Ζητήματα αναγνώρισης των πολυλεκτικών σύμπλοκων όρων στον τομέα της βιοϊατρικής

Άννα Ιορδανίδου, Μαβίνα Πανταζάρα, Έλενα Μάντζαρη, Γιώργος Ορφανός
Αριστείδης Βαγγελάτος, Βασίλης Παπαπαναγιώτου

ΠΕΡΙΛΗΨΗ

Αντικείμενο της ανακοίνωσης αποτελεί η αναλυτική περιγραφή των δομών με βάση τις οποίες σχηματίζονται οι πολυλεκτικοί σύμπλοκοι όροι στον τομέα της βιοϊατρικής, με στόχο την αυτόματη αναγνώριση ή/και εξαγωγή τους από κείμενα. Η παρούσα εργασία στηρίζεται σε έρευνα που διεξάγεται στο πλαίσιο του ερευνητικού έργου ΙΑΤΡΟΛΕΞΗ.

About recognition of Greek multi-word complex terms in the biomedical domain

Anna Iordanidou, Mavina Pantazara, Elena Mantzari, George Ophanos, Aristides Vagelatos, Vassilis Papapanagiotou

SUMMARY

The paper aims at describing the morpho-syntactic and semantic formation patterns of multi-word complex terms in the biomedical domain, for their automatic recognition or/and extraction from texts. This work is based on the research that is carried out within the framework of the R&D project IATROLEXI.

0. Εισαγωγή - Πλαίσιο της έρευνας

Η παρούσα ανακοίνωση στηρίζεται σε έρευνα που διεξάγεται στο πλαίσιο του ερευνητικού έργου ΙΑΤΡΟΛΕΞΗ[1], στόχος του οποίου είναι η δημιουργία πληροφορικής υποδομής (υπολογιστικών εργαλείων και γλωσσικών πόρων) για τη διαχείριση και επεξεργασία ελληνικών βιοϊατρικών κειμένων. Πιο συγκεκριμένα, στα τελικά προϊόντα του έργου θα περιλαμβάνονται: *σώμα ελληνικών βιοϊατρικών κειμένων* (corpus of Greek biomedical texts), *λεξικό βιοϊατρικών όρων με οντολογική πληροφορία* και μια *σειρά εργαλείων για το δομικό, μορφοσυντακτικό και σημασιολογικό σχολιασμό* (structural, morpho-syntactic and semantic annotation) των κειμένων.

Αντικείμενο της ανακοίνωσης αποτελεί η αναλυτική περιγραφή των πολυλεκτικών σύμπλοκων όρων (multi-word complex terms) στον τομέα της βιοϊατρικής, με στόχο την αυτόματη αναγνώριση ή/και εξαγωγή τους από κείμενα. Μελετήθηκαν όροι από διάφορες σημασιολογικές κατηγορίες, όπως: *γονίδια* (π.χ. *γονίδιο pol*), *ασθένειες* (π.χ. *πνευμονικό οίδημα*), *σύνδρομα* (π.χ. *σύνδρομο Munchausen*), *όργανα ή μέρη του σώματος* (π.χ. *ιγνυακή αρτηρία*), *συμπτώματα* (π.χ. *παρεγκεφαλιδική αταξία*), *ιοί* (π.χ. *ιός ηπατίτιδας Β*), *χημικές και φαρμακευτικές ουσίες* (π.χ. *ισότοπο ιωδίου, πενικιλίνη G, εμβόλιο AIDS*), *διαδικασίες θεραπείας και πρόληψης* (π.χ. *πλύση στομάχου*, *προληπτική ορθοδοντική*).

Στόχος της μελέτης είναι η γλωσσική επεξεργασία των πολυλεκτικών όρων μέσα από αυτόνομα επίπεδα ανάλυσης (συντακτικό, σημασιολογικό κτλ.), τα οποία αλληλοσυμπληρώνονται και αλληλοενισχύονται ώστε να επιτευχθεί μεγαλύτερη ακρίβεια στη διαδικασία αναγνώρισής τους. Τα αυτόνομα αυτά επίπεδα ανάλυσης, όπως θα παρουσιαστούν στη συνέχεια, αφορούν συγκεκριμένα: τον εντοπισμό των γενικών μορφοσυντακτικών δομών σχηματισμού των πολυλεκτικών όρων (απλών και επαυξημένων), το λεπτομερέστερο προσδιορισμό των λεξικών στοιχείων που τις απαρτίζουν, τη διαχείριση της μορφολογικής ποικιλίας και τον εντοπισμό σημασιολογικών σχημάτων συνεμφάνισης των συνθετικών τους για ορισμένες σημασιολογικές κατηγορίες.

1. Συγκέντρωση του γλωσσικού υλικού

Για τη συγκέντρωση των πολυλεκτικών (σύμπλοκων) όρων της βιοϊατρικής ακολουθήθηκαν οι εξής διαδικασίες:

[1] Το ερευνητικό έργο ΙΑΤΡΟΛΕΞΗ (www.iatrolexi.gr) χρηματοδοτείται μερικώς από τη Γενική Γραμματεία Έρευνας και Τεχνολογίας (ΓΓΕΤ) – στο πλαίσιο του Μέτρου 3.3 του Προγράμματος «Κοινωνία της Πληροφορίας».

1. **Επεξεργασία του ελληνικού λεξιλογίου βιοϊατρικών όρων του MeSH-Hellas** [1]. Ο στόχος αυτής της εργασίας ήταν διττός: αφενός, να εντοπιστεί το βασικό λεξιλόγιο της βιοϊατρικής ορολογίας που εκφράζεται μέσω πολυλεκτικών δομών και, αφετέρου, να συγκεντρωθεί ένα ικανό και επαρκές σώμα όρων το οποίο θα χρησιμεύσει ως βάση για τον προσδιορισμό μορφοσυντακτικών σχημάτων (morphosyntactic patterns) με στόχο την αναγνώριση και εξαγωγή πολυλεκτικών όρων από κείμενα.
2. **Έλεγχος των πολυλεκτικών όρων του MeSH-Hellas στο Σώμα Ελληνικών Βιοϊατρικών Κειμένων του ΙΑΤΡΟΛΕΞΗ** (11,5 εκ. λέξεων). Το σώμα κειμένων επιβεβαίωσε ως υπαρκτό ένα μεγάλο ποσοστό των μορφοσυντακτικών σχημάτων που είχαν εντοπιστεί στο λεξιλόγιο του MeSH, ενώ ταυτόχρονα συντέλεσε στην καταγραφή της μορφολογικής και συντακτικής τους ποικιλίας.

Ο προσδιορισμός των μορφοσυντακτικών σχημάτων βασίστηκε στην ανάλυση των σχετικών δομών περίπου 6.000 βιοϊατρικών πολυλεκτικών όρων. Η διάκριση σε δομές, ανάλογα με τον αριθμό των οροστοιχείων που περιέχουν, επέτρεψε:

1. Την ακριβή μέτρηση του αριθμού των αντίστοιχων όρων και την ποσοστιαία συμμετοχή τους στο σύνολο του εξεταζόμενου υλικού. Οι μετρήσεις συνοψίζονται ως εξής:

- Το **82**% των πολυλεκτικών όρων είναι *διλεκτικοί*
- Το **15**% *τριλεκτικοί*
- Το **1,9**% *τετραλεκτικοί.*

Τα συγκεκριμένα αποτελέσματα μπορούν να αξιοποιηθούν στις στατιστικές μετρήσεις που θα εφαρμοστούν κατά την αναγνώριση και την εξαγωγή των πολυλεκτικών όρων από τα κείμενα.

2. Τη μελέτη των πρωτογενών δομών, οι οποίες εμφανίζονται κυρίως στους διλεκτικούς όρους, καθώς και των επαυξημένων δομών τους που εμφανίζονται στους τριλεκτικούς και τετραλεκτικούς όρους.

2. Προσδιορισμός του πολυλεκτικού όρου

Στη γενική γλώσσα, μία πολυλεκτική λεξική μονάδα διαφοροποιείται από μία ελεύθερη φράση ως προς τη μεγαλύτερη συνοχή που παρουσιάζει σε επίπεδο φωνολογίας, γραφής, μορφολογίας, σημασιολογίας, σύνταξης και χρήσης [2]. Ο βαθμός λεξικοποίησης μιας

πολυλεκτικής λεξικής μονάδας μπορεί να ποικίλλει, εφόσον «το μόνο σοβαρό κριτήριο για να διαπιστωθεί η λεξικοποίηση μιας μη μονολεκτικής λεξικής μονάδας είναι το κριτήριο του ενός και σταθερού αντικειμένου αναφοράς, το οποίο συνεπάγεται διατύπωση ορισμού για τη λεξική φράση» [2, σ. 142].

Στις ειδικές γλώσσες, *πολυλεκτικός σύμπλοκος όρος* ονομάζεται κάθε όρος που αποτελείται από δύο ή περισσότερες λέξεις, τα *οροστοιχεία*[2] του, και ο οποίος αντιπροσωπεύει μία αντίστοιχη έννοια, ειδική ή γενική, του σχετικού θεματικού πεδίου [3, 4, 5].

Ο πολυλεκτικός όρος αποτελεί ένα δυαδικό σχηματισμό που μπορεί να αναλυθεί σε δύο μέρη, τα άμεσα συνθετικά, τα οποία αποτελούνται από ένα ή περισσότερα οροστοιχεία το καθένα. Συνήθως, το ένα από τα άμεσα συνθετικά είναι το *προσδιοριζόμενο* και το άλλο το *προσδιοριστικό*.

Στο πλαίσιο της παρούσας μελέτης, οι πολυλεκτικοί όροι που εξετάζουμε είναι ονοματικές φράσεις, π.χ. *νωτιαίος μυελός*, *ιός της λύσσας*, που αντιστοιχούν σε καθιερωμένες έννοιες του βιοϊατρικού τομέα.

3. Συντακτική ανάλυση των πολυλεκτικών όρων

Ο καθορισμός των δομών σχηματισμού των πολυλεκτικών όρων αφορά:

α) τον προσδιορισμό των επιτρεπόμενων συντακτικών δομών τους και τη διατύπωση κανόνων για το συνδυασμό των οροστοιχείων τους βάσει μορφοσυντακτικών χαρακτηριστικών και περιορισμών (π.χ. γραμματικών κατηγοριών, πτωτικών σχέσεων, ύπαρξη ή μη λειτουργικών λέξεων κτλ.),

β) την αποτύπωση της εσωτερικής δομής τους βάσει της ορολογικής τους σύστασης, δηλαδή βάσει της πληροφορίας για το ποια από τα οροστοιχεία τους είναι όροι ή λέξεις της γενικής γλώσσας.

3.1. Σχηματισμός και ιδιότητες των διλεκτικών όρων

Οι διλεκτικοί όροι που μελετήθηκαν κατατάσσονται στις τέσσερις γενικές κατηγορίες πολυλεκτικών ονοματικών σύνθετων της νέας ελληνικής:

[2] Οροστοιχείο είναι κάθε μορφολογικό μέρος ενός σύμπλοκου όρου (μονολεκτικού ή πολυλεκτικού) το οποίο είναι φορέας σημασίας ή έχει σημασιολογικά διαφοροποιητική αξία.

1. επίθετο+ουσιαστικό, π.χ. *ρυθμιστικό γονίδιο*, *οφθαλμικό διάλυμα*, *βραχιόνιο πλέγμα*
2. ουσιαστικό+ουσιαστικό σε γενική, π.χ. *συνθάση χιτίνης, κοιλίες του εγκεφάλου, ιός της λύσσας*
3. ουσιαστικό+πρόθεση+ουσιαστικό, π.χ. *αντισυλληπτικά από το στόμα*, *δηλητηρίαση από σταφυλόκοκκο*
4. ουσιαστικό+ουσιαστικό (ομοιόπτωτο), π.χ. *βάκιλος μεγαθήριο*, *μέλος φάντασμα*.

Οι σχέσεις ανάμεσα στα συνθετικά τους είναι είτε σχέσεις *παράταξης* (δομή 4) είτε σχέσεις *εξάρτησης* (δομές 1, 2, 3). Στη βιοϊατρική, οι όροι της δομής 4 εκφράζουν κυρίως μεταφορά. Οι όροι των δομών 1, 2, 3 αποτελούν *ενδοκεντρικά σύνθετα* (composés endocentriques), με άλλα λόγια το προσδιοριζόμενο ουσιαστικό αποτελεί την κεφαλή· η σχέση ανάμεσα στα δύο συνθετικά είναι συνήθως προσδιοριστική (π.χ. *ερυθρά αιμοσφαίρια*), αν και σε μερικές περιπτώσεις παρατηρείται σχέση κατηγορήματος (π.χ. *μεταδοτικό νόσημα → νόσημα που μεταδίδεται*).

Οι πολυλεκτικοί όροι που μελετήσαμε υπόκεινται στους ακόλουθους περιορισμούς, γεγονός που αποδεικνύει υψηλό βαθμό παγίωσης [6]:

- κανένα συνθετικό δεν μπορεί να προσδιοριστεί ξεχωριστά (π.χ. *το νευρικό σύστημα / *το πιο νευρικό σύστημα, το ρυθμιστικό γονίδιο / *το πολύ ρυθμιστικό γονίδιο / *το ρυθμιστικό αυτοάνοσο γονίδιο*),
- δεν είναι δυνατή η παρατακτική ή συμπλεκτική σύνδεση ενός συνθετικού με ένα ξένο στοιχείο (π.χ. *άμεσες καρδιακές μαλάξεις / *άμεσες, καρδιακές μαλάξεις, το κεντρικό νευρικό σύστημα / *το κεντρικό και νευρικό σύστημα*),
- υπάρχουν περιορισμοί ως προς την παρουσία / απουσία άρθρων (π.χ. *δηλητηρίαση από σταφυλόκοκκο / *δηλητηρίαση από το σταφυλόκοκκο*),
- υπάρχουν περιορισμοί στον αριθμό (π.χ. *νεογνικός ίκτερος* / **νεογνικοί ίκτεροι*, *δίαυλοι καλίου* / **δίαυλος καλίου*),
- το προσδιοριστικό επίθετο δεν μπορεί να λειτουργήσει ως κατηγορούμενο (**το οστό είναι ιερό*, **το σύμπλεγμα είναι βιταμινών Β*, **η ευπάθεια είναι σε νόσο*, **το μέλος είναι φάντασμα*),
- σε ορισμένα συμφραζόμενα είναι δυνατή η απαλοιφή κάποιου συνθετικού (π.χ. *νουκλεοσίδιο Q → Q*, *το ουροποιητικό σύστημα → το ουροποιητικό*),

- δεν είναι δυνατή συνήθως η αντικατάσταση ενός συνθετικού από κάποιο συνώνυμο ή αντώνυμό του (π.χ. *υποδοχείς αντιγόνου* / **δέκτες αντιγόνου*, *οδοντική εξαγωγή* / **οδοντική εισαγωγή*).

Δεδομένου ότι, σε αντίθεση με τις λέξεις της γενικής γλώσσας, οι όροι σχηματίζονται με στόχο όχι μόνο την κατασήμανση (designation), αλλά και την κατηγοριοποίηση και ταξινόμηση εννοιών, κατά κανόνα, το προσδιοριστικό συνθετικό προσθέτει κάποιο χαρακτηριστικό που μετατρέπει τη *γένια έννοια* (generic concept) του προσδιοριζόμενου συνθετικού σε *είδια έννοια* (specific concept) του σύμπλοκου όρου. Το προσδιοριστικό συνθετικό συντελεί είτε στη *δυαδική ταξινόμηση* του προσδιοριζόμενου συνθετικού (π.χ. *λευκά / ερυθρά αιμοσφαίρια*, *σακχαρώδης / άποιος διαβήτης*) ή στην *πολλαπλή ταξινόμηση* (π.χ. *ακουστικό / οπτικό / γευστικό / οσφρητικό νεύρο*) είτε στη *μοναδική ταξινόμηση* (π.χ. *ομφάλιος λώρος, διαβητική κετοξέωση, λοιμώδης μονοπυρήνωση*).

3.2. Σχηματισμός των τριλεκτικών και τετραλεκτικών όρων

Το βασικό μοντέλο σχηματισμού δομών που περιγράψαμε παραπάνω έχει τη δυνατότητα να αναπτύσσεται και να γίνεται όλο και πιο περίπλοκο. Έτσι, μία διλεκτική μονάδα μπορεί να αποτελέσει τη βάση για το σχηματισμό μίας νέας, μεγαλύτερης πολυλεκτικής μονάδας. Πιο αναλυτικά, η παραγωγή των επαυξημένων (τριλεκτικών και τετραλεκτικών) δομών γίνεται με βάση τις ακόλουθες διεργασίες [7, 8]:

- προσθήκη νέου προσδιοριστικού

  νευρικό σύστημα → κεντρικό νευρικό σύστημα

  λιπαρά οξέα → ακόρεστα λιπαρά οξέα

- προσθήκη νέου προσδιοριζόμενου

  ερυθρά αιμοσφαίρια → όγκος ερυθρών αιμοσφαιρίων

- σύνδεση προσδιοριστικών (παρατακτική, συμπλεκτική, διαζευκτική)

  άνω κοιλία, κάτω κοιλία → άνω-κάτω κοιλία

  αποικία κοκκιοκυττάρων, αποικία μακροφάγων → αποικία κοκκιοκυττάρων-μακροφάγων

  κιρσοί οισοφάγου, κιρσοί στομάχου → κιρσοί οισοφάγου και στομάχου

μεσοκολπικό διάφραγμα, μεσοκοιλιακό διάφραγμα → μεσοκολπικό ή μεσοκοιλιακό διάφραγμα

- σύνδεση προσδιοριζόμενων

  χολικά οξέα, χολικά άλατα → χολικά οξέα και άλατα

  σημεία λοίμωξης, συμπτώματα λοίμωξης → σημεία και συμπτώματα λοίμωξης

3.3. Αποτύπωση της ορολογικής σύστασης των πολυλεκτικών όρων

Η απλή περιγραφή της συντακτικής δομής των πολυλεκτικών όρων, η οποία ακολουθεί τους γενικούς κανόνες σχηματισμού των ονοματικών συνθέτων, και η διατύπωση κανόνων φραστικής δομής δε φαίνεται να επαρκούν για τον εντοπισμό και την αναγνώρισή τους μέσα σε κείμενα, δεδομένου ότι μία επιτρεπόμενη συντακτική δομή δε συνιστά απαραίτητο κριτήριο για το σχηματισμό βιοϊατρικού όρου. Στόχος αυτού του επιπέδου ανάλυσης είναι ο επαναπροσδιορισμός των συντακτικών δομών λαμβανομένης υπόψη της εσωτερικής τους σύστασης ως προς τη συνεμφάνιση *όρων* (terms) και *μη όρων* (non-terms). Η πληροφορία για το αν ένα ουσιαστικό είναι *μονολεκτικός όρος* (ΜΟρ) αντλείται από το λεξικό βιοϊατρικών όρων του ΙΑΤΡΟΛΕΞΗ.

Η Maynard [9], κατά τη μελέτη των δομών σχηματισμού (formation patterns) των βιοϊατρικών πολυλεκτικών όρων στα αγγλικά, αναγνωρίζει δύο είδη συνθετικών: τα *επίθετα* και τις *ονοματικές φράσεις*, οι οποίες μπορεί να είναι όροι ή κοινές λέξεις. Επιπλέον, διακρίνει τους όρους που αποτελούν συστατικά άλλων όρων σε τρεις κατηγορίες: *διλεκτικούς* (ΔΟρ), *τριλεκτικούς* (ΤΡΟρ) και *τετραλεκτικούς*. Με βάση το μοντέλο της Maynard, διαμορφώθηκε ο παρακάτω πίνακας, στον οποίο καταγράφονται οι δυνατοί συνδυασμοί όρων και μη όρων προκειμένου για το σχηματισμό βιοϊατρικών πολυλεκτικών όρων από δύο έως τέσσερις λέξεις για τα ελληνικά:

Μήκος	**Σχήμα**	**Παράδειγμα**
2	Ουσ+Ουσ	παράγοντας συμπληρώματος
2	Ουσ +ΜΟρ	απώλεια δοντιών, όγκος παλμού
2	ΜΟρ +Ουσ	καθετήρες διαρκείας
2	ΜΟρ+ ΜΟρ	ιός της λύσσας, έλκος κνημών
2	Ε+Ουσ	υαλώδες σώμα, αυξητικός παράγοντας, ερυθρός πυρήνας

3	Ε+Ε+Ουσ	βραδεία αρνητική μεταβολή
3	Ε+Ε+ΜΟρ	προσωρινή μερική οδοντοστοιχία
3	Ε+ΔΟρ	κεντρικό νευρικό σύστημα, ανοικτός αρτηριακός πόρος
3	ΜΟρ+ΔΟρ	ιός δάγκειου πυρετού
4	Ουσ+ ΤΡΟρ	χρόνος πήξης ολικού αίματος, παρόξυνση Χ.Α.Π.
4	Ε+ ΤΡΟρ	ανοσοβλαστικό λέμφωμα μεγάλων κυττάρων
4	ΤΡΟρ+ ΜΟρ	λείες μυϊκές ίνες αγγείων
4	ΜΟρ+ ΤΡΟρ	σύνδρομο απόσυρσης τοξικής ουσίας
4	ΔΟρ+ ΔΟρ	υποφυσιακές ορμόνες οπίσθιου λοβού, φακοί επαφής συνεχούς χρήσης

4. Κατηγορίες των οροστοιχείων

Εκτός από λέξεις της νέας ελληνικής (επίθετα, ουσιαστικά, προθέσεις, άρθρα, σύνδεσμοι), αναγνωρίστηκαν οι εξής κατηγορίες οροστοιχείων:

1. *Λεξικά στοιχεία*
   - λόγιοι τύποι (π.χ. *έσω ους, ιερό οστούν*)
   - ελληνικά ακρώνυμα (π.χ. *Χ.Α.Π.*, *Ο.Μ.Λ.*)
   - λέξεις με λατινικούς χαρακτήρες (π.χ. *laser*)
   - ονόματα με λατινικούς χαρακτήρες (π.χ. *σύνδρομο Down*)
   - ακρώνυμα με λατινικούς χαρακτήρες (π.χ. *DNA, RNA*)
2. *Μη λεξικά στοιχεία*
   - αριθμοί (π.χ. *ιντερλευκίνη 1, ημιχολίνιο 3*)
   - ελληνικά και λατινικά γράμματα (π.χ. *βιταμίνη C, ακτίνες βήτα*)
   - σημεία στίξης, όπως παύλες, κόμματα, παρενθέσεις (π.χ. *1,4-άλφα-γλυκοσιδάση γλυκάνης*)

Συχνά παρατηρείται συνδυασμός αριθμών και γραμμάτων (π.χ. *ωμέγα-3 λιπαρά οξέα, κύτταρα 3Τ3*), καθώς και συνδυασμός μη λεξικών και λεξικών στοιχείων (π.χ. *Ν-αποθεμελιωτικές οξειδορεδουκτάσες, 2-αμινοαδιπικό οξύ, κυκλικά Ρ-οξείδια*).

5. Μορφολογική ποικιλία

Ποικιλία (term variation) υπάρχει όταν μία έννοια εκφράζεται με πολλούς συνώνυμους όρους ή με *παραλλαγές* (variants) του ίδιου όρου. Το φαινόμενο της ποικιλίας απορρέει από τη δυνατότητα της φυσικής γλώσσας να εκφράζει μία έννοια με περισσότερους τρόπους και είναι ιδιαίτερα συχνό στην ορολογία. Υπολογίζεται ότι το 37% των όρων που απαντούν σε ένα κείμενο αποτελούν παραλλαγές [10]. Ο εντοπισμός και η αναγνώριση αυτών των παραλλαγών είναι κεφαλαιώδους σημασίας για εφαρμογές όπως η εξαγωγή και η ανάκτηση πληροφορίας (information extraction and retrieval) αλλά και γενικότερα για την ορολογική τεκμηρίωση, αφού είναι σημαντικό να διακρίνεται αν και κατά πόσο διαφορετικές μορφές όρων αναφέρονται στην ίδια έννοια ή δε σχετίζονται μεταξύ τους [11].

Στον τομέα της βιοϊατρικής, όπως και στη γενική γλώσσα, εντοπίζονται τα εξής είδη ποικιλίας: ορθογραφική, μορφολογική, συντακτική και λεξικο-σημασιολογική (βλ. [9]). Στην παρούσα μελέτη δε θα ασχοληθούμε με τη σημασιολογική ποικιλία, δηλαδή με την ύπαρξη περισσότερων συνώνυμων όρων για την κατασήμανση της ίδιας έννοιας, αλλά θα περιοριστούμε στην εξέταση των παραλλαγών της μορφής του ίδιου όρου. Με βάση το υλικό που επεξεργαστήκαμε καταγράψαμε τις ακόλουθες περιπτώσεις ποικιλίας:

α) σε επίπεδο γραφής (χρήση/απουσία παύλας, πεζά/κεφαλαία, αναπαράσταση ελληνικών γραμμάτων, αναπαράσταση αριθμητικών, μεταγραφή ξένων ονομάτων), π.χ. ***ωμέγα-3*** / ***ωμέγα 3*** *λιπαρά οξέα,* ***1,4-άλφα-γλυκοσιδάση*** / ***1, 4 άλφα γλυκοζιδάση*** *γλυκάνης,* ***β-αλανίνη*** / ***βήτα αλανίνη***, ***α*** *1-αντιθρυψίνη* / ***άλφα*** *1-αντιθρυψίνη,* ***enterococcus*** / ***εντερόκοκκος*** *faecium,* ***διοξυγονάση*** / ***διοξυγενάση*** *4-υδροξυφαινυλοπυροσταφυλικού* κτλ.

β) σε επίπεδο ορθογραφίας (παλαιότερη/νεότερη ορθογραφία, παράλληλοι τύποι), π.χ. ***μπορέλια*** / ***μπορρέλια*** *burgdorferi,* ***βάκιλος*** / ***βάκιλλος*** *άνθρακα,* ***διοξείδιο*** / ***διοξίδιο*** *του θείου* κτλ.

γ) σε επίπεδο μορφολογίας (παραγωγή, σύνθεση, σχηματισμός ελληνικών και ξένων ακρώνυμων, λόγιοι τύποι, αλλαγή πτώσης, αλλαγή γραμματικής κατηγορίας), π.χ. *χρόνια αναπνευστική πνευμονοπάθεια / Χ.Α.Π., υποδοχείς* ***ανοσοσφαιρίνης Ε*** */ υποδοχείς* ***IgE,*** *λιποπρωτεΐνες* ***υψηλής πυκνότητας*** */ λιποπρωτεΐνες* ***HDL,*** *διοξείδιο του* ***άνθρακα*** */ διοξείδιο του* ***άνθρακος,*** *αρτηρίες* ***του εγκεφάλου*** / ***εγκεφαλικές*** *αρτηρίες,* ***πρωτεϊνοκινάση*** *C /* ***πρωτεϊνική κινάση*** *C* κτλ.

6. Σημασιολογική ανάλυση

Η προσθήκη σημασιολογικής πληροφορίας στο σύστημα αναγνώρισης πολυλεκτικών όρων έχει στόχο να ενισχύσει τα προηγούμενα επίπεδα γλωσσικής ανάλυσης. Στο βαθμό που η παρούσα μελέτη αφορά το ειδικό λεξιλόγιο της βιοϊατρικής, και άρα ένα περιορισμένο λεξιλόγιο σε σχέση με τη γενική γλώσσα, η σημασιολογική ανάλυση αναδεικνύει τους συγκεκριμένους τρόπους με τους οποίους συγκεκριμένες σημασιολογικές κατηγορίες λέξεων ή όρων συνδυάζονται μεταξύ τους (βλ. και [12]).

Η ανάλυση που ακολουθήθηκε περιλαμβάνει δύο στάδια:

α) τη σημασιολογική κατηγοριοποίηση των όρων: η γνώση των σημασιολογικών κατηγοριών στις οποίες ανήκουν οι έννοιες, και κατ' επέκταση οι όροι που τις κατασημαίνουν, αντλήθηκε από την οντολογία του ΙΑΤΡΟΛΕΞΗ, η οποία στηρίχτηκε στο Μεταθησαυρό του UMLS (Unified Medical Language System) [13]. Για τα συνθετικά των όρων, κυρίως επίθετα, που δεν περιλαμβάνονται ως αυτόνομες έννοιες ούτε στην οντολογία του ΙΑΤΡΟΛΕΞΗ ούτε στο Μεταθησαυρό του UMLS χρειάστηκε να δημιουργηθούν νέες σημασιολογικές κατηγορίες.

β) τον προσδιορισμό των σχημάτων συνεμφάνισης των σημασιολογικών κατηγοριών: συγκεκριμένες υποκατηγορίες εννοιών συνδυάζονται σε συγκεκριμένα σχήματα συνεμφάνισης ώστε να σχηματιστούν οι πολυλεκτικοί όροι και αντίστοιχα η σημασιολογική τους κατηγορία. Θα πρέπει να διευκρινιστεί ότι η συγκεκριμένη προσέγγιση ισχύει μόνο για τους πολυλεκτικούς όρους των οποίων η σημασία μπορεί να εξαχθεί από τη σημασία των επιμέρους συνθετικών τους (composiotionality).

Ενδεικτικά, παρουσιάζεται η ανάλυση για τους πολυλεκτικούς όρους που αντιστοιχούν στη σημασιολογική κατηγορία «Ασθένεια, διαταραχή ή σύνδρομο», και στον πίνακα 1 παρουσιάζονται οι σημασιολογικές κατηγορίες των συνθετικών τους.

Σημασιολογική κατηγορία προσδιοριστικού	**Παράδειγμα**
Ασθένεια, διαταραχή ή σύνδρομο	νεφρωσικό **σύνδρομο**, οφθαλμική **υπόταση**
Ανατομική ανωμαλία	**ανεύρυσμα** αορτής
Σημασιολογική κατηγορία προσδιοριστικού	
Μέρος του σώματος (Μσώματος)	**δερματικό** νόσημα, αδένας **μαστού**
Τόπος ή Θέση	**οπίσθιο** γάγγλιο
ΘέσηΜσώματος	**μεσοκολπικό** διάφραγμα
Λειτουργία	**αιμοφόρα** αγγεία
Σχήμα	**αστεροειδές** γάγγλιο
Υφή	**ινώδης** χιτώνας
Μέγεθος	**μεγάλος** μαστός
Ηλικία	**νεανικός** μαστός
Φύλο	**ανδρικά** γεννητικά όργανα
Ιδιότητα	**εμπύρετη** λοίμωξη
Βαθμός	**οξεία** λοίμωξη
Σειρά	**τρίτος** γομφίος
Αιτία	**ρευματικός** πυρετός, λοίμωξη **από** σαλμονέλα
Άτομο ή ομάδα	νόσος των **πτηνών**, νοσήματα **αγροτών**
Οργανική λειτουργία	**διατροφή** βρέφους
Ιστός	διαταραχές κινητικότητας κροσσών **επιθηλίου**
Ιός	λοιμώξεις από **ρεοϊούς**
Μύκητας	λοιμώξεις από **μικρόσπορα**
Βακτήριο	λοιμώξεις από **μπορέλια**

Πίνακας 1: Σημασιολογικές κατηγορίες προσδιοριζόμενων και προσδιοριστικών συνθετικών

Ο πίνακας 2 περιγράφει κάποια χαρακτηριστικά σημασιολογικά σχήματα συνεμφάνισης των συνθετικών των πολυλεκτικών όρων που ανήκουν στην κατηγορία «Ασθένεια, διαταραχή ή σύνδρομο». Όπως γίνεται φανερό, σε μία μορφοσυντακτική δομή μπορούν να αντιστοιχούν περισσότερα σχήματα συνεμφάνισης σημασιολογικών κατηγοριών, αφού η σημασιολογική κατηγορία του πολυλεκτικού όρου μπορεί να εκφράζεται μέσα από διαφορετικούς σημασιολογικούς συνδυασμούς του προσδιοριζόμενου και του προσδιοριστικού.

Μορφοσυντακτική δομή	**Σημασιολογική δομή**	**Παράδειγμα**
Ε+Ουσ	*Μσώματος ή Ιστός ή Κύτταρο ή Γονίδιο+Ασθένεια*	μυοσκελετικά νοσήματα, οισοφαγική καντιντίαση, κολπικό ανεύρυσμα κυτταρική νέκρωση
	ΘέσηΜσώματος+Ασθένεια	επιχείλιος έρπητας
	Αιτία+Ασθένεια	αλλεργική επιπεφυκίτιδα, διαβητικό κώμα, ρευματική καρδιοπάθεια, νεφρογενής υπέρταση
	Ιδιότητα+Ασθένεια	εκφυλιστική αρθρίτιδα ανίατη νόσος
Ουσ+ΟυσΓεν	*Ασθένεια+Μσώματος*	νοσήματα εγκεφάλου
	ΑνατΑνωμαλία+Μσώματος	κύστη παγκρέατος πρόπτωση ορθού στένωση τραχείας
	Ασθένεια+Ιστός ή Κύτταρο ή Γονίδιο	νοσήματα αίματος
	Ασθένεια+Οργανική λειτουργία	νοσήματα μεταβολισμού επιπλοκές εγκυμοσύνης
	Ασθένεια+Άτομο ή Ομάδα	νοσήματα αγροτών νοσήματα αιγοειδών
Ουσ+Όνομα	*Ασθένεια ή Σύνδρομο+Όνομα*	νόσος Wolman, νόσος Crohn, σύνδρομο Chiari-Frommel, σύνδρομο Down
Ε+Ε+Ουσ	*Αιτία+Ασθένεια*	ιογενή δερματικά νοσήματα, μυκητιασικά πνευμονικά νοσήματα
	Σύμπτωμα ή Βλάβη +Ασθένεια	εκζεματικά δερματικά νοσήματα, αποφρακτικά πνευμονικά νοσήματα
	Ιδιότητα+Ασθένεια	αυτοάνοσα δερματικά νοσήματα
Ουσ+ΠΡΟΘαπό+ΟυσΑιτ	*Ασθένεια+Ιός ή Μύκητας ή Βακτήριο ή Χλαμύδια*	λοιμώξεις από ουρεόπλασμα, μηνιγγίτιδα από λιστέρια, λοιμώξεις από σαλμονέλα
Ουσ+ΕΠΙΡτύπου+Σύμβολο	*Ασθένεια+Σύμβολο*	διαβήτης τύπου Ι, υπερλιποπρωτεϊναιμία τύπου V

Πίνακας 2: Σημασιολογικές δομές συνεμφάνισης προσδιοριζόμενων και προσδιοριστικών συνθετικών για τους όρους της κατηγορίας «Ασθένεια, διαταραχή ή σύνδρομο»

Ανεξάρτητα από τη μορφοσυντακτική δομή, το *προσδιοριζόμενο συνθετικό* ανήκει στην κατηγορία «Ασθένεια, διαταραχή ή σύνδρομο» (π.χ. *οισοφαγική* ***καντιντίαση***, ***νοσήματα*** *εγκεφάλου,* ***σύνδρομο*** *Down*, ***μηνιγγίτιδα*** *από λιστέρια*), εκτός από το σχήμα **ΑνατΑνωμαλία+Μσώματος**, όπου το προσδιοριζόμενο ανήκει στην κατηγορία «Ανατομική

ανωμαλία» (π.χ. ***στένωση*** *τραχείας*). Επίσης, στις απλές δομές το προσδιοριζόμενο που δηλώνει ασθένεια μπορεί να είναι κάποια *γένια έννοια* (π.χ. *νόσημα*, *νόσος*, *σύνδρομο*, *ασθένεια*, *λοίμωξη*, *διαταραχή* κτλ.) ή κάποια *είδια έννοια* (π.χ. *νεφρίτιδα*, *βαρηκοΐα*, *οίδημα*, *έλκος*, *νευρίτιδα*, *υπόταση*, *αρθρίτιδα*, *βρογχοκήλη*, *νευροπάθεια*, *κολίτιδα*, *έρπης*, *εξάνθημα*, *θυρεοειδής*, *αναιμία*, *σύφιλη*, *απόφραξη* κτλ.). Στην επαυξημένη δομή *Ε+Ε+Ουσ*, το προσδιοριζόμενο είναι συνήθως διλεκτικός όρος (π.χ. *αγγειακά* ***δερματικά νοσήματα***, *μυκητιασικές* ***οφθαλμικές λοιμώξεις***).

Αντίθετα, η σημασιολογική υφή του *προσδιοριστικού συνθετικού* ποικίλλει ανάλογα με τη μορφοσυντακτική δομή ή/και τη γραμματική ή λεξική πραγμάτωσή του:

- Ως *επίθετο* στην απλή δομή δηλώνει το «Μέρος του σώματος» (π.χ. ***καρδιακά*** *νοσήματα)* ή ειδικότερα τα μέρη των οργάνων ή υποσυστημάτων στα οποία εντοπίζεται μια ανώμαλη λειτουργία (π.χ. ***στεφανιαίο***, ***αορτικό***, ***κολπικό*** ή ***μεσοκολπικό*** *ανεύρυσμα),* τη «Θέση του μέρους σώματος» (π.χ. ***περιοδοντικό*** *απόστημα),* την «Αιτία», (π.χ. ***ρευματική*** *καρδιοπάθεια)*, την «Ιδιότητα» (π.χ. ***κληρονομική*** *νόσος),* ενώ στην επαυξημένη δομή *Ε+Ε+Ουσ* το «Σύμπτωμα ή βλάβη» με την οποία γίνεται αισθητή η ασθένεια (π.χ. ***εκζεματικά*** *δερματικά νοσήματα).*

- Ως *ουσιαστικό σε γενική* δηλώνει επίσης το «Μέρος του σώματος»[3] (π.χ. *στένωση* ***οισοφάγου***), τον «Ιστό» (π.χ. *λιπωμάτωση* ***περιτόναιου)***, την «Οργανική λειτουργία» (π.χ. *διαταραχές* ***αναπνοής***) ή την «Ομάδα ή το Άτομο» που προσβάλλεται από μια ασθένεια (π.χ. *νοσήματα* ***πτηνοτρόφων***). Στην περίπτωση κάποιων παγιωμένων δομών συνεμφάνισης, όπου το προσδιοριζόμενο είναι κάποιος από τους γένιους όρους *σύνδρομο*, *νόσος* και *διαταραχή,* το όνομα, συνήθως με λατινικούς χαρακτήρες, που ακολουθεί δηλώνει τον επιστήμονα που ανακάλυψε τη συγκεκριμένη παθολογική λειτουργία (π.χ. *νόσος* ***Crown***) ή ορισμένες φορές το άτομο που έχει ιδιότητες που ταυτίζονται ή παρομοιάζονται με τις ιδιότητες της σχετικής ασθένειας (π.χ. *σύνδρομο* ***Munchausen***). Το προσδιοριζόμενο, όπως φαίνεται και στον πίνακα 2, όταν αποτελεί συμπλήρωμα της προθετικής φράσης που εισάγεται με την πρόθεση *από* και δηλώνει την αιτία της ασθένειας, ανήκει σε μία από τις σημασιολογικές κατηγορίες «Ιός ή Μύκητας ή

[3] Παρατηρήθηκε εναλλαγή των δομών *Ε+ΟΥΣ* και *ΟΥΣ+ΟΥΣΓεν* στις περιπτώσεις όπου το προσδιοριστικό, *Ε* ή *ΟΥΣ*, ανήκει στην κατηγορία «Μέρος του σώματος» ή «Ιστός», π.χ. *νοσήματα* ***αίματος*** = ***αιματολογικά*** *νοσήματα, απόστημα* ***δέρματος*** = ***δερματικό*** *απόστημα*.

Βακτήριο ή Χλαμύδια» (π.χ. *πνευμονία από **μυκόπλασμα***, *παράλυση από **κρότωνες***).

7. Συμπεράσματα

Προκειμένου για την αυτόματη αναγνώριση ή/και εξαγωγή πολυλεκτικών σύμπλοκων όρων της βιοϊατρικής, παρουσιάστηκε μια πρόταση γλωσσικής ανάλυσης που στηρίζεται στον εντοπισμό σχημάτων συνεμφάνισης των συνθετικών τους, βάσει μορφοσυντακτικής, λεξικής, ορολογικής και σημασιολογικής πληροφορίας. Δεδομένου ότι ο τομέας της βιοϊατρικής είναι ιδιαίτερα ευρύς και φαίνεται να αποτελείται από περισσότερες υπογλώσσες με ιδιαίτερα χαρακτηριστικά η καθεμία (π.χ. κλινική υπογλώσσα, βιομοριακή υπογλώσσα κτλ.), θα πρέπει η ανάλυση να εξειδικευτεί σε καθεμία από αυτές προκειμένου να χειριστεί τις ιδιαιτερότητές τους με μεγαλύτερη ακρίβεια και αποτελεσματικότητα.

Βιβλιογραφία

[1] ΙΑΤΡΟΤΕΚ (1997). *Ελληνοαγγλικό και Αγγλοελληνικό Λεξικό Βιοϊατρικών Όρων (MeSH Hellas)*. Εταιρία Ιατρικών Σπουδών.

[2] Αναστασιάδη-Συμεωνίδη, Ά., (1986). *Η Νεολογία στην Κοινή Ελληνική*. Διδακτορική Διατριβή, Επιστημονική Επετηρίδα της Φιλοσοφικής Σχολής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, Παράρτημα Αρ. 65, Θεσσαλονίκη.

[3] Βαλεοντής Κ., Ζερίτη Κ., Νικολάκη Κ. (1999). *Το προσδιοριστικό συνθετικό του ελληνικού σύμπλοκου όρου και η χρήση της γενικής. Πρακτικά 2ου Συνεδρίου «Ελληνική Γλώσσα και Ορολογία»*. Αθήνα, σ. 283-316.

[4] Βαλεοντής, Κ., Ζερίτη Κ., Νικολάκη Ά., (2000). «Ο ελληνικός σύμπλοκος όρος και η χρήση της Γενικής ως προσδιοριστικού συνθετικού». (Εργασία που της απονεμήθηκε τιμητική διάκριση στο πλαίσιο των Διεθνών Βραβείων 2000 του Διεθνούς Κέντρου Πληροφοριών Ορολογίας (Infoterm).)

[5] ΕΛΟΤ 561-1:2006, Ορολογική εργασία – Λεξιλόγιο – Μέρος 1: Θεωρία και εφαρμογή. Αντίστοιχο του ISO 1087-1:2000, *Terminology work – Vocabulary – Part 1: Theory and application.*

[6] Αναστασιάδη-Συμεωνίδη, Ά. (1996). «Η νεοελληνική σύνθεση». *Ζητήματα νεοελληνικής γλώσσας: Διδακτική προσέγγιση*, (επιμ.) Γ. Κατσιμαλή, Φ. Καβουκόπουλος. Πανεπιστήμιο Κρήτης: Ρέθυμνο, σ. 97-120.

[7] Spasic, I., (2004). *A machine learning approach to term classification.* Διδακτορική διατριβή, Information Systems Research Centre, School of Computing, Science and Engineering University of Salford, Salford, UK.
[8] Γαβριηλίδου, Μ., Λαμπροπούλου Π., (2004). «Εξαγωγή όρων από κείμενα: Μια υβριδική μέθοδος». *Ελληνική ορολογία: Έρευνα και εφαρμογές*, (επιμ.) Κατσογιάννου, Μ. και Ε. Ευθυμίου. Εκδόσεις Καστανιώτη: Αθήνα, σ. 313-326.
[9] Maynard, D., (2000). *Term Recognition using Combined Knowledge Sources*. Διδακτορική διατριβή, Manchester Metropolitan University, Manchester, UK.
[10] Jacquemin, Chr., (2001). *Spotting and Discovering Terms through Natural Language Processing.* MIT Press.
[11] Ανανιάδου, Σ., Ζερβάνου, Κ. (2004). «Αναγνώριση όρων σε υπολογιστικά συστήματα: προβλήματα και μέθοδοι». *Ελληνική ορολογία: Έρευνα και εφαρμογές*, (επιμ.) Κατσογιάννου Μ. και Ε. Ευθυμίου. Εκδόσεις Καστανιώτη: Αθήνα, σ. 283-312.
[12] Friedman, C., Kra, P., Rzhetsky, A., (2002). "Two biomedical sublanguages: a description based on the theories of Zellig Harris". *Journal of Biomedical Informatics* (35), σ. 222-235. Διαθέσιμο στον ιστότοπο: www.sciencedirect.com
[13] Unified Medical Language System Documentation (UMLS), National Library of Medicine (NLM), U.S. Department of Health and Human Services. (Έκδοση 2006). Διαθέσιμο στον ιστότοπο: www.nlm.nih.gov/research/umls

Άννα Ιορδανίδου, Αν. καθηγήτρια ΠΤΔΕ Παν. Πατρών (A.Iordanidou@upatras.gr)
Μαβίνα Πανταζάρα, Δρ. γλωσσολόγος, Neurosoft A.E. (mavina@neurosoft.gr)
Έλενα Μάντζαρη, γλωσσολόγος, Neurosoft A.E. (emantz@tee.gr)
Γιώργος Ορφανός, Δρ. μηχανικός Η/Υ, Neurosoft A.E. (orphan@neurosoft.gr)
Αριστείδης Βαγγελάτος, Δρ. μηχανικός Η/Υ, ΕΑΙΤΥ (vagelat@cti.gr)
Βασίλης Παπαπαναγιώτου, ιατρός-καρδιολόγος (papapanagiotouv@yahoo.gr)