα μορφοσυντακτικά μοτίβα να αναγνωρίζονται ως πιθανοί όροι. Τα πιο συνηθισμένα γλωσσικά φίλτρα τέτοιου τύπου είναι αυτά που επιτρέπουν την αναγνώριση διαφόρων τύπων ονοματικών φράσεων που θα μπορούσαν να προταθούν ως πιθανοί όροι. Ένα γλωσσικό φίλτρο, για παράδειγμα, θα μπορούσε να έχει τη μορφή: (*Adj* | *Noun*)$^{+}$ *Noun*, δηλαδή να επιτρέπει την αναγνώριση ονοματικών φράσεων που αποτελούνται από ένα τουλάχιστον ή περισσότερα, επίθετα ή ουσιαστικά ακολουθούμενα από ένα ουσιαστικό.
Ένα σύστημα που βασίζεται σε τέτοιου είδους μορφοσυντακτικές πληροφορίες για την αναγνώριση πολυλεκτικών όρων στη Γαλλική γλώσσα είναι το σύστημα LEXTER (Bourigault et al., 1996). Χρησιμοποιώντας επιφανειακή μορφοσυντακτική ανάλυση και γλωσσικούς κανόνες βασισμένους σε παρατήρηση (*heuristics*) των κειμένων, το LEXTER επιχειρεί να εντοπίσει και να εξάγει τις μέγιστες δυνατόν ονοματικές φράσεις. Στη συνέχεια, αυτές οι φράσεις διαχωρίζονται σε μικρότερες φράσεις και προτείνονται ως πιθανοί όροι αν εμφανίζονται ως ονοματικές φράσεις και σε άλλα σημεία του κειμένου. Σε διαφορετική περίπτωση υποβάλλονται σε νέα διχοτόμηση.
Οι γραμματικοί κανόνες του LEXTER χρησιμοποιούν μορφοσυντακτικές και μορφολογικές πληροφορίες για τον εντοπισμό αυτών των μέγιστων δυνατόν ονοματικών φράσεων. Για τη διχοτόμησή τους χρησιμοποιείται τοπικά συντακτικός αναλυτής. Το LEXTER είναι ένα καθαρά γλωσσολογικό σύστημα, δεν χρησιμοποιεί στατιστικές μετρήσεις όπως συχνότητα εμφάνισης ή περιορισμούς *οριακής τιμής* (*threshold*). Κατά αυτό τον τρόπο, ακόμη και ονοματικές φράσεις που εμφανίζονται μία φορά στο κειμενικό σώμα (*άπαξ λεγόμενα*) μπορούν να εντοπιστούν ως πιθανοί όροι. Μειονέκτημα όμως μιας τέτοιας στρατηγικής είναι η αύξηση του *θορύβου* στα αποτελέσματα, δηλαδή η εξαγωγή, στην προκειμένη περίπτωση, ονοματικών φράσεων που δεν είναι όροι. Επίσης, μέθοδοι, όπως το LEXTER, που βασίζονται σε μεγάλο βαθμό σε γλωσσικούς κανόνες είναι εξαιρετικά ειδικευμένες σε μια γλώσσα και συχνά είναι δύσκολη ή και αδύνατη η προσαρμογή και εφαρμογή τους σε άλλες γλώσσες.

Πρόσφατες εξελίξεις των εφαρμογών εξαγωγής πληροφοριών έδωσαν το έναυσμα σε νέες γλωσσολογικές προσεγγίσεις της αυτόματης αναγνώρισης όρων, βασισμένες σε μεθοδολογίες και υπάρχοντα συστήματα εξαγωγής πληροφοριών. Δύο τέτοια συστήματα είναι και τα συστήματα EMPathIE και PASTA για την αναγνώριση όρων Μοριακής Βιολογίας (Gaizauskas et al., 2000). Τα συστήματα αυτά βασίζονται στο σύστημα εξαγωγής πληροφοριών LaSIE (Gaizauskas et al. 1995). Μπορούν να επεξεργαστούν κείμενα δομικά αναλυμένα (*structured texts*) με επισημειώσεις SGML, αλλά και απλό, καθαρό κείμενο. Στην δεύτερη περίπτωση, πριν από την *tokenisation* εφαρμόζεται *sectioniser,* δηλαδή ένα σύστημα διαχωρισμού του κειμένου σε ζώνες. Ο *tokeniser* είναι προσαρμοσμένος στις ιδιαίτερες ανάγκες ενός συστήματος αναγνώρισης όρων, ούτως ώστε να λαμβάνει υπόψη του ιδιαιτερότητες, όπως ο ρόλος των σημείων στίξης και των μονάδων που αποτελούνται από γράμματα και αριθμούς, ή περιέχουν παύλες κτλ. Για τη μορφολογική ανάλυση χρησιμοποιείται μορφολογικός αναλυτής εμπλουτισμένος με πληροφορίες για ορολογικά προσφύματα. Στη συνέχεια, η αναγνώριση και κατηγοριοποίηση των όρων βασίζεται σε ένα συνδυασμό πληροφοριών ορολογικών λεξικών και γλωσσικών κανόνων. Σε τελικό στάδιο εφαρμόζονται οι κανόνες για αναγνώριση παραλλαγών όρων.

Για τη δημιουργία των ορολογικών λεξικών χρησιμοποιούνται ορολογικές βάσεις του τομέα Βιοχημείας όπως οι SWISS-PROT, CATH και SCOP, καθώς επίσης και επισημειωμένα κείμενα. Ένα μέρος των γλωσσικών κανόνων για την αναγνώριση δημιουργήθηκε με ημιαυτόματο τρόπο, βασισμένο στις πληροφορίες των λεξικών. Οι κανόνες για τον εντοπισμό ορολογικής ποικιλίας εστιάζουν στην αναγνώριση κυρίως ακρωνύμων και συντετμημένων όρων. Αυτοί οι κανόνες βασίζονται σε παρατήρηση κειμένου (*heuristics*) και επιχειρούν να ταυτίσουν τις σχηματομορφές (*pattern matching*) αναγνωρισμένων όρων και μη αναγνωρισμένων, συντετμημένων όρων ή ακρωνύμων.
Μέθοδοι τέτοιου είδους, που βασίζονται σε υπάρχοντα συστήματα εξαγωγής πληροφοριών έχουν ως άμεσο πλεονέκτημα την χρήση μιας πλήρους αναπτυγμένης πλατφόρμας γλωσσικών εργαλείων, όπως μορφολογικούς και μορφοσυντακτικούς αναλυτές, βάσεις δεδομένων και υπάρχοντα λεξικά. Μπορούν λοιπόν να αναπτυχθούν και να προσαρμοστούν σχετικά γρήγορα και εύκολα, ούτως ώστε όχι μόνο να αναγνωρίζουν όρους αλλά και να εξάγουν πληροφορίες από επιστημονικά ή τεχνικά κείμενα. Επιπλέον, μπορούν να χρησιμοποιήσουν για τον εντοπισμό όρων, δοκιμασμένες τεχνικές και έρευνα πάνω στην αναγνώριση ονοματικών οντοτήτων (*named entities*), ένα από τα σχετικά εύκολα κι αποτελεσματικά *επιμέρους έργα* (*tasks*) της εξαγωγής πληροφοριών. Όμως τα προβλήματα και οι στόχοι στην εξαγωγή πληροφοριών διαφέρουν από αυτά της αναγνώρισης όρων. Τα μορφοσυντακτικά μοτίβα πολυλεκτικών και ποικίλων όρων είναι πολύ πιο σύνθετα από αυτά που υποδηλώνουν ονοματικές οντότητες. Επίσης η μορφολογική ανάλυση, όπως είδαμε, παίζει ένα καθοριστικό ρόλο στην αναγνώριση όρων, ενώ κάτι τέτοιο δεν είναι εξίσου απαραίτητο στην αναγνώριση ονοματικών οντοτήτων. Κατά συνέπεια ο βαθμός δυσκολίας για αναγνώριση πολυλεκτικών και παραλλαγών όρων με γλωσσικούς κανόνες είναι κατά πολύ μεγαλύτερος. Για αυτό το λόγο, ένα τέτοιο σύστημα βασίζεται κυρίως στη χρήση ποικίλων, ήδη διαθέσιμων ορολογικών πηγών και λιγότερο σε γλωσσικούς μορφοσυντακτικούς κανόνες.

2.4 Υβριδικές Μέθοδοι

Οι υβριδικές μέθοδοι προσέγγισης του προβλήματος της αυτόματης αναγνώρισης όρων αποτελούν συνδυασμούς των μεθόδων που προαναφέρθηκαν. Συνδυάζουν κυρίως γλωσσικές πληροφορίες, όπως γλωσσικά φίλτρα και επιφανειακούς συντακτικούς αναλυτές με στατιστικές μετρήσεις.

Το σύστημα TERMS (Justeson & Katz, 1995) έχει ως στόχο την αναγνώριση πολυλεκτικών όρων σε κείμενα οποιουδήποτε αντικειμένου και τύπου. Έρευνες δείχνουν ότι το 92-99% των όρων είναι πολυλεκτικές ονοματικές φράσεις. Έτσι βασικές θεωρητικές υποθέσεις στην προσέγγιση του TERMS είναι ότι καταρχήν οι ονοματικές φράσεις - όροι επαναλαμβάνονται σε κείμενα τεχνικής φύσεως συχνότερα από άλλες ονοματικές φράσεις, και ότι οι όροι έχουν ιδιαιτερότητες ως προς τη δομή που τους διαφοροποιούν από άλλες ονοματικές φράσεις. Δηλαδή για παράδειγμα δεν επιδέχονται παρεμβολή επιρρημάτων ή άρθρων ενώ οι απλές ονοματικές εκφράσεις διαθέτουν πιο ευέλικτες δομές κι επιδέχονται ένα ευρύ φάσμα συντακτικών τροποποιήσεων.
Σύμφωνα με αυτές τις αρχές, οι ονοματικές φράσεις χαμηλής συχνότητας εμφάνισης απορρίπτονται μέσω στατιστικού φίλτρου που ορίζει κατώτατες οριακές τιμές (*threshold*) συχνότητας, προκειμένου να ληφθεί υπόψη κάποια πολυλεκτική δομή. Στη συνέχεια εφαρμόζεται ένα γλωσσικό μορφοσυντακτικό φίλτρο για την επιλογή των πολυλεκτικών όρων:
((Adj|Noun)+ | ((Adj|Noun)*(Noun- Prep)?) (Adj|Noun)*) Noun
Για την μορφοσυντακτική ανάλυση το TERMS βασίζεται σε πληροφορίες μορφολογικής λεξικολογικής βάσης δεδομένων και σε επισημειωμένο κειμενικό σώμα κι όχι σε μορφοσυντακτικό αναλυτή, γιατί υποστηρίζεται ότι έτσι περιορίζεται το ποσοστό σφάλματος. Για την αντιμετώπιση της ασάφειας χρησιμοποιείται προτιμησιακός αλγόριθμος (*preference algorithm*) . Το σύστημα TERMS κατά την αξιολόγηση αναφέρεται ότι επιτυγχάνει ποσοστά ανάκλησης (*recall*) της τάξης του 71% και ακρίβειας (*precision*) 67-

96% συγκριτικά με δεδομένα αναγνώρισης όρων στο ίδιο κειμενικό σώμα από ειδικούς επιστήμονες.
Παρά το γεγονός ότι το TERMS επιτυγχάνει ικανοποιητική απόδοση, το σύστημα δεν φέρεται να επιλύει ικανοποιητικά το πρόβλημα των παραλλαγών όρων (Jaquemin, 2001). Επίσης δεν γίνεται σαφής ο τρόπος με τον οποίο αναλύονται μορφοσυντακτικά λέξεις που δεν υπάρχουν στο λεξικό και θεωρείται ότι η χρήση μορφοσυντακτικού αναλυτή θα μπορούσε να βελτιώσει την ανάλυση, παρά το ενδεχόμενο ποσοστό σφάλματος (ibid.).

Ένα άλλο υβριδικό σύστημα που συνδυάζει στατιστικές και γλωσσολογικές μεθόδους είναι το ACABIT (Daille et al. 1994). Το ACABIT είναι ένας υβριδικός αναλυτής που αποτελείται από επιφανειακό μορφοσυντακτικό αναλυτή και στατιστικά φίλτρα για την αναγνώριση όρων. Μετά από εκτενείς μελέτες της μορφοσυντακτικής δομής των όρων στη Γαλλική γλώσσα οι (Daille et al. 1994) καταλήγουν σε σχετικά απλά και «ανοικτά» γλωσσικά φίλτρα που εστιάζουν κυρίως στον εντοπισμό δι-λεκτικών όρων, θεωρώντας ότι οι πολυλεκτικοί όροι αποτελούν συνθέσεις αυτών των δι-λεκτικών ένθετων όρων. Τα απλά γλωσσικά φίλτρα που χρησιμοποιεί αντισταθμίζονται με χρήση στατιστικών φίλτρων για τα επιτρεπόμενα μορφοσυντακτικά μοτίβα, αποφεύγοντας έτσι την υπερβολική ανάκληση προτεινόμενων όρων και παραλλαγών τους.
Χρησιμοποιεί στατιστικές μετρήσεις, όπως συντελεστή Φ^2, συντελεστή λογαριθμικού τύπου και συχνότητας εμφάνισης. Θεωρείται ότι, όροι, αποτελούμενοι από δύο λέξεις, τείνουν να συνεμφανίζονται συχνότερα από ότι άλλοι που τυγχάνουν να εμφανίζονται μαζί. Επίσης μετρήσεις ποικιλότητας επιχειρούν να προσδιορίσουν πόσο συχνά ένας προσδιορισμός και μια *κεφαλή* όρου συνδυάζονται με άλλες λέξεις κατά τον ίδιο τρόπο. *Κεφαλές όρων* με μεγάλη ποικιλότητα συνδυασμών θεωρείται ότι δηλώνουν σημαντικές, βασικές έννοιες στο συγκεκριμένο επιστημονικό / τεχνικό αντικείμενο, ενώ το αντίστοιχο φαινόμενο στους προσδιορισμούς όρων υποδηλώνει κοινές έννοιες, που δεν έχουν ορολογικό ενδιαφέρον.
Τα δεδομένα εισόδου στο ACABIT είναι κείμενα, μορφοσυντακτικά αναλυμένα και το αποτέλεσμα της επεξεργασίας εμφανίζεται με μορφή λίστας δι-λεκτικών προτεινόμενων όρων. Το ACABIT λαμβάνει υπόψη κάποια είδη ορολογικής ποικιλότητας, κυρίως αυτά που προέρχονται από παρατακτική σύνδεση, παρεμβολή προσδιορισμών και υπερ-σύνθεση (overcomposition) καθώς επίσης και κάποια είδη μορφολογικής ποικιλότητας που απλοποιεί στη βασική τους μορφή.
Σημαντικό πλεονέκτημα της μεθόδου αυτής είναι ότι η γλωσσικές πηγές και εργαλεία που χρησιμοποιεί δεν είναι τέτοια που να περιορίζουν τη χρήση της σε μια συγκεκριμένη γλώσσα. Από την άλλη πλευρά, όπως σε όλες τις στατιστικές μετρήσεις, η ακριβής ρύθμιση των παραμέτρων των στατιστικών φίλτρων είναι δύσκολο να επιτευχθεί. Κι αυτό γιατί, αντίθετα με τα γλωσσικά φίλτρα που απομακρύνουν μοτίβα που δεν ανταποκρίνονται στους κανόνες, τα στατιστικά φίλτρα απλώς πριμοδοτούν ή όχι τις τιμές που ορίζουν τη σειρά με την οποία οι προτεινόμενοι όροι εμφανίζονται στην τελική λίστα.

Η *C/NC value* μέθοδος (Frantzi & Ananiadou, 1999) είναι μια μέθοδος που μπορεί να εφαρμοστεί σε οποιοδήποτε γνωστικό αντικείμενο και είδος κειμένου, για αναγνώριση πολυλεκτικών και ένθετων όρων. Συνδυάζει στατιστικές μετρήσεις με γλωσσικές πληροφορίες. Μετά από μορφοσυντακτική προ-επεξεργασία εφαρμόζονται γλωσσικά φίλτρα με στόχο τον εντοπισμό κυρίως ονοματικών φράσεων. Η μέθοδος χρησιμοποιεί τις στατιστικές μετρήσεις των τιμών *C* και *NC*, με στόχο το συνδυασμό τόσο εγγενών γλωσσικών πληροφοριών για τον όρο, όπως τα μορφοσυντακτικά μοτίβα σχηματισμού του, όσο και εξωγενών πληροφοριών, δηλαδή αυτών που προέρχονται από το γλωσσικό του περιβάλλον. Τα δεδομένα εξόδου του συστήματος έχουν τη μορφή λίστας προτεινόμενων όρων, ταξινομημένων με βάση την C/NC τιμή τους, χωρίς καμία ανθρώπινη παρέμβαση.
Η μέθοδος φαίνεται να αποδίδει αρκετά καλά στην αναγνώριση πολυλεκτικών και ένθετων όρων και η ενσωμάτωση πληροφοριών από το γλωσσικό περιβάλλον (*NC*) δείχνει να βελτιώνει σημαντικά την ακρίβεια των αποτελεσμάτων.

Επέκταση αυτής της μεθόδου αποτελεί η μέθοδος TRUCKS (*Term Recognition Using Combined Knowledge Sources*) (Maynard & Ananiadou, 2000b), που εστιάζει στον εντοπισμό πληροφοριών για το γλωσσικό περιβάλλον ενός όρου μέσω διαφόρων πηγών, με στόχο την περαιτέρω βελτιστοποίηση των στατιστικών μετρήσεων για αναγνώριση όρων. Βασισμένη στην έννοια της *ορολογικής οικειότητας* (*terminological acquaintance*) προσπαθεί να μελετήσει τους όρους σε σχέση με συγκειμενικές πληροφορίες (*contextual information*) και όχι ως μεμονωμένες φράσεις. Διακρίνει τρία είδη συγκειμενικής πληροφορίας: *μορφοσυντακτική, σημασιολογική* και *ορολογική*. Παρατηρείται ότι κάποιες μορφοσυντακτικές κατηγορίες λέξεων μπορούν να οδηγήσουν σε προβλέψεις πιθανής εμφάνισης όρου μέσα στο άμεσο γλωσσικό τους περιβάλλον. Επίσης χρησιμοποιεί σημασιολογικές πληροφορίες μέσω *μετρήσεων ομοιότητας* (*similarity measures*) και χρήση ορολογικών θησαυρών, όπως ο θησαυρός ιατρικής ορολογίας UMLS. Τέλος, χρησιμοποιεί ορολογικές πληροφορίες, θεωρώντας ότι συγκείμενες λέξεις που είναι επίσης όροι είναι σημαντικοί στον εντοπισμό άλλων όρων. Ο συσχετισμός και η σημασία όλων αυτών των παραμέτρων κωδικοποιείται μέσω συντελεστών βαρύτητας για τη μέτρηση του συνολικού *συντελεστή σπουδαιότητας* (*Importance Weight*). Ο συντελεστής σπουδαιότητας υπολογίζεται με πολλαπλασιασμό του ορολογικού συντελεστή βαρύτητας επί της τιμής του σημασιολογικού συντελεστή και στη συνέχεια πρόσθεση του γινόμενου στο συνολικό μορφοσυντακτικό συντελεστή όλων των συγκειμένων λέξεων ή όρων, του όρου προς αναγνώριση (Maynard & Ananiadou, 2000a).

Στο σύνολό της, η μέθοδος βασίζεται σε τρία επίπεδα, ένα βασικό στατιστικό επίπεδο που καθορίζεται από τον υπολογισμό της τιμής *C*, ένα μεσαίο που καθορίζει το συντελεστή βαρύτητας συγκειμένου μέσω της τιμής *NC* και τέλος, ένα ανώτερο επίπεδο που ορίζει το συντελεστή σπουδαιότητας. Τα τρία αυτά επίπεδα συνδυάζονται για τον υπολογισμό της τελικής τιμής *SNC*, δηλαδή το συνδυασμό της τιμής *NC* και του συντελεστή σπουδαιότητας.

Σημαντική καινοτομία της μεθόδου TRUCKS αποτελεί ο τρόπος με τον οποίο αποκτάται σημασιολογική πληροφορία για το γλωσσικό περιβάλλον, δηλαδή η χρήση μετρήσεων ομοιότητας των συγκειμένων όρων με τους πιθανούς όρους, μέσω ορολογικού θησαυρού. Θεωρείται ότι ένας συγκείμενος όρος που προσεγγίζει σημασιολογικά έναν υποψήφιο όρο, είναι πιο πιθανό να παίζει ρόλο στον εντοπισμό του από ότι ένας συγκείμενος όρος που εμφανίζει χαμηλότερες τιμές σημασιολογικής ομοιότητας. Επίσης έρευνα πάνω στο ορολογικό γλωσσικό περιβάλλον χρησιμοποιήθηκε για την αποσαφήνιση όρων και τη βελτίωση της αναγνώρισης, βασισμένη σε *ομαδοποίηση* (*clustering*) του γλωσσικού τους περιβάλλοντος.

Το γραφικό περιβάλλον εργασίας ATRACT βασίζεται στη μέθοδο *C/NC* για εξόρυξη κειμένου και γνώσης από το Διαδίκτυο, συνδυάζοντας εφαρμογές όπως η αναγνώριση όρων, η αυτόματη ομαδοποίηση όρων με βάση το γλωσσικό τους περιβάλλον, η ανάκτηση κειμένων, η εξαγωγή πληροφοριών και η ευφυής πρόσβαση σε διάφορες ετερογενείς βάσεις δεδομένων από το χώρο της βιολογίας και της γενετικής (Mima et al. 2001a, 2001b, Ananiadou et al. 2001).

2.5 Μέθοδοι Μηχανικής Μάθησης

Τα συστήματα που θα παρουσιάσουμε χρησιμοποιούν εποπτευόμενες μεθόδους μάθησης για αναγνώριση όρων στον τομέα βιοτεχνολογίας. Χρησιμοποιούν αρκετά δημοφιλείς στον χώρο της επεξεργασίας γλώσσας τεχνικές μηχανικής μάθησης, αυτές που βασίζονται σε δένδρα αποφάσεων και σε Μαρκοβιανά μοντέλα.. Ανήκουν στην κατηγορία των εποπτευόμενων (*supervised*) μεθόδων γιατί απαιτούν δεδομένα εκμάθησης (*training data*). Η εκμάθηση συνίσταται στον εντοπισμό κανόνων με βάση τα δεδομένα εκμάθησης και ένα σύνολο διακριτικών χαρακτηριστικών τους, προκειμένου να προβλέπεται στη συνέχεια επιτυχώς η κατηγορία νέων, αγνώστων δεδομένων. Η απόδοση αυτών των συστημάτων εξαρτάται, τόσο από τις ιδιαιτερότητες του αλγόριθμου που εφαρμόζεται κάθε φορά, όσο και από τον όγκο των δεδομένων εκμάθησης και την σωστή επιλογή διακριτικών χαρακτηριστικών για την κατηγοριοποίηση. Λανθασμένα αποτελέσματα μπορεί να οφείλονται σε έλλειψη επαρκών δεδομένων εκπαίδευσης (*sparse data*) αλλά και σε χρήση χαρακτηριστικών που δεν

επηρεάζουν το φαινόμενο που προσπαθούμε να προβλέψουμε, και που η εμφάνισή τους είναι μάλλον συμπτωματική. Οι μέθοδοι μηχανικής μάθησης για αναγνώριση όρων χρησιμοποιούν μορφοσυντακτικές και γραφηματικές πληροφορίες μεταξύ άλλων χαρακτηριστικών.

Μια μέθοδος εποπτευόμενης εκμάθησης βασισμένη σε Κρυφά Μαρκοβιανά Μοντέλα (*Hidden Markov Models*) έχει χρησιμοποιηθεί από τους Collier et al. (2000) για τον εντοπισμό όρων μοριακής βιολογίας σε κείμενα από το MEDLINE. Το σύστημα εκπαιδεύτηκε στον εντοπισμό των ορολογικών ζευγών (*bigrams*) που εμφανίζονται σε μικρό κειμενικό σώμα, με χαρακτηριστικά εκμάθησης (*features*) που περιελάμβαναν λεξικολογική και γραφηματική πληροφορία. Θεωρήθηκε ότι τα γραφηματικά χαρακτηριστικά των λέξεων μπορούν να βοηθήσουν στην κατηγοριοποίηση των όρων. Αυτά τα χαρακτηριστικά μπορούν στη συνέχεια, να συμβάλουν στη δημιουργία υποδειγμάτων (*models*) για την αναζήτηση ομοιοτήτων μεταξύ γνωστών όρων, δηλαδή λέξεων που έχουν αναγνωριστεί στο σύνολο εκπαίδευσης (*training set*), και αγνώστων όρων, δηλαδή λέξεων με μηδενική συχνότητα εμφάνισης στο σύνολο εκπαίδευσης. Παραδείγματα γραφηματικών χαρακτηριστικών που επιλέχθηκαν στην εκμάθηση είναι:

- Δυο Κεφαλαία: RalGDS
- Συνδυασμός μικρών – κεφαλαίων: kappaB
- Αρχικό κεφαλαίο: Interleukin
- Ελληνικό γράμμα: alpha

Τα ζεύγη όρων που εντοπίζονται, επισημειώνονται από ειδικούς του γνωστικού αντικειμένου με πληροφορίες σχετικά με την ορολογική κατηγορία στην οποία ανήκουν, για παράδειγμα, *protein*, *DNA*, κτλ.

Ο στόχος της διαδικασίας εκμάθησης είναι ο εντοπισμός της πιο πιθανής αλληλουχίας ορολογικών κατηγοριών σε μια δεδομένη αλληλουχία λέξεων. Δηλαδή αν θεωρήσει κανείς ως P την πιθανότητα εμφάνισης μιας αλληλουχίας κατηγοριών, και C αυτή την αλληλουχία δεδομένης μιας W αλληλουχίας λέξεων, ο στόχος του συστήματος εκφράζεται ως η μεγιστοποίηση της συνάρτησης $P(C|W)$.

Class	***No of XML tagged terms***	***Example***	***Description***
PROTEIN	2125	JAK kinase	Proteins, protein groups, complexes & substructures
DNA	358	IL-2 promoter	DNAs, DNA groups, regions and genes

Πίνακας 1: Παράδειγμα κατηγοριών ζευγών όρων και των χαρακτηριστικών τους.

Ο αλγόριθμος εκμάθησης που χρησιμοποιείται σχεδιάστηκε ούτως ώστε να περιορίζει προβλήματα που οφείλονται σε μη επαρκή δεδομένα εκμάθησης (*sparse data*). Το σύστημα υποδεικνύει μια *κατάσταση* (*state*) ανά κατηγορία όρου και δύο ειδικές κατηγορίες για την αρχή και το τέλος μιας πρότασης αντίστοιχα. Για τον εντοπισμό πιθανών κατηγοριοποιήσεων οι Colier, et al. (2000) χρησιμοποιούν τον αλγόριθμο Viterbi (Viterbi, 1967). Στο τελικό στάδιο επεξεργασίας και αφού ο εντοπισμός και η επισημείωση των κατηγοριών έχει ολοκληρωθεί, δημιουργείται μια λίστα συχνότητας εμφάνισης όρων και κατηγοριών για ένα κείμενο, βασισμένη στις πιο συχνές κατηγορίες που εντόπισε ο αλγόριθμος.

No of Texts	80	40	20	10	5
F score	0.728	0.705	0.647	0.594	0.534

Πίνακας 2: Ενδεικτικά αποτελέσματα αξιολόγησης
όπου F score ο συνδυασμός ακρίβειας (Precision) και ανάκλησης (Recall)

Σε γενικές γραμμές, το σύστημα φέρεται να αποδίδει ικανοποιητικά και χωρίς την ανάγκη δημιουργίας μεγάλου συνόλου εκμάθησης. Ένα άλλο πλεονέκτημα έγκειται στην επιλογή

των χαρακτηριστικών εκμάθησης. Τα γραφηματικά χαρακτηριστικά που επιλέγονται είναι όχι μόνο αποτελεσματικά αλλά και αρκετά γενικά. Για αυτό το λόγο η μέθοδος απαιτεί ελάχιστη παρέμβαση για την προσαρμογή της σε κάποιο νέο γνωστικό πεδίο. Ένα σημαντικό πρόβλημα του συστήματος από την άλλη πλευρά, είναι ο εντοπισμός της ακριβούς έκτασης του όρου και η αντιμετώπιση πιθανόν ασαφών δομών, όπως για παράδειγμα, στην περίπτωση παρατακτικής σύνδεσης. Για την αντιμετώπιση αυτών των περιορισμών είναι αναγκαίος ο εντοπισμός και άλλων διακριτικών χαρακτηριστικών εκμάθησης.

Μια άλλη μέθοδος (Nobata et al. 1999) εποπτευόμενης εκμάθησης που έχει χρησιμοποιηθεί για αναγνώριση όρων βασίζεται σε *δένδρα αποφάσεων* (*decision trees*) και χρησιμοποιεί τον αλγόριθμο C4.5 (Quinlan, 1993). Τα δεδομένα εκμάθησης που χρησιμοποιούν είναι σώμα κειμένων επισημειωμένο με μη αυτόματο τρόπο. Τα διακριτικά χαρακτηριστικά για την κατηγοριοποίηση περιλαμβάνουν μορφοσυντακτικές και γραφηματικές πληροφορίες που είναι καταχωρημένες σε λίστες όρων. Για την μορφοσυντακτική ανάλυση χρησιμοποιείται ο αναλυτής MXPOST (Ratnaparkhi, 1996). Επίσης διακρίνονται 28 γραφηματικές κατηγορίες λέξεων, που βασίζονται στους χαρακτήρες (κεφαλαία-μικρά), την ύπαρξη αριθμών και συμβόλων.

Βιβλιογραφικές αναφορές

Ananiadou, S. (1988): *Towards a Methodology for Automatic Term Recognition,* PhD thesis, University of Manchester Institute of Science and Technology, 1988.

Ananiadou, S. (1994): «A Methodology for Automatic Term Recognition», στο *Proceedings of Coling94,* σ.1034-1038.

Ananiadou, S., Albert, S., Schuhmann, D. (2000): «Evaluation of Automatic Term Recognition of Nuclear Receptors from Medline», *Genome Informatics Series*, vol.11.

Ananiadou, S., Nenadic, G. (2001): «A Terminology Management Workbench for Molecular Biology», στο *Proceedings of the 19th Twente Workshop on Language Theory, Workshop on Information Extraction in Molecular Biology*, σ. 7-13.

Bourigault, D., Gonzalez-Mullier, I. & Gros, C.(1996): «LEXTER, a Natural Language Processing tool for Terminology Extraction» στο *Proceedings of the 7th Euralex International Congress,* Goteborg.

Collier, N., Nobata, C. & Tsujii, J. (2000): «Extracting the Names of Genes and Gene Products with a Hidden Markov Model», στο *Proceedings of Coling2000,* σ.201-207.

Dagan, I. & Church, K. (1995): «Termight: Identifying and translating technical terminology», στο *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, EACL'95,* σ.34-40.

Daille, B., Gaussier, E. & Lange, J. (1994): «Towards automatic extraction of monolingual and bilingual terminology», στο *Proceedings of Coling94,* σ. 515-521.

Damerau, F.J. (1993): «Evaluating domain-oriented multi-word terms from texts», *Information Processing and Management,* 29(4), σ.433-447.

Dillon, M. & Gray, A. (1983): «FASIT: Fully automatic syntax-based indexing», *Journal of the American Society for Information Science*, 34(2), σ. 99-108.

Dunning, T. (1993): «Accurate methods for the statistics of surprise and coincidence», *Computational Linguistics,* 19(1) σ. 61-74.

Enguehard, C., & Pantera, L. (1994): «Automatic natural acquisition of a terminology», *Journal of Quantitative Linguistics,* 2(1), σ.27-32.

Evans, D.A. & Lefferts, R.G. (1995): «CLARIT-TREC Experiments», *Information Processing and Management,* 31(3), σ. 385-389.

Fano, R.M. (1961): *Transmission of Information. A statistical theory of communications*. MIT Press, Ma.

Frantzi, K.T. & Ananiadou, S. (1995): «Statistical measures for terminological extraction», στο *Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data*, *(JADT'95),* σ.297-308.

Frantzi, K.T. & Ananiadou, S. (1999): «The C/NC value domain independent method for multi-word term extraction», *Journal of Natural Language Processing,* 6(3), σ. 145-180.

Gaizauskas, R., Demetriou, G. & Humphreys, K. (2000): «Term Recognition in Biological Science Journal articles», στο Ananiadou, S. & Maynard, D. (eds.) *Workshop on Computational Terminology for Medical and Biological Applications (NLP2000, Patras, Greece),* σ. 37-44.

Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H. & Wilks, Y. (1995): «Description of the LaSIE system as used for MUC-6.» στο *Proceedings of the Sixth Message Understanding Conference (MUC-6),* Morgan Kaufmann, σ. 207-220.

Gale, W.A. & Church, K.W. (1991): «Concordances for parallel texts», στο *Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research Using Corpora*, σ. 40-62.

ISO704 (1986): *Principles and Methods of Terminology,* International Organization for Standardization.

Jaquemin, C. & Tzoukermann, E. (1999): «NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax», στο Strzalkowski, T. (ed.) *Natural Language Information Retrieval*, Kluwer, Boston, MA, σ.25-74.

Jaquemin, C. (2001): *Spotting and Discovering Terms through Natural Language Processing,* The MIT Press, Cambridge, MA.

Justeson, J.S., Katz, S.M. (1995): «Technical terminology: some linguistic properties and an algorithm for identification in text», *Natural Language Engineering,* 1(1), σ.9-27.

Kita, K., Kato, Y., Omoto, T. & Yano, Y. (1994): «A comparative study of automatic extraction of collocations from corpora: Mutual Information vs. cost criteria.», στο *Journal of Natural Language Processing,* 1(1), σ. 21-33.

Lauriston, A. (1994): «Automatic recognition of complex terms: Problems and the Termino solution», *Terminology,*1(1), σ. 147-170.

Maynard, D. & Ananiadou, S. (2000a): «Identifying Terms by their Family and Friends», στο *Proceedings of the 18th International Conference on Computational Linguistics, COLING 2000*, σ. 530-536.

Maynard, D. & Ananiadou, S. (2000b): «TRUCKS: A Model for Automatic Multi-Word Term Recognition», *Journal of Natural Language Processing*, 8(1), σ. 101-125.

Mima, H., Ananiadou, S., Nenadic, G. (2001a): «Improving Knowledge Acquisition Through Automatic Term Recognition», στο *Proceedings of PC Human-Computer Interaction Conference, (PC-HCI 2001), Patras, Greece.*

Mima, H., Ananiadou, S., Nenadic, G. (2001b): «The ATRACT Workbench: Automatic Term Recognition and Clustering for Terms», στο *Text, Speech and Dialogue, TSD 2001,* Lecture Notes in Artificial Intelligence, Springer Verlag.

Nobata, C., Collier, N. & Tsujii, J. (1999): «Automatic term identification and classification in biology texts», στο *Proceedings of the Natural Language Pacific Rim Symposium (NLPRS'2000),* σ. 369-375.

Porter, M.F. (1980): «An algorithm for suffix stripping», *Program*, 14(3) σ.130-137.

Quinlan, R.J.(1993): *C4.5 Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.

Ratnaparkhi, A. (1996): «A Maximum Entropy Part-Of-Speech Tagger», στο *Proceedings of the Empirical Methods in Natural Language Processing Conference*, *May 17-18*, University of Pennsylvania.

Schulze-Kremer, S. (1998): «Ontologies for Molecular Biology», στο *Proceedings of the Third Pacific Symposium on Biocomputing*, AAAI Press, σ.693-704.

Vivaldi, J. & Rodriquez, H. (2000): «Improving term extraction by combining different techniques» στο Ananiadou, S. & Maynard, D. (eds.) *Workshop on Computational Terminology for Medical and Biological Applications (NLP2000, Patras, Greece),* σ. 61-68.