ΕΛΕΤΟ – 1° Συνέδριο «Ελληνική Γλώσσα και Ορολογία»
Αθήνα, 30, 31 Οκτωβρίου και 1 Νοεμβρίου 1997

ΜΕΘΟΔΟΣ ΗΜΙ-ΑΥΤΟΜΑΤΗΣ ΕΞΑΓΩΓΗΣ ΟΡΩΝ

Μ. Γαβριηλίδου και Π. Λαμπροπούλου

ΠΕΡΙΛΗΨΗ

Αντικείμενο της παρούσας ανακοίνωσης είναι μία μέθοδος ημι-αυτόματης εξαγωγής όρων από κείμενα. Στόχος είναι ο εντοπισμός λέξεων και πολυλεκτικών σχηματισμών οι οποίοι, με βάση συγκεκριμένες ενδείξεις, έχουν υψηλή πιθανότητα να είναι όροι. Η τελική επικύρωση (ή όχι) των υποψήφιων όρων γίνεται από τον ειδικό του γνωστικού τομέα. Η μέθοδος αυτή περιλαμβάνει δύο στάδια : καταρχήν, την επεξεργασία ειδικών κειμένων με ***γλωσσικά*** και ***στατιστικά εργαλεία,*** των οποίων η λειτουργία είναι η άντληση γνώσης από τα κείμενα και η ταξινόμηση και μέτρηση των δεδομένων, και στη συνέχεια, την εφαρμογή ***στατιστικο-γραμματικών φίλτρων*** για την ανάδειξη των υποψήφιων όρων.

ABSTRACT

M. Gavrilidou and P. Labropoulou : A METHOD FOR SEMI-AUTOMATIC TERM EXTRACTION

In this paper a method for corpus-based semi-automatic term extraction is presented. *The* aim of this method is to locate words and multi-word constructions, which, on the basis of specific indications, present a high probability of termhood. *The* knowledge domain specialist is responsible for the validation (or not) of candidate terms. *This* method entails two stages : initially, the ***linguistic*** and ***statistical processing*** of specialised texts, based on tools whose function is the extraction of knowledge from texts and the classification and quantification of the data, and, secondly, the application of ***statistico-grammatical filters*** on these data yielding the candidate terms.

0. ΕΙΣΑΓΩΓΗ

Στην παρούσα ανακοίνωση παρουσιάζεται μία μέθοδος **ημι-αυτόματης εξαγωγής όρων** από επιστημονικά ή τεχνικά κείμενα, η οποία βασίζεται σε συνδυασμό **γλωσσικών και στατιστικών κριτηρίων**.

Η μέθοδος χαρακτηρίζεται ως ημι-αυτόματη δεδομένου ότι τελικός στόχος είναι η κατάρτιση ενός επεξεργασμένου καταλόγου με υποψήφιους όρους, ο οποίος τίθεται υπόψη των ειδικών για επισκόπηση και έγκριση. Ο κατάλογος αυτός είναι το αποτέλεσμα εφαρμογής υπολογιστικών γλωσσικών και στατιστικών εργαλείων σε ειδικό σώμα κειμένων (specialised corpus) ενός συγκεκριμένου τομέα.

Παράλληλα προς την παραδοσιακή μέθοδο δημιουργίας ορολογικών πόρων, δηλαδή την αποδελτίωση, συλλογή και καταγραφή όρων από ειδικούς επιστήμονες, οι μέθοδοι που

συνήθως χρησιμοποιούνται για την αυτόματη εξαγωγή όρων από κείμενα εμπίπτουν σε δύο βασικές κατηγορίες[1] :

- χρήση ***μορφο-συντακτικής ή επιφανειακής συντακτικής ανάλυσης*** (1), και
- χρήση ***στατιστικών μετρήσεων***, κυρίως για την αναγνώριση πολυλεκτικών όρων (2, 3).

Η μέθοδος που προτείνουμε διαφοροποιείται από τις προηγούμενες. Χαρακτηρίζεται ως **υβριδική**, δεδομένου ότι αντλεί στοιχεία και από τις δύο προσεγγίσεις[2].

1. ΟΡΙΣΜΟΣ ΤΟΥ ΑΝΤΙΚΕΙΜΕΝΟΥ

Όροι θεωρούνται οι λεξικές πραγματώσεις (μονολεκτικές ή πολυλεκτικές) των εννοιών ενός επιστημονικού ή τεχνικού τομέα. Οι λέξεις που είναι φορείς σημασίας (ουσιαστικά, επίθετα, ρήματα και επιρρήματα) μπορούν να είναι μονολεκτικοί όροι ή συστατικά πολυλεκτικών όρων. Οι λειτουργικές λέξεις (άρθρα, σύνδεσμοι, προθέσεις, κ.τ.λ.) δηλώνουν σχέσεις μεταξύ των συστατικών αλλά δεν συνεισφέρουν στη σημασία του όρου. Οι συνδυασμοί γραμματικών κατηγοριών με τους οποίους εκφέρονται οι πολυλεκτικοί όροι σχηματίζουν συγκεκριμένες δομές, οι οποίες διαφοροποιούνται ανάλογα με τη γλώσσα, και, πιθανόν, και ανάλογα με την υπογλώσσα.

Η μελέτη μας στο σώμα κειμένων κατέδειξε ότι οι δομές με τις οποίες εκφέρονται οι διλεκτικοί όροι στην ελληνική γλώσσα είναι οι ακόλουθες[3] :

επίθετο + ουσιαστικό	*αγροτικό προϊόν*
ουσιαστικό + ουσιαστικό σε γενική	*φόρος εισοδήματος*
ουσιαστικό + [πρόθ.] + ουσιαστικό	*απαλλαγή [από] φόρο*
ουσιαστικό + ουσιαστικό (ομοιόπτωτα)	*κράτος μέλος*
ρήμα + ουσιαστικό	*τηρώ βιβλία*

Στις ανωτέρω δομές, το ουσιαστικό λειτουργεί ως κεφαλή του συντάγματος, εκτός από την τελευταία δομή, της οποίας κεφαλή είναι το ρήμα. Στην περίπτωση των δομών που αποτελούνται από δύο ουσιαστικά, τον ρόλο της κεφαλής έχει το ουσιαστικό που βρίσκεται πιο αριστερά.

---

[1] Μία συνοπτική επισκόπηση των μεθόδων αυτόματης εξαγωγής όρων αυτόματης αναγνώρισης όρων παρουσιάζεται στο (6).

[2] Μέθοδος με βάση τις ίδιες αρχές, αλλά με έμφαση στις στατιστικές μετρήσεις, έχει προταθεί και από τους Daille et al. (4).

[3] Στις δομές αυτές δεν αναφέρονται οι λειτουργικές λέξεις που πιθανόν παρεμβάλλονται, εκτός εάν κρίνονται βασικές για την κατανοήση της δομής (π.χ. η πρόθεση στην τρίτη δομή).

Σε ό,τι αφορά την τυπολογία των τριλεκτικών και τετραλεκτικών όρων, παρατηρούνται ελάχιστοι όροι οι οποίοι είναι "εκ γενετής" τριλεκτικοί όροι, ενώ οι περισσότεροι παράγονται με πυρήνα υπάρχοντες διλεκτικούς συνδυασμούς προσαυξημένους με βάση τις εξής διαδικασίες :

- ανάπτυξη προς τα αριστερά
  - ⇒ προσθήκη νέας κεφαλής *έκπτωση φόρου* → *δικαίωμα έκπτωσης φόρου*
  - ⇒ προσθήκη προσδιοριστή *οικονομική υπηρεσία* → *δημόσια οικονομική υπηρεσία*
- παρεμβολή προσδιοριστή *εξαγωγή προϊόντος* → *εξαγωγή αγροτικού προϊόντος*
- συνένωση δύο διλεκτικών όρων *σκάφος αναψυχής + ιδιωτική χρήση* → *σκάφος αναψυχής ιδιωτικής χρήσης*
- παράταξη δύο διλεκτικών όρων *Κώδικας Βιβλίων + Κώδικας Στοιχείων* → *Κώδικας Βιβλίων και Στοιχείων*

Με βάση τις προαναφερθείσες διαδικασίες και την τυπολογία επιτρεπτών δομών των διλεκτικών όρων, οι αντίστοιχες επιτρεπτές δομές για τους τριλεκτικούς όρους είναι :

ουσιαστ. + επίθετο + ουσιαστ. σε γεν.	*εισαγωγή αγροτικού προϊόντος*
επίθ. + επίθ. + ουσιαστ. (ομοιόπτ.)	*Δημόσια Οικονομική Υπηρεσία*
ουσιαστ. + ουσιαστ. σε γεν. + ουσιαστ. σε γεν.	*δικαίωμα έκτπωσης φόρου*
ουσιαστ. + [πρόθ.] + ουσιαστ. + ουσιαστ. σε γεν.	*απαλλαγή [από] φόρο εισοδήματος*
ρήμα + επίθετο + ουσιαστικό	*εισάγω αγροτικά προϊόντα*
ρήμα + ουσιαστικό + ουσιαστικό σε γενική	*παρέχω δικαίωμα έκπτωσης*

Οι επιτρεπτές δομές για τους τετραλεκτικούς όρους ακολουθούν τις προηγούμενες αρχές.

2. ΑΡΧΕΣ ΤΗΣ ΜΕΘΟΔΟΥ

Η μέθοδος που προτείνουμε για την εξαγωγή όρων από κείμενα βασίζεται στην ***ποσοτικοποίηση και οργάνωση των δεδομένων***. Τα δεδομένα μας δεν είναι το ίδιο το κείμενο, αλλά οι λέξεις που περιέχονται σε αυτό.

Η μέθοδος, στο πρώτο στάδιο, επεξεργάζεται τα κείμενα, και επιλέγει, οργανώνει σε καταλόγους και ταξινομεί εκείνες τις λέξεις ή/και τους συνδυασμούς λέξεων, που πληρούν την ελάχιστη συνθήκη της γραμματικής κατηγορίας ή επιτρεπτής δομής στις οποίες μπορεί

να ανήκει ένας όρος. Τα δεδομένα αυτά ποσοτικοποιούνται με βάση στατιστικές μετρήσεις (συχνότητα εμφάνισης ή/και συνεμφάνισης).

Το δεύτερο στάδιο επεξεργάζεται τους καταλόγους που προέκυψαν από το πρώτο στάδιο, αξιολογεί τα δεδομένα με βάση στατιστικο-γραμματικά κριτήρια, και προκρίνει εκείνες τις λέξεις ή/και τους συνδυασμούς λέξεων που παρουσιάζουν μεγάλη πιθανότητα να είναι όροι.

Η χρήση γλωσσικών παραμέτρων δεν συνεπάγεται την εξάρτηση της μεθόδου από μία ορισμένη γλώσσα ή οικογένεια γλωσσών, αλλά απλά την προσαρμογή των γλωσσικών κριτηρίων στις ιδιοσυγκρασίες της εκάστοτε γλώσσας. Στην συγκεκριμένη περίπτωση η μέθοδος εφαρμόστηκε στην ελληνική γλώσσα. Η εφαρμογή σε άλλη γλώσσα απαιτεί τη διατύπωση των κατάλληλων γλωσσικών παραμέτρων.

Η μέθοδος που προτείνεται στην παρούσα ανακοίνωση έχει τις βάσεις της σε αντίστοιχη μέθοδο που αναπτύχθηκε και εφαρμόστηκε (στο πλαίσιο του προγράμματος RENOS) σε σώμα κειμένων φορολογικής νομοθεσίας συνολικού μεγέθους 161.815 λέξεων με στόχο την κατάρτιση καταλόγου υποψήφιων όρων του συγκεκριμένου γνωστικού τομέα. Με βάση την εμπειρία αυτή, η παρούσα μέθοδος συστηματοποιεί τη συσχέτιση στατιστικών και γλωσαικών κριτηρίων με στόχο τον μεγαλύτερο βαθμό αυτοματοποίησης της διαδικασίας εξεύρεσης όρων. Η προσεχής υλοποίηση της μεθόδου θα καταδείξει τον βαθμό αποτελεσματικότητάς της.

3. ΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΩΝ

3.1. ΓΛΩΣΣΙΚΑ ΕΡΓΑΛΕΙΑ

Η συχνότητα μίας λέξης στο κείμενο είναι το άθροισμα των εμφανίσεων όλων των λεκτικών της μορφών. Οικονομικότερη λύση από την άθροιση των μεμονωμένων λεκτικών τύπων αποτελεί η αναγωγή τους σε λήμμα και η μέτρηση της συχνότητας του λήμματος. Για παράδειγμα, το ρήμα *τηρώ* απαντά στο σώμα κειμένων με τους εξής τύπους :

Λεκτικός τύπος	**Συχνότητα εμφάνισης**
τηρούν	20
τηρούσαν	6
τηρήθηκαν	6
τηρεί	4
ΤΗΡΩ	**36**

Δεδομένου ότι μόνον οι λέξεις - φορείς σημασίας είναι υποψήφιοι μονολεκτικοί όροι, οι λειτουργικές λέξεις αποκλείονται από τη στατιστική μέτρηση. Επίσης, όπως προαναφέραμε, μόνον ορισμένες γραμματικές δομές παράγουν όρους, και, συνεπώς, οι συνδυασμοί λέξεων που δεν εμπίπτουν σε αυτές τις κατηγορίες αποκλείονται.

Την ***αναγωγή σε λήμμα*** και ***την αναγνώριση γραμματικής κατηγορίας*** επιτελεί ο Λημματοποιητής - Γραμματικός Χαρακτηριστής που έχει αναπτυχθεί στο ΙΕΛ. Το εργαλείο αυτό δέχεται ως είσοδο ένα κείμενο και, με βάση υπολογιστικό Μορφολογικό Λεξικό, αναγνωρίζει το λήμμα από το οποίο προέρχεται κάθε λεκτικός τύπος του κειμένου, και τον χαρακτηρίζει ως προς τη γραμματική κατηγορία και τα μορφολογικά χαρακτηριστικά. Εάν κάποιος τύπος είναι αμφίσημος, δίνονται όλες οι πιθανές ερμηνείες.

3.2. ΣΤΑΤΙΣΤΙΚΑ ΕΡΓΑΛΕΙΑ

Τα στατιστικά εργαλεία, τα οποία εφαρμόζονται στο αποτέλεσμα του Λημματοποιητή - Γραμματικού Χαρακτηριστή, επιτελούν τις εξής εργασίες :

- ***εξαγωγή*** και ***μέτρηση συχνότητας*** των λεκτικών τύπων του σώματος κειμένων και των λημμάτων στα οποία ανάγονται,
- υπολογισμό της ***αναλογικής συχνότητας*** των λεκτικών τύπων και των λημμάτων ανά 1000 λέξεις,
- ***εξαγωγή*** και ***μέτρηση συνεμφάνισης*** συνδυασμών δύο, τριών και τεσσάρων λέξεων ή λημμάτων. Οι συνδυασμοί αυτοί αναγνωρίζονται είτε σε αυστηρή σειριακή ακολουθία είτε με παρεμβολή άλλων λέξεων. Χωρίς αυτή την πρόβλεψη, ο διλεκτικός συνδυασμός "*απαλλαγή [από το] φόρο*" δεν θα λαμβανόταν υπόψη, καθώς οι λειτουργικές λέξεις απομακρύνουν τα δύο συστατικά κατά δύο θέσεις. Συγκεκριμένα, η αναζήτηση γίνεται σε κυλιόμενο παράθυρο πέντε λέξεων, θεωρώντας ότι δύο λέξεις δεν μπορούν να αποτελούν συστατικά του ίδιου όρου εφόσον απέχουν μεταξύ τους περισσότερο από τρεις λέξεις.

Το αποτέλεσμα των στατιστικών εργαλείων είναι κατάλογοι ταξινομημένοι με βάση τις γραμματικές κατηγορίες των λέξεων και των συνδυασμών τους. Συνεπώς, δημιουργούνται τέσσερις κατάλογοι μονολεκτικών σχηματισμών (ουσιαστικά, επίθετα, ρήματα και επιρρήματα) και όλοι οι επιτρεπτοί συνδυασμοί των τεσσάρων γραμματικών κατηγοριών. Οι κατάλογοι αυτοί περιέχουν τις εξής πληροφορίες :

- οι κατάλογοι των μονολεκτικών σχηματισμών τη συχνότητα εμφάνισης κάθε λεκτικού τύπου και του αντίστοιχου λήμματος και τις αναλογικές συχνότητες,
- οι κατάλογοι των πολυλεκτικών σχηματισμών όλα τα προηγούμενα και την επιπλέον πληροφορία της συχνότητας συνεμφάνισης.

4. ΕΠΕΞΕΡΓΑΣΙΑ ΚΑΤΑΛΟΓΩΝ

Για τον εντοπισμό των πιθανότερων υποψήφιων όρων σε αυτούς τους καταλόγους εφαρμόζονται ***στατιστικο-γραμματικά φίλτρα***. Τα φίλτρα αυτά έχουν διττή λειτουργία :

- να αποκλειστούν λέξεις ή συνδυασμοί λέξεων που έχουν πολύ μικρή πιθανότητα να είναι όροι, και
- να προβληθούν λέξεις ή συνδυασμοί λέξεων με μεγάλη πιθανότητα να είναι όροι.

Ήδη η επιλογή των καταλόγων πολυλεκτικών σχηματισμών που περιέχουν μόνο τις επιτρεπτές δομές αποτελεί ένα πρώτο φίλτρο για τον αποκλεισμό συνδυασμών λέξεων που είναι αδύνατο να είναι όροι. Έτσι, με τη διαδικασία αυτή αποκλείεται ο συνδυασμός "επίθετο + επίθετο" που δεν είναι επιτρεπτή δομή. Για να σχηματιστεί όρος, η δομή αυτή απαιτεί την παρουσία ενός ουσιαστικού-κεφαλής.

Το βασικότερο στατιστικό φίλτρο είναι η ***συχνότητα εμφάνισης***. Η υπόθεση εργασίας μας είναι ότι, καθώς τα ειδικά κείμενα παρουσιάζουν πυκνή συγκέντρωση ορολογίας, η υψηλή συχνότητα ενός λήμματος αποτελεί τη βασική ένδειξη ότι το λήμμα αυτό είναι όρος. Η υπόθεσή μας αυτή επιβεβαιώνεται από μελέτες που συγκρίνουν διάφορες στατιστικές μεθόδους και καταλήγουν υπέρ της χρήσης της συχνότητας (4 και 5).

Έτσι, λέξεις που ανήκουν στη γραμματική κατηγορία των ουσιαστικών και καταλαμβάνουν τις υψηλότερες θέσεις στον κατάλογο έχουν μεγάλη πιθανότητα να είναι είτε μονολεκτικοί όροι είτε συστατικά πολυλεκτικών όρων. Ανάλογα, επίθετα με υψηλή συχνότητα εμφάνισης είναι πολύ πιθανό να είναι συστατικά πολυλεκτικών όρων που δρουν ως προσδιοριστές.

Για τον αποκλεισμό όσο το δυνατόν περισσότερων μη-όρων χρήσιμος κρίνεται ο προσδιορισμός ενός κατωφλιού συχνότητας εμφάνισης. Λέξεις που εμφανίζονται με μικρότερη συχνότητα από αυτό το κατώφλι κρίνεται ότι έχουν μικρή πιθανότητα να είναι όροι και, συνεπώς, αποκλείονται από τον τελικό κατάλογο που θα δοθεί για επισκόπηση στον ειδικό.

Η μελέτη μας, καθώς και άλλες μελέτες, έδειξε ότι στις υψηλότερες θέσεις των καταλόγων παρατηρείται υψηλή συγκέντρωση όρων και μικρός αριθμός λημμάτων. Όσο μειώνεται η συχνότητα εμφάνισης, αυξάνεται ο αριθμός των λημμάτων και μειώνεται η συγκέντρωση όρων, ώσπου στις τελευταίες θέσεις του καταλόγου η πιθανότητα εμφάνισης όρων να είναι αμελητέα. Συγκεκριμένα, εμφανίζονται περισσότερα λήμματα με συχνότητα 8 παρά με συχνότητα 95, αλλά, ταυτόχρονα, στις χαμηλότερες συχνότητες ο λόγος όρων και μη-όρων τείνει να μειώνεται.

Ο ορισμός ενός καθολικού κατωφλιού δεν είναι δυνατός ούτε επιθυμητός, καθώς εξαρτάται από το μέγεθος και το είδος των κειμένων.

Σημαντική πληροφορία για τον βαθμό σύνδεσης δύο λέξεων δίνεται από τη ***σύγκριση της συχνότητας συνεμφάνισής τους σε σχέση με τη συχνότητα εμφάνισης της κάθε λέξης*** ανεξάρτητα. Παρατηρείται ότι στην περίπτωση των στενά συνδεδεμένων λέξεων η συχνότητα συνεμφάνισής τους τείνει προς τη συχνότητα εμφάνισης τουλάχιστον της μίας λέξης. Ο βαθμός σύνδεσης, παραδείγματος χάρη, του διλεκτικού σχηματισμού *φορολογία εισοδήματος* είναι ισχυρός :

Λέξη 1	**Συχνότητα Λέξης 1**	**Λέξη 2**	**Συχνότητα Λέξης 1**	**Συχνότητα συνεμφάνισης**
φορολογία	21	εισοδήματος	13	10
φόρου	173	εισοδήματος	13	3

Οι τρεις φορές που η λέξη *εισοδήματος* δεν εμφανίστηκε με το *φορολογία* ήταν με το *φόρος*, λέξη συγγενή.
Η πληροφορία αυτή δεν αποτελεί από μόνη της ένδειξη ότι ο συνδυασμός αυτός είναι όρος. Στην περίπτωση, όμως, που μία από τις λέξεις του πολυλεκτικού σχηματισμού έχει ήδη εντοπιστεί ως υποψήφιος μονολεκτικός όρος, η υψηλή συχνότητα συνεμφάνισής της με άλλες λέξεις αποτελεί πολύ ισχυρή ένδειξη ότι και ο σχηματισμός αυτός είναι όρος.

Οι στατιστικές ενδείξεις συνδυάζονται με γλωσσικά κριτήρια. Τα κριτήρια αυτά εξαρτώνται από τα μορφο-συντακτικά χαρακτηριστικά κάθε γλώσσας και αναδεικνύονται από τους ταξινομημένους καταλόγους. Για την Ελληνική γλώσσα, συγκεκριμένα, ένα τέτοιο κριτήριο αποτελεί η ***σύγκριση της συχνότητας εμφάνισης των λεκτικών τύπων που εμφανίστηκαν στα κείμενα και της συνολικής συχνότητας εμφάνισης του λήμματος*** στο οποίο ανάγονται.
Όταν η συχνότητα εμφάνισης ενός συγκεκριμένου λεκτικού τύπου στα ειδικά κείμενα τείνει προς τη συνολική συχνότητα εμφάνισης του λήμματός του, το λήμμα αυτό έχει μεγάλη πιθανότητα να είναι όρος ή συστατικό πολυλεκτικού όρου. Παρατηρείται ότι οι ειδικές γλώσσες προκρίνουν ορισμένους λεκτικούς τύπους ενός λήμματος. Στη συγκεκριμένη εφαρμογή, το λήμμα *προϊόν* εμφανίστηκε συνολικά 91 φορές, από τις οποίες 44 ως *προϊόντα* (ονομαστική ή αιτιατική πληθυντικού) και 46 ως *προϊόντων* (γενική πληθυντικού).

Αξίζει νασημειωθεί ότι ο ενικός αριθμός αντιπροσωπεύεται μόνο από τη γενική *προϊόντος* που εμφανίζεται 1 φορά, ενώ η ονομαστική και η αιτιατική ενικού απουσιάζουν εντελώς.

Λεκτικός τύπος	**Συχνότητα εμφάνισης**
προϊόν	0
προϊόντος	1
προϊόντα	44
προϊόντων	46
ΣΥΝΟΛΟ	**91**

Στην υπογλώσσα της φορολογικής νομοθεσίας (όπως αντικατοπτρίζεται στα συγκεκριμένα κείμενα) προτιμάται η χρήση του όρου *προϊόν* στον πληθυντικό. Η απουσία ενικού αριθμού υποδεικνύει ότι το λήμμα αυτό υπακούει σε ιδιοσυγκρατικό τρόπο χρήσης στο γνωστικό τομέα της φορολογίας, ο οποίος αποκλίνει από τον τρόπο χρήσης στη γενική γλώσσα. Δεδομένου ότι η γενική γλώσσα δεν παρουσιάζει τόσο έντονη πόλωση προς κάποιο αριθμό, η λέξη θεωρείται ότι έχει αποκτήσει εξειδικευμένη θέση στην εννοιολογική ιεραρχία του τομέα.

Η πολωμένη εμφάνιση ενός συγκεκριμένου λεκτικού τύπου έναντι άλλων του ίδιου λήμματος μας προσφέρει επιπλέον ενδείξεις για το αν πρόκειται για μονολεκτικό όρο ή συστατικό πολυλεκτικού όρου. Για την ελληνική γλώσσα, η συχνότερη χρήση της γενικής έναντι των άλλων πτώσεων ενός λήμματος αποτελεί ένδειξη ότι πιθανόν να πρόκειται για το δεύτερο συστατικό πολυλεκτικού όρου που αντιστοιχεί στη δομή "ουσιαστικό + ουσιαστικό σε γενική".

υπηρεσιών 128 εμφανίσεις
ΥΠΗΡΕΣΙΑ 204 εμφανίσεις
➔ *παροχή υπηρεσιών*

Αντίστοιχα, η συστηματική προτίμηση συγκεκριμένων λεκτικών τύπων των συστατικών που απαρτίζουν ένα πολυλεκτικό σχηματισμό, αποτελεί επίσης ισχυρή ένδειξη ότι ο εν λόγω σχηματισμός είναι υποψήφιος όρος. Στην περίπτωση του *δικαίωμα έκπτωσης*, έχουμε :

Λέξη 1	Συχνότητα Λέξης 1	Λέξη 2	Συχνότητα Λέξης 2	Συχνότητα Συνεμφάνισης
δικαίωμα	50	έκπτωσης	46	36
δικαιώματος	14	έκπτωσης	46	3
δικαιώματα	7	έκπτωσης	46	0
δικαιωμάτων	3	έκπτωσης	46	0
δικαίωμα	50	εκπτώσεων	2	0
δικαιώματος	14	εκπτώσεων	2	0
δικαιώματα	7	εκπτώσεων	2	0
δικαιωμάτων	3	εκπτώσεων	2	0

Παρατηρούμε δηλαδή δύο λήμματα, τα οποία ενώ χρησιμοποιούνται ελεύθερα σε όλους τους λεκτικούς τύπους, όταν συνεμφανίζονται, και τα δύο χρησιμοποιούνται αποκλειστικά στον ενικό (*δικαίωμα / δικαιώματος έκπτωσης*).

Στην περίπτωση των τριλεκτικών και τετραλεκτικών σχηματισμών, η υψηλή συχνότητα συνεμφάνισης αποτελεί ισχυρότερη ένδειξη από ό,τι στους διλεκτικούς. Η κατ' επανάληψη χρήση συνδυασμού τριών ή τεσσάρων λέξεων σε μικρή απόσταση έχει χαμηλή πιθανότητα να είναι τυχαία.

Σύμφωνα με όσα αναφέρθηκαν στην ενότητα 1, σχετικά με την παραγωγή τριλεκτικών και τετραλεκτικών όρων με βάση διλεκτικούς όρους, ισχυρή ένδειξη ότι ένας τριλεκτικός ή τετραλεκτικός σχηματισμός είναι όρος αποτελεί ***ο εντοπισμός δύο συστατικών του στον κατάλογο των υποψήφιων διλεκτικών όρων***. Προϋπόθεση για την ασφαλή εφαρμογή του κριτηρίου αυτού είναι ο τριλεκτικός ή τετραλεκτικός όρος να παράγεται σύμφωνα με τις επιτρεπτές διαδικασίες παραγωγής όρων. Τσυτόχρονα, ο εγκιβωτισμός ενός διλεκτικού σχηματισμού σε ένα μεγαλύτερο ενισχύει την πιθανότητα να είναι και αυτός όρος.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η μέθοδος που παρουσιάστηκε είναι **ημι-αυτόματη** ως προς την προσέγγιση, **υβριδική** ως προς τις τεχνικές που χρησιμοποιεί και **τυπολογική** ως προς την οργάνωση των δεδομένων :

- Ο ημι-αυτόματος χαρακτήρας της συνίσταται στο ότι αποσκοπεί να εξαγάγει υποψήφιους όρους που θα αξιολογηθούν από ειδικούς.
- Ο υβριδικός χαρακτήρας της συνίσταται στο συνδυασμό αντικειμενικών στατιστικών μετρήσεων και γλωσσικών φίλτρων που βασίζονται σε μελέτη των ειδικών γλωσσών και αποκρυσταλλώνουν τη μορφο-συντακτική συμπεριφορά των όρων. Με αυτόν τον τρόπο, η ποσοτικοποίηση των δεδομένων συναρτάται από ποιοτικές παραμέτρους.
- Τέλος, ο τυπολογικός της χαρακτήρας συνίσταται στην οργάνωση των δεδομένων με βάση κοινή γραμματική κατηγορία και κοινή συντακτική συμπεριφορά.

Βιβλιογραφία

1. Bourigault, D. (1992) "Surface grammatical analysis for the extraction of terminological noun phrases". In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING-92)*, Nantes, France.
2. Calzolari, N. and R. Bindi (1990) "Acquisition of lexical information from a large textual Italian corpus". In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki, Finland.
3. Church, K. and P. Hanks (1990) "Word Association norms, mutual information and lexicography". In *Computational Linguistics*, vol. 16:1.
4. Daille, B., E. Gaussier and J.-M. Lange (1994) "Towards Automatic Extraction of Monolingual and Bilingual Terminology". In *Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-94)*, Kyoto, Japan.
5. Daille, B. (1994) "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology". In *The Balancing Act : Combining Symbolic and Statistical Approaches to Languages*, Workshop at the 32nd Annual Meeting, Association for Computational Linguistics, Las Cruces, New Mexico.
6. Kageura, K. and B. Umino (1996) "Methods of automatic term recognition : A review". In *Terminology 3:2*.

Μαρία Γαβριηλίδου, Γλωσσολόγος Πένυ Λαμπροπούλου, Γλωσσολόγος,
Ινστιτούτο Επεξεργασίας Λόγου, Μάργαρη 22, 11525 Αθήνα