ΠΟΛΥΛΕΚΤΙΚΟΙ ΟΡΟΙ ΚΑΙ ΑΠΟΚΛΙΣΕΙΣ ΣΤΗΝ ΑΥΤΟΜΑΤΗ ΕΞΑΓΩΓΗ ΤΩΝ ΟΡΩΝ

Κατερίνα Κεχαγιά – Σοφία Ανανιάδου

ΠΕΡΙΛΗΨΗ

Η αναγνώριση των επιστημονικών και τεχνικών όρων είναι πρωταρχικής σημασίας στην ηλεκτρονική επεξεργασία κειμένων λόγω του πλήθους των ηλεκτρονικών κειμένων στο διαδίκτυο. Είναι επομένως εμφανής η ανάγκη δημιουργίας μεθόδων για την αυτόματη αναγνώριση και εξαγωγή των όρων αυτών. Η μέθοδος *C/ NC – value* είναι μια τέτοια μέθοδος, ιδιαίτερα αποτελεσματική και η οποία εφαρμόζεται σε κείμενα διαφορετικών επιστημονικών τομέων και γλωσσών. Ωστόσο τα αποτελέσματα της εξαγωγής όρων μπορούν να βελτιωθούν με την αναγνώριση συγγενών όρων – αποκλίσεων του ίδιου όρου και την διασύνδεσή τους.

Στην παρούσα εργασία εστιάζουμε στην απόκλιση όρων που οφείλονται σε συντακτικές δομές, με έμφαση σε δομές που απορρέουν από την συμπλεκτική σύνδεση όρων, δομές συχνές στα κείμενα Μοριακής Βιολογίας που αποτελούν τον χώρο της έρευνάς μας.

MULTI-WORD TERMS AND TERM VARIATION IN AUTOMATIC TERM EXTRACTION

Katerina Kehagia – Sofia Ananiadou

SUMMARY

Technical terms are important for text mining, especially as vast amounts of multilingual documents are available on Internet. Thus, a domain and language-independent method for term recognition is necessary to automatically recognize terms from documents. The C/NC-value method is an efficient domain-independent multi-word term recognition method, which combines linguistic and statistical knowledge. Nevertheless, the results of term extraction should be further analysed is order to identify term variants and perform term normalization.

In this paper we will concentrate on syntagmatic (syntactic) term variation, and specifically coordination, a rather frequently occurring example of syntagmatic terminological variation. We applied the C/NC value technique into a collection of 2000 abstracts from MEDLINE. As biomedical terms are in the process of being standardized they show high variation.

We applied a set of simple linguistic filters on the tagged corpus based on commonly used formation patterns. Term variation was applied on the output of C/NC-value. Finally, an evaluation of the syntagmatic term variant conflation is also provided.

0. ΑΥΤΟΜΑΤΗ ΕΞΑΓΩΓΗ ΟΡΩΝ

Δεν θα επεκταθούμε σε λεπτομέρειες σχετικά με την Αυτόματη Εξαγωγή Όρων. Θα περιοριστούμε μόνο να στην αναφορά των βασικών προσεγγίσεων που έχουν εφαρμοστεί στο χώρο από τους ερευνητές: βάσει της στατιστικής, της γλωσσολογίας και εκείνες που συνδυάζουν στοιχεία και από

τις δυο προαναφερθείσες πλευρές, *υβριδικές*. Το σύστημα που χρησιμοποιήσαμε ***C/NC- μέτρο*** στηρίζεται τόσο στην στατιστική προσέγγιση (αμοιβαία πληροφόρηση), με ενσωμάτωση όμως περικειμενικής πληροφορίας (πληροφορία από το άμεσο περιβάλλον του όρου) καθώς και γλωσσολογικής πληροφορίας (συντακτικός χαρακτηρισμός λέξεων, εφαρμογή γλωσσολογικών φίλτρων που ορίζουν και περιορίζουν τους τύπους των εξαγομένων όρων (γίνονται 'δεκτά' ουσιαστικά, επίθετα, προθέσεις, κατηγορίες στις οποίες ανήκουν οι περισσότεροι όροι). Το μέτρο ***C/NC*** λαμβάνει υπ' όψιν του την εσωτερική δομή των όρων. Στόχος είναι η εξαγωγή ένθετων (nested) όρων που συνδέεται στενά με τους πολυλεκτικούς όρους. [Frantzi & Ananiadou 1999, Maynard & Ananiadou, 2000].

1. ΟΡΙΣΜΟΣ ΠΟΛΥΛΕΚΤΙΚΩΝ ΟΡΩΝ

Οι όροι πραγματώνουν γλωσσικά τις έννοιες ενός επιστημονικού τομέα, με άλλα λόγια δηλώνουν (αποτελούν την δήλωση) μια(ς) συγκεκριμένη(ς) έννοια(ς) μιας ειδικής γλώσσας μέσω μιας γλωσσικής έκφρασης. Είναι δυνατή η διάκριση μεταξύ *όρου* (*term*) και *μορφής του όρου* (*termform*). Ο *όρος* δεν είναι παρά η ένωση/ το σημείο τομής δύο συνόλων, ενός *εννοιολογικού* (το ορισμένο σημασιολογικό περιεχόμενο) και ενός *γλωσσικού* (η έκφραση, ή η μορφή του όρου) [Lauriston, 1994]. Από την παραπάνω διάκριση συνεπάγεται ότι ενώ ένας όρος δεν μπορεί να είναι πολύσημος, οι διάφορες μορφές του (γλωσσολογικές πραγματώσεις του) μπορούν να έχουν περισσότερες σημασίες. Η διάκριση αυτή, αν και πολύ σημαντική δεν θα μας απασχολήσει περαιτέρω, αφού είναι σημαντική όταν η αναγνώριση όρων συνδέεται με την σημασιολογική ανάλυσή τους, δηλαδή με την σύνδεσή τους με τις έννοιες του τομέα. Ωστόσο, εφόσον η προσέγγιση μας δεν είναι σημασιολογική, οι δυο όροι (*όρος, μορφή του όρου*) θα χρησιμοποιηθούν στο εξής ως συνώνυμοι.

Οι περισσότεροι όροι εμφανίζουν σύνθετη μορφή, αποτελούνται δηλαδή από περισσότερες από μια λέξεις. Οι πολυλεκτικοί ή σύνθετοι ή συνταγματικοί όροι (όροι – συντάγματα) θεωρούνται ως 'οι προτιμώμενες μονάδες / οντότητες των όρων-εννοιών' [Sager, 1990]. Αν και οι πολυλεκτικοί όροι συνήθως εκφράζουν σύνθετες σχέσεις μεταξύ εννοιών, η σχέση μεταξύ της πολυπλοκότητας μιας έννοιας και του μήκους του αντίστοιχου όρου δεν είναι πάνα αυτονόητη/ αυτόματη/ δεδομένη.

Ο ορισμός του πολυλεκτικού όρου και κατά συνέπεια η εξαγωγή του από ένα σώμα δεδομένων δεν είναι ούτε απλές ούτε απολύτως αποσαφηνισμένες διαδικασίες. Ωστόσο, όσον αφορά στην **δομή** των πολυλεκτικών όρων έχει παρατηρηθεί ότι οι περισσότεροι έχουν ως κεφαλή είτε:

α) ένα *όνομα ουσιαστικό* (αυτή είναι και η συνηθέστερη κατηγορία). Επιπλέον οι όροι μπορεί να περιλαμβάνουν:

α) μόνον ουσιαστικά, π.χ. *bcl-2 protein level, head squamous cell carcinoma patients*

β) επίθετα και ουσιαστικά, π.χ. *cellular retinoic acid, synthetic agonistic retinoids*

γ) άλλες λεξικές κατηγορίες, όπως μετοχές ενεστώτα π.χ. *ligand binding studies* ή αορίστου, π.χ. *naked DNA template*, αριθμούς, π.χ. *bcl-2 RNA level, occupied 9 S ER,* επιρρήματα, π.χ. *early-stage cancer,* προθετικές φράσεις e*xpression of PCAF.*

δ) άλλα στοιχεία: ελληνικές λέξεις ή χαρακτήρες: *RAR **alpha** messenger protein level, plasma 17 **beta**-E concentration,* ακρωνύμια, π.χ. *DNA, IGF-binding proteins,* άλλα σημεία, όπως παύλες, παρενθέσεις, κλπ, π.χ. *all-trans retinoic acid, RAR/RXR heterodimers,* κόμματα *immuno-, steroid-, site-specific DNA-affinity chromatography,* παρατακτικούς συνδέσμους, π.χ *Fos and Jun expression vectors*. Οι τελευταίοι μάλιστα συνδέονται στενά με την παρατακτική σύνδεση των όρων (Coordination). Και βέβαια είναι δυνατή ο συνδυασμός/ συνεμφάνιση περισσότερων του ενός χαρακτηριστικών.

β) ένα *επίθετο* που συνήθως προηγείται του ουσιαστικού (-ων), π.χ. *heat resistant, light sensitive*. Στις περιπτώσεις αυτές, ωστόσο, δεν είναι απολύτως σαφές αν ο 'επιθετικός' πολυλεκτικός όρος αποτελεί όντως όρο ή μέρος ευρύτερου/ μεγαλύτερου όρου.

2. ΑΝΑΓΝΩΡΙΣΗ ΠΟΛΥΛΕΚΤΙΚΩΝ ΟΡΩΝ ΚΑΙ ΣΧΕΤΙΚΕΣ ΔΥΣΚΟΛΙΕΣ

Ένα από τα βασικότερα προβλήματα που άμεσα συνδέονται με τους πολυλεκτικούς όρους είναι η **μεταβλητότητα**, οι **αποκλίσεις** (**variations**) που παρατηρούνται στην χρήση των όρων αυτών. Οι διάφορες μορφές της ίδιας έννοιας, οι τρόποι δηλαδή πραγμάτωσης και έκφρασής της, που είναι δυνατόν να διαφέρουν λιγότερο ή περισσότερο, αποτελούν τις αποκλίσεις του όρου/ έννοιας. Το φαινόμενο αυτό, κοινό σε όλους σχεδόν τους επιστημονικούς τομείς, είναι ιδιαιτέρως σύνηθες στον χώρο της Βιολογίας και μάλιστα της Μοριακής Βιολογίας (Molecular Biology) σε κείμενα της οποίας εστιάζεται η έρευνα μας. Οι λόγοι για την εμφάνιση της μεταβλητότητας είναι πολλοί και διαφορετικής υφής. Η συνεχής εξέλιξη της επιστήμης, ιδιαίτερα μετά την αποκρυπτογράφηση του DNA, η πιεστική ανάγκη να 'ονομαστούν' οι νέες έννοιες, τα νέα γονιδιακά προϊόντα που συνεχώς ανακαλύπτονται, τα ασαφή όρια μεταξύ των όρων αποτελούν μερικά από τα αιτία εμφάνισης της μεταβλητότητας. Σ' αυτά πρέπει να προστεθεί η χρήση και αναφορά στις έννοιες αυτές από μεγάλο αριθμό ερευνητών από διαφορετικές εθνικότητες και τομείς που με την σειρά τους καθιστούν δυσκολότερη την τυποποίηση και την ομοφωνία στην χρήση της ορολογίας ενισχύοντας κατά συνέπεια την ποικιλομορφία των όρων. Εξάλλου δεν είναι σπάνιο φαινόμενο η χρήση παρόμοιων

όρων οι οποίοι ωστόσο αναφέρονται στη ίδια έννοια από έναν επιστήμονα και μάλιστα μέσα στο ίδιο κείμενο. Αυτό συμβαίνει κυρίως στην περίπτωση των πολυλεκτικών όρων, όπου είναι δυνατόν μετά από μια πρώτη - και πλήρη - αναφορά στην έννοια, να επιλέγεται η μετέπειτα αναφορά σε αυτήν με μια ή και περισσότερες διαφοροποιημένες μορφές της, π.χ. συντετμημένη, με χρήση μόνο των αρχικών των λέξεων που αποτελούν τον όρο, κλπ. Αυτή ακριβώς η ποικιλομορφία αποτελεί το σημείο αναφοράς μας, αφού η διαφοροποίηση των όρων φαίνεται να είναι αποτέλεσμα κάποιας γλωσσικής διαδικασίας (είτε μορφολογικής, είτε συντακτικής, είτε σημασιολογικής). Η ποικιλομορφία ως απόρροια διαφορετικών γλωσσολογικών διαδικασιών και μορφών παρουσιάζει έτσι ενδιαφέρον για τον γλωσσολόγο. Μετά από μια σύντομη αναφορά στα αίτια της μεταβλητότητας των όρων, ας δούμε λοιπόν αναλυτικότερα τους τρόπους που γλωσσικά πραγματώνονται οι αποκλίσεις αυτές. Έτσι λοιπόν, οι πολλαπλές μορφές της ίδιας έννοιας απορρέουν είτε από [Jacquemin, 1999]:

1) μορφολογικά συγγενείς τύπους, π.χ.

ενικός- πληθυντικός: *Biochemical study - biochemical studies*
παράγωγος τύπος: *grammar category - grammatical category*
έλλειψη τυποποίησης: *all-trans retinoic acid - all trans retinoic acid*

2) συντακτικά συγγενείς τύπους, π.χ. προσδιορισμός: *human clones - human cDNA clones*
παρατακτική σύνδεση *human ribophorin I – human ribophorins I and II*
προθετική φράση σε θέση επιθετικού προσδιορισμού:
SF-1 transcriptional activity – transcriptional activity of SF-1

3) σημασιολογικά συγγενείς τύποι, π.χ. *rigid lenses – hard lenses*

4) συνδυασμό των 1-2-3, π.χ. μορφοσυντακτικά συγγενείς τύποι: *gene located on – gene location.*

Επομένως, ένα πλήθος μορφολογικών και συντακτικών φαινομένων μπορούν δυνητικά να προκαλέσουν επιπλοκές και να θέσουν εμπόδια στην αναγνώριση των όρων. Τα προβλήματα ωστόσο δεν σταματούν εδώ για την αυτόματη αναγνώριση των όρων, αν ληφθεί υπ' όψιν ότι η χρήση των διαφόρων μορφών και η ταύτιση τους με την ίδια έννοια συνιστά ιδιαίτερα πολύπλοκη και επίπονη διαδικασία στα πλαίσια των υπολογιστικών συστημάτων δεδομένων των πάντα πεπερασμένων δυνατοτήτων των Η/Υ τα όρια των οποίων είναι περιορισμένα σε σχέση με τις δυνατότητες της φυσικής γλώσσας. Με τα παραπάνω άμεσα συνδέεται η περιορισμένη χρήση

σημασιολογικής ανάλυσης που σε μεγάλο βαθμό θα επέτρεπε την προσπέλαση των κείμενων και κατά συνέπεια την ταύτιση των διαφορετικών μορφών ενός όρου.

3. ΥΛΟΠΟΙΗΣΗ

Η περιγραφή των πολυλεκτικών ορών και η - εν συντομία - αναφορά στα προβλήματα που συνδέονται με την χρήση τους και κατά συνέπεια με την αναγνώριση και την αυτόματη εξαγωγή τους, καθιστούν –πιστεύουμε - εμφανή την ανάγκη οι όροι αυτοί να μην εξάγονται απλώς από τα κείμενα, αλλά -τουλάχιστον- να επιχειρείται η ***συγχώνευση*** (***conflation***) των (μορφολογικά, συντακτικά ή σημασιολογικά) συγγενών όρων.

Για τον λόγο αυτό χρησιμοποιήσαμε ένα εργαλείο Υπολογιστικής Γλωσσολογίας που στοχεύει τόσο στην εξαγωγή των πολυσύνθετων όρων, όσο και στην συγχώνευση, την σύνδεση δηλαδή, συγγενών όρων, το ***FASTR.***

3. 1. ΠΕΡΙΓΡΑΦΗ ΤΟΥ FASTR

Όπως ήδη αναφέραμε, το ***FASTR*** [Jacquemin et al., 1999] είναι ένα εργαλείο Υπολογιστικής Γλωσσολογίας το οποίο κάνει εξαγωγή πολυσύνθετων όρων και κατόπιν συγχώνευση των όρων αυτών με άλλους όρους οι οποίοι απαντώνται στα κείμενα και αποτελούν αποκλίσεις των πρώτων βάσει των κοινών χαρακτηριστικών τους.

Το ***FASTR*** αποτελείται από:

1. έναν tagger για την αναγνώριση των μερών του λόγου όλων των λέξεων του κειμένου
2. ένα σύστημα μορφολογικής ανάλυσης (Finite State Automaton FSA) που στοχεύει τόσο στην κλίση όσο και στην παραγωγή
3. μια λίστα πολυσύνθετων όρων των οποίων η δομή δίδεται με την μορφή συντακτικών κανόνων/ σχημάτων
4. έναν parser (συντακτικό αναλυτή) που αποτελείται από *μετακανόνες (metarules)*, όπως χαρακτηριστικά τους ορίζει ο Jacquemin, οι οποίοι περιγράφουν τους μορφοσυντακτικούς μετασχηματισμούς.

3. 2. ΤΥΠΟΙ ΠΑΡΑΛΛΑΓΩΝ

Οι μετακανόνες μπορούν να διακριθούν ανάλογα με τους τύπους των παραλλαγών που περιγράφουν:

1. παραλλαγές όρων – αποτέλεσμα *κλιτικών* ή *συντακτικών* διαδικασιών
2. παραλλαγές/ αποκλίσεις όρων - απόρροια πιο σύνθετων μετασχηματισμών *μορφοσυντακτικής* υφής, όταν δηλαδή οι δύο όροι ή - ορθότερα - οι παραλλαγές ενός όρου

/ έννοιας έχουν προέλθει από έναν συνδυασμό μορφολογικών και συντακτικών διαδικασιών, όπως είναι, για παράδειγμα, οι παραγωγικοί μετασχηματισμοί, π.χ. *gene located on – gene location.*

3. 2. 1. ΣΥΜΠΛΕΚΤΙΚΗ ΣΥΝΔΕΣΗ ΟΡΩΝ

Το είδος των δομών που κυρίως μας απασχόλησαν είναι όροι που συνδέονται κατά παράταξη με συμπλεκτικό ή διαχωριστικό σύνδεσμο (*και* ή ή) ή ίσως κόμμα (,) και ανήκουν στο πρώτο είδος παραλλαγών, αφού είναι αποτέλεσμα αποκλειστικά συντακτικών διαδικασιών. Ας δούμε κάποια παραδείγματα από τα κείμενα της έρευνάς μας:

cellular retinoic acid binding proteins I and II,
RAR alpha messenger RNA and protein levels,
cysteine and arginine residues,
immuno-, steroid-, and site-specific DNA-affinity chromatography.
naked DNA or cromatin template
LNCaP and COS-1 cell lines
human neuroblastoma and melanoma cell lines

Για την πιο επαρκή αντιμετώπιση του φαινομένου διακρίναμε **δύο** επιμέρους **σχήματα σύζευξης**:

1. Σύνδεση των προσδιορισμών / των συμπληρωμάτων ενώ η κεφαλή είναι κοινή και για τους δύο όρους. Είναι ωστόσο απαραίτητες ορισμένες διευκρινίσεις – και ανάλυση – για τις δομές, αφού και μέσα στην κάθε κατηγορία εμφανίζονται περαιτέρω διαφοροποιήσεις. Ας δούμε ορισμένα σχετικά παραδείγματα:

   naked DNA or cromatin template → naked DNA template
   → cromatin template

   Εδώ η κεφαλή που βρίσκεται στο τέλος (*template*) συνοδεύεται από διαφορετικό αριθμό προσδιορισμών στους δύο συνδεόμενους όρους.

   Ενώ στο παράδειγμα που ακολουθεί η κεφαλή είναι ήδη ένας πολυλεκτικός τύπος (*cell lines*).

   LNCaP and COS-1 cell lines → LNCaP cell lines
   → COS-1 cell lines

   Από την άλλη το παράδειγμα που ακολουθεί, και ανήκει επίσης στην ίδια ευρύτερη κατηγορία, είναι το πιο πολύπλοκο από όλα. Ορισμένα από τα συμπληρώματα / προσδιορισμούς είναι κοινά (*human)* και κάποια όχι (*neuroblastoma - melanoma*).

human neuroblastoma and melanoma cell lines → human neuroblastoma cell lines
→ human melanoma cell lines

Ίσως στην περίπτωση αυτή να παίζει ρόλο το είδος των συμπληρωμάτων: ο επιθετικός προσδιορισμός (*human*) είναι ίσως λιγότερο στενά 'δεμένος' με την κεφαλή σε σχέση με τα ουσιαστικά – συμπληρώματα (*neuroblastoma - melanoma*).

2. Στο είδος αυτό οι προσδιορισμοί μένουν κοινοί ενώ συνδέονται οι *κεφαλές,* οι 'πυρήνες' των δύο όρων. Και στην περίπτωση αυτή ωστόσο είναι απαραίτητος ο διαχωρισμός επιμέρους περιπτώσεων. Ας παρακολουθήσουμε δύο χαρακτηριστικά παραδείγματα:

Cellular retinoic acid binding proteins I and II, immuno-, steroid-, and site-specific DNA-affinity chromatography

Παρατηρούμε ότι στην πρώτη περίπτωση είναι μάλλον ασαφές ποια είναι ακριβώς η κεφαλή του όρου: οι λατινικοί αριθμοί οι οποίοι και άμεσα συνδέονται συμπλεκτικά (*I, II*) ή ίσως μια σωστότερη ανάλυση θα συμπεριλάμβανε στις κεφαλές και την λέξη *proteins* που εξάλλου, σημασιολογικά, αποτελεί με τους λατινικούς αριθμούς μία ενότητα;

Το δεύτερο παράδειγμα αποτελεί χαρακτηριστικό παράδειγμα των δυσκολιών με τις οποίες οι ερευνητές έρχονται αντιμέτωποι στην προσπάθεια αυτόματης αναγνώρισης των όρων επιστημονικών ή τεχνολογικών τομέων κοινό χαρακτηριστικό των οποίων είναι οι πολυσύνθετες γλωσσικές δομές με τις οποίες πραγματώνονται. Πιο συγκεκριμένα, εδώ το πολυσύνδετο σχήμα θέτει επιπλέον δυσκολίες στην περιγραφή, και κατά συνέπεια στην μεταγραφή του στον φορμαλισμό του προγράμματος, λόγω της ιδιαίτερης μορφής των όρων. Οι παύλες δηλώνουν τα όρια του κάθε όρου, καθιστώντας τα ωστόσο δυσδιάκριτα, αφού ο ρόλος τους είναι διττός: διαχωρίζουν τα συνθετικά του κάθε όρου, ενώ συγχρόνως συντελούν στην κατά παράταξη σύνδεση των πολυσύνθετων όρων. Υπάρχει 'κίνδυνος' να οριστεί ως β' συνθετικό το *specific- DNA -affinity chromatography*, αντί για το ορθό *affinity chromatography.*

Αυτά είναι μερικά μόνο από τα προβλήματα που απορρέουν από τους πολυσύνθετους όρους και την προσπάθεια αναγνώρισης τους στα κείμενα. Σε μεγάλο βαθμό τα προβλήματα περιγραφής πρέπει να αντιμετωπιστούν σε αρχικό στάδιο, δηλαδή κατά την εισαγωγή λέξεων στον μορφολογικό αναλυτή (tagger) και τον χαρακτηρισμό τους. Στην προαναφερθείσα περίπτωση, για παράδειγμα, θα πρέπει οι όροι με την παύλα να αποτελούν ξεχωριστά λήμματα (π.χ. *immuno-*) ή χρειάζεται πιο αναλυτική μορφολογική-ορθογραφική πληροφορία;

3. 3. ΜΕΤΑΓΡΑΦΗ ΤΩΝ ΔΟΜΩΝ ΣΤΟΝ ΦΟΡΜΑΛΙΣΜΟ ΤΟΥ FASTR

3. 3. 1. ΣΥΝΔΕΣΗ ΠΡΟΣΔΙΟΡΙΣΜΩΝ

Ας δούμε πώς τα προαναφερθέντα φαινόμενα περιγράφονται με την τυπολογία του ***FASTR***. Ας πάρουμε για παράδειγμα τον όρο *LNCaP and COS-1 cell lines → LNCaP cell lines*

→ COS-1 cell lines

Ως αρχική δομή, από την οποία θα παραχθεί ο πολυλεκτικός όρος με τον συμπλεκτικό σύνδεσμο και ο οποίος κατά συνέπεια αποτελεί απόκλιση του 'βασικού' όρου, ορίζεται εκείνη του όρου *LNCaP cell lines*. Ο όρος αυτός βρίσκεται στην λίστα των πολυλεκτικών όρων η οποία, όπως έχει ήδη αναφερθεί, είναι απαραίτητη για την αναγνώριση των άλλων πολυλεκτικών όρων που συνδέονται με τους πρώτους. Ο όρος μεταγράφεται ως εξής στο ***FASTR***:

$N_1 \rightarrow N_2\ N_3\ N_4$

Η δομή που απορρέει από την σύνδεση των όρων περιγράφεται με τον εξής κανόνα

$^{1}X_1 \rightarrow N_2\ (C\ N)\ N_3\ N_4$

Όπως φαίνεται, η σύνδεση (δηλαδή ο σύνδεσμος και το άλλο όνομα – ουσιαστικό) δίνονται στην παρένθεση και παρεισφρύουν στην αρχική δομή έχοντας ως αποτέλεσμα την δημιουργία ενός νέου τύπου που ορίζεται ως X_1, δηλαδή ως μεταβλητή που μπορεί να πάρει οποιαδήποτε τιμή, αφού το μέρος του λόγου του νέου τύπου δεν μας ενδιαφέρει.

Ας περάσουμε σε ένα πιο πολύπλοκο παράδειγμα με σύνδεση περισσότερων των δύο στοιχείων που για την σύνδεσή τους γίνεται χρήση τόσο συμπλεκτικού συνδέσμου όσο και ασύνδετου σχήματος (*κόμμα -,-).*

both RAR alpha, RAR beta and RAR gamma mRNAs → RAR alpha mRNA

→ RAR beta mRNA

→ RAR gamma mRNA

Ως μορφή του αρχικού πολυλεκτικού τύπου ορίζεται η εξής:

$N_1 \rightarrow N_2\ N_3\ N_4$

ενώ η πολυσύνθετη δομή μεταγράφεται ως εξής:

$X_1 \rightarrow N_2\ N_3\ (^{2}\text{Punc N C N})\ N_4$

3. 3. 2. ΣΥΝΔΕΣΗ ΚΕΦΑΛΩΝ

[1] Το X αναφέρεται σε κάθε είδους φράση (συνήθως βέβαια πρόκειται για ΟΦ), το N(nominal) αναφέρεται σε κάθε όρο που έχει λειτουργία ονόματος, ενώ το C αναφέρεται στον συμπλεκτικό σύνδεσμο (coordinating conjunction)

[2] Punc(punctuation) σημαίνει στίξη και συνήθως στις δομές που μελετώνται είναι το κόμμα (,).

Σε αυτό το είδος σύνδεσης οι προσδιορισμοί μένουν κοινοί, ενώ συνδέονται οι κεφαλές, οι 'πυρήνες' των δύο όρων. Ας παρακολουθήσουμε κάποια παραδείγματα και την μεταγραφή τους σε 'κανόνες' του ***FASTR*** που καταδεικνύουν διάφορα προβλήματα γλωσσολογικής υφής.

Στο παράδειγμα που ακολουθεί φαινομενικά συνδέονται μόνο μέρη της κεφαλής (οι λατινικοί αριθμοί *I* και *II*) ενώ όλη η κεφαλή είναι ένας πολυλεκτικός τύπος.

Cellular retinoic acid binding proteins I and II → Cellular retinoic acid proteins I

→ Cellular retinoic acid proteins II

Να πώς μεταγράφονται οι όροι στο ***FASTR***:

$N_1 \rightarrow A_2\ A_3\ N_4\ A_5\ N_{6+7}$

Το ουσιαστικό N_{6+7} δηλώνει ακριβώς την στενή σύνδεση του ουσιαστικού (*proteins*) με τον λατινικό αριθμό (*I*). Στον τύπο αυτόν προστίθενται ο συμπλεκτικός σύνδεσμος και ένα ακόμη ουσιαστικό (ο λατινικός αριθμός *II*)

$X_1 \rightarrow A_2\ A_3\ N_4\ A_5\ N_{6+7}$ (C N)

3. 4. ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΜΕΤΑΚΑΝΟΝΩΝ

Μετά την σύντομη περιγραφή κάποιων δομών και την νύξη ορισμένων από τα προβλήματα που απορρέουν, ας δούμε τους *μετακανόνες* οι οποίοι ορίζουν και περιγράφουν τους γλωσσικούς μετασχηματισμούς από τους οποίους απορρέουν οι παραλλαγές των τύπων. Οι κανόνες δίνονται με την μορφή **τυπικών εκφράσεων** (**regular expressions**).

Metarule Coor ($X_1 \rightarrow X_2\ N_3$) =

$X_1 \rightarrow X_2$ $PUNC_4$ <{A|N|Np|V}> $PUNC_5$ <{A|N|Np|V} PUNC? >

C6 <{A|N|Np|V}> $N_{3:}$

<X_2 num >! plu.

Ο κανόνας για την **σύνδεση των προσδιορισμών** δηλώνει ότι σε μια τέτοια δομή ο τύπος που συνδέεται κατά παράταξη με τον αρχικό τύπο - ή αλλιώς τον **τύπο αναφοράς (reference type)** - με κεφαλή το N_3 προστίθεται ή παρεισφρύει στον αρχικό τύπο. X_2 είναι μια μεταβλητή για τους προσδιορισμούς που προηγούνται του όρου – κεφαλής. Δεν επιδρούν στην μορφή του δεύτερου - παραγόμενου – όρου, οπότε δεν μας ενδιαφέρουν οι τιμές που μπορεί να πάρει. Ωστόσο στο τέλος του κανόνα δίδεται ένας **μορφολογικός / γραμματικός περιορισμός** (grammatical constraint): αυτό/ - ά τα γλωσσικά στοιχεία δεν μπορεί να είναι πληθυντικού αριθμού. Ο κανόνας περιορισμός ισχύει για την αγγλική γλώσσα, όπου πολλά ουσιαστικά (όχι όμως στον πληθυντικό) αποτελούν προσδιορισμούς άλλων ονομάτων. Ο περιορισμός αυτός απορρίπτει την αποδοχή δομών όπως

tumors and/ or K562 cells όπου έχουμε δύο διαφορετικούς όρους (*tumors, K562 cells)* που συνδέονται παρατακτικά και όχι σύνδεση προσδιορισμών μιας κοινής κεφαλής.
Ιδιαίτερο ενδιαφέρον έχει η 2η γραμμή του κανόνα όπου δίδεται η μορφή του τύπου – απόρροια του συντακτικού μετασχηματισμού: ενός τριγωνικών αγκυλών δίδεται η δομή που 'παρεισφρύει' στον αρχικό όρο, με την μορφή τυπικής έκφρασης (regular expression) ώστε να είναι όσον το δυνατόν πιο γενικευμένη και απλουστευμένη: ο σύνδεσμός μπορεί να ακολουθείται από ένα επίθετο (A), ουσιαστικό (N), ή ολόκληρη ΟΦ (NP) ή ρήμα ή κάποιο συνδυασμό τους (δηλώνεται με τις αγκύλες {}) 0 – 3 φορές.

Παρόμοιος είναι και ο κανόνας για την **σύνδεση κεφαλών**. Εδώ ενδεικτικά δίνουμε τον κανόνα που περιγράφει πιο πολύπλοκη δομή, εκείνη της σύνδεσης περισσότερων των δύο όρων με την εισαγωγή σημείων στίξης, π.χ. κόμμα (,). Το ερωτηματικό (?)δηλώνει την μη υποχρεωτική παρουσία του στοιχείου του οποίου έπεται.

Metarule Coor (X_1→ X_2 N_3) =
X_1→ X_2 $PUNC_4$ <{A|N|Np|V}> $PUNC_5$ <{A|N|Np|V} PUNC?>
C6 <{A|N|Np|V}> N_3:
<X_2 num >! plu.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ – ΜΕΛΛΟΝΤΙΚΕΣ ΕΦΑΡΜΟΓΕΣ

Εφαρμόσαμε το ***FASTR*** σε ένα σώμα δεδομένων 113.428 λέξεων. Το ***FASTR*** εκτέλεσε 5046 προσπάθειες για μια λίστα 899 δεδομένων όρων. Αναγνωρίστηκαν 488 αποκλίσεις των όρων αυτών (μορφολογικά και συντακτικά). Μετά από μία ταξινόμηση των αποτελεσμάτων πήραμε τα εξής αποτελέσματα: 132 από τους 899 δεδομένους τύπους βρέθηκαν να έχουν 259 αποκλίσεις. Επομένως 14% των όρων είχαν αποκλίσεις. Οι πιο ενδιαφέρουσες αποκλίσεις αφορούν το συνταγματικό επίπεδο, στο οποίο ανήκει η σύνδεση κατά παράταξη. Στις 63 πρώτες αποκλίσεις βρέθηκαν 5 δομές με συμπλεκτική σύνδεση. Ο πίνακας που ακολουθεί δίνει ένα παράδειγμα των εξαγομένων όρων από το ***FASTR***.

AhR heterodimer	*AhR and Arnt heterodimer*	*COOR*
estrogen receptor	*estrogen, glucocorticoid, and*	*COOR*

	progesterone receptors	

Ασχοληθήκαμε με ορισμένα από τα προβλήματα της **μεταβλητότητας των όρων** (**Term Variation**) στον τομέα της Μοριακής Βιολογίας. Αναλύσαμε ορισμένες από τις μορφές με τις οποίες οι πραγματώνονται οι παραλλαγές πολυλεκτικών όρων (συμπλεκτική σύνδεση) χρησιμοποιώντας το ***FASTR***. Βεβαίως παραμένουν ακόμη και άλλα είδη μετασχηματισμών, τόσο *συντακτικής – μορφολογικής* όσο και *σημασιολογικής υφής*. Για την τυποποίηση αυτής της τελευταίας κατηγορίας (σημασιολογικές αποκλίσεις) είναι απαραίτητη η χρήση **θησαυρών (thesauri)** και άλλων ηλεκτρονικών πηγών πληροφοριών (resources).

Απώτερος στόχος μας είναι η εφαρμογή, η ενσωμάτωση των αποτελεσμάτων του ***FASTR*** στο πρόγραμμα εξαγωγής όρων που χρησιμοποιούμε (C/ NC - μέτρο), με σκοπό την βελτίωση των αποτελεσμάτων, δηλαδή των όρων που εξάγουμε από τα κείμενα. Αν οι μορφολογικά - συντακτικά – σημασιολογικά συγγενείς μορφές λαμβάνονται ως τέτοιες, δηλαδή παρέχεται η δυνατότητα σύνδεσής τους, τότε τα αποτελέσματα αυτόματης εξαγωγής όρων θα είναι πολύ πιο ακριβή, αφού συγγενείς όροι (παραλλαγές του ίδιου όρου) θα συνδέονται, δεν θα εξάγονται ως εντελώς διαφορετικοί, με αποτέλεσμα την μετατροπή του ποσοστού εμφάνισής τους.

Απώτερος στόχος είναι και η αυτόματη δημιουργία θησαυρών για τους όρους των κειμένων που θα επιτρέπει την άμεση και εύκολη διεπαφή του χρήστη με κείμενα από σώματα δεδομένων των διαφόρων επιστημονικών τομέων.

5. ΒΙΒΛΙΟΓΡΑΦΙΑ

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

Jacquemin, C. (1999). Syntagmatic and paradigmatic representations of term variation, in Proc. of *37th Annual Meeting of the Association for Computational Linguistics (ACL'99),* University of Maryland, pages 341-348

Jacquemin, C., et Tzoukermann, E. (1999). NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax, in T. Strzalkowski, editor, *Natural Language Information Retrieval,* pages 25-74, Kluwer, Boston, MA.

Lauriston, A. (1994) Automatic recognition of complex terms: problems and the TERMINO solution. Terminology, 1, 1:147-170.

Maynard, D., Ananiadou, S. (2000) Identifying Terms by their Family and Friends, in Proceedings of *the 18th International Conference on Computational Linguistics, COLING 2000*: 530-536, Saarbrucken, Germany.

Sager, J. (1990) A Practical Course in Terminology Processing. John Benjamins.

Katerina Kehagia and **Sophia Ananiadou**

Computer Science, School of Sciences
University of Salford, Salford M5 4WT U.K.

K.Kehagia@pgr.salford.ac.uk, S. Ananiadou@salford.ac.uk
telephone: +44.161.295.0480
Computer Science, School of Science
Newton Building
University of Salford
Salford M5 4WT
UK