ΔΗΜΙΟΥΡΓΙΑ ΕΝΟΣ ΠΟΛΥΓΛΩΣΣΟΥ ΘΗΣΑΥΡΟΥ ΜΕ ΒΑΣΗ ΔΙΕΘΝΗ ΠΡΟΤΥΠΑ

Μαρίνα Βασιλείου, Στέλλα Μαρκαντωνάτου

ΠΕΡΙΛΗΨΗ

Η παρούσα ανακοίνωση έχει ως θέμα τη δημιουργία ενός πολύγλωσσου θησαυρού όρων, ο οποίος συνιστά αποτέλεσμα συγχώνευσης ορολογικών βάσεων σε διάφορες γλώσσες και χρησιμοποιείται από εταιρείες για ευρετηρίαση (indexing) και αναζήτηση (searching). Ο συγκεκριμένος θησαυρός, η κατασκευή του οποίου δεν έχει ακόμα ολοκληρωθεί, περιέχει όρους από ποικίλες θεματικές περιοχές, καλύπτει έξι (6) γλώσσες (Αγγλικά, Ελληνικά, Γερμανικά, Ιταλικά, Γαλλικά, Ισπανικά) και μέχρι στιγμής εκτείνεται ιεραρχικά σε βάθος οκτώ (8) επιπέδων. Περιέχει σχέσεις υπωνυμίας - κατά κύριο λόγο - και μερωνυμίας.

Θα αναφερθούμε στη διαδικασία συγχώνευσης των ορολογικών βάσεων, καθώς και στην προσπάθειά μας να καταστήσουμε το θησαυρό συμβατό με υπάρχοντα διεθνή πρότυπα ταξινόμησης όρων. Επιπλέον, θα αναφερθούμε στις που ακολουθήσαμε στην αντιστοίχηση όρων από διαφορετικές γλώσσες και όρων διαφορετικού μήκους (μονολεκτικοί όροι vs. πολυλεκτικοί όροι). Τέλος, θα σχολιάσουμε τον ευέλικτο χαρακτήρα του θησαυρού αυτού τόσο σε επίπεδο κάλυψης όρων διαφορετικής γλώσσας όσο και σε επίπεδο επεκτασιμότητας με προσθήκη νέων όρων ή περισσότερων επιπέδων.

DEVELOPMENT OF A MULTILINGUAL THESAURUS BASED ON INTERNATIONAL STANDARDS

Marina Vassiliou, Stella Markantonatou

ABSTRACT

In this paper we report on the development of a multilingual thesaurus, the result of a merging procedure of terminological databases in various languages, which is used by companies for indexing and searching purposes. The terms of the specific thesaurus cover various thematic domains and belong to six (6) different languages (English, Greek, German, Italian, French, Spanish). The thesaurus hierarchy currently reaches the depth of eight (8) levels and the relations represented are those of hyponymy and meronymy.

We will refer to the merging procedure of the terminological databases, as well as to our attempt to render the thesaurus compatible with already existing international term classification standards. Moreover, we will refer to the solutions that we followed in achieving correspondence between terms from different languages and between terms of variant length (one-word vs. multi-word terms). Finally, we will emphasise on the flexibility of the specific thesaurus, as regards both the coverage of additional languages and its extensibility to additional hierarchical levels.

1. Η ΑΝΑΓΚΗ ΓΙΑ ΕΝΑΝ ΠΟΛΥΓΛΩΣΣΟ ΘΗΣΑΥΡΟ

Η προσπάθεια δημιουργίας ενός πολύγλωσσου θησαυρού εντάσσεται στα πλαίσια του έργου e-Content "ML-Images! A Multilingual Search System for Exploring Large Image Databases". Στόχος του έργου είναι η αναζήτηση και η ανάκτηση ψηφιακών εικόνων, οι οποίες βρίσκονται σε γεωγραφικά διάσπαρτες βάσεις. Οι εικόνες αυτές, παροχείς των οποίων είναι εταιρείες πώλησης ψηφιακών εικόνων μέσω διαδικτύου, που μετέχουν στο έργο, συνήθως συνοδεύονται από κάποιου είδους σχολιασμό (annotation), ο οποίος γίνεται σε μία ή το πολύ δύο γλώσσες και περιλαμβάνει τα ακόλουθα:

- πληροφοριακά στοιχεία (π.χ. προέλευση, χρονολογία, δημιουργός, κάτοχος πνευματικών δικαιωμάτων κλπ)
- τεχνικά χαρακτηριστικά (π.χ. ανάλυση εικόνας, μέγεθος, χρώμα, προσανατολισμός)
- περιγραφή περιεχομένου (τίτλος, λέξεις-κλειδιά, σύντομο κείμενο)

Δεδομένου ότι η απόληξη του έργου είναι η κατασκευή μίας δικτυακής πύλης, όπου ο τελικός χρήστης θα μπορεί να πληκτρολογεί λέξεις / φράσεις αναζήτησης σε διάφορες γλώσσες, ανακτώντας εικόνες με μονόγλωσσο σχολιασμό, προέβαλε ως αναγκαία η παράμετρος της πολυγλωσσικότητας.

1.1 Πρακτική των εταιρειών

Σε αυτό το σημείο θα πρέπει να γίνει μία σύντομη αναφορά στις μεθόδους σχολιασμού και καταχώρησης, που ακολουθούν οι προαναφερθείσες εταιρείες. Οι εταιρείες αυτές δεν έχουν κάποια συγκεκριμένη πρακτική ευρετηρίασης, η οποία να περιλαμβάνει τη συστηματική χρήση θησαυρών ή οντολογιών. Ο σχολιασμός των εικόνων είναι περισσότερο προσωπική υπόθεση του εκάστοτε σχολιαστή (annotator), ο οποίος δεν εργάζεται βάσει κάποιων αυστηρών προδιαγραφών ούτε είναι υποχρεωμένος να χρησιμοποιεί αποκλειστικά τους όρους κάποιας ορολογικής βάσης.

1.2 Το πρωτογενές υλικό

Στα πλαίσια του έργου είχαμε στη διάθεσή μας όρους από τρεις διαφορετικές πηγές:

1. ένα σύνολο περίπου 2000 λέξεων οργανωμένων σε θεματικές κατηγορίες και βάθους τριών επιπέδων (στη Γερμανική γλώσσα) π.χ.
   (1) *Architektur* (= *Αρχιτεκτονική*)
   (2) *Stile* (= *Ρυθμός*)
   (3) *Klassizismus* (= *Κλασσικισμός*)

2. ένα σύνολο περίπου 500 λέξεων (στην Ιταλική & Αγγλική γλώσσα), συνοδευόμενων από συναφείς όρους (στην Ιταλική γλώσσα) π.χ.
   cosmesi – cosmetics (= *Καλλυντικά*)
   cosmetico (= *καλλυντικό*), *crema* (= *κρέμα*), *pomata* (= *πομάδα*), *unguento* (= *αλοιφή*)
3. έναν ιεραρχικά δομημένο θησαυρό όρων με βάση το πρότυπο ISO 2788 [4] για την κατασκευή μονόγλωσσων θησαυρών (στη Γαλλική γλώσσα) π.χ.
   TERME: résineux (= ρητινώδες δέντρο)
   SYN: conifère (= κωνοφόρο)
   NT: cèdre (= κέδρος)
   NT: pin (= πεύκο)
   RT: pinède (= πευκώνας)
   NT: sapin (= έλατο)

Τα παραπάνω συνιστούν ένα αρκετά ετερογενές υλικό, δύσκολο στην επεξεργασία του, όχι μόνο λόγω της ποικιλότητας των γλωσσών, αλλά και εξαιτίας της διαφορετικότητας δόμησης των όρων και του γεγονότος ότι οι έννοιες καλύπτουν ένα ευρύ φάσμα θεματικών περιοχών (Τέχνη, Βιολογία, Ιστορία, Αθλητισμός, Πολιτική, Επιστήμες, Μόδα κλπ.). Δεδομένου αυτού του ετερόκλιτου υλικού, η πολυγλωσσικότητα θα μπορούσε να επιτευχθεί μόνο με τη δημιουργία ενός πολύγλωσσου θησαυρού όρων, ο οποίος θα καλύπτει όλα τα παραπάνω και θα αντανακλά τις επιμέρους μονόγλωσσες ορολογικές βάσεις. Ο θησαυρός αυτός, που φέρει τον τίτλο "Multilingual ML-Images! Matrix" (MMM), είναι ακόμα εν εξελίξει και προς το παρόν εκτείνεται σε βάθος οκτώ επιπέδων.

2. ΔΗΜΙΟΥΡΓΙΑ ΚΑΙ ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΘΗΣΑΥΡΟΥ MMM

Ο θησαυρός MMM έχει σχετικά απλή δομή, η οποία αντανακλά τις ιεραρχικές σχέσεις υπωνυμίας και μερωνυμίας μεταξύ των όρων. Η πρώτη στήλη κάθε εγγραφής περιέχει ένα μοναδικό αριθμό "**ταυτοποίησης έννοιας**" (**Concept id**), ο οποίος καταδεικνύει τη θέση του δεδομένου όρου στη θεματική ιεραρχία. Οι επόμενες στήλες περιέχουν τον ίδιο τον όρο σε έξι (6) γλώσσες[1]:

[1] Οι όροι δεν έχουν ακόμα μεταφερθεί στην Ισπανική γλώσσα. Όπως ήδη αναφέρθηκε, ο θησαυρός τελεί υπό κατασκευή.

Concept id	English	German	Greek	Italian	French
01	architecture	Architektur	αρχιτεκτονική	architettura	architecture
0108	building	Bauwerk	οικοδόμημα	edifici	édifice
0108002	church	Kirche	εκκλησία	chiesa	temple
01080020001	basilica	Basilika	βασιλική	basilica	
01080020002	cathedral	Kathedrale	καθεδρικός ναός	cattedrale	cathédrale
01080020003	chapel	Kapelle	παρεκκλήσι	cappella	chapelle
01080020004	cupola	Kuppel	τρούλος	cupola	coupole
01080020005	dome	Dom	μητρόπολη	duomo	métropole

Κανονικά οι όροι θα πρέπει να δίνονται σε λημματική μορφή. Ενδεχομένως, όμως, ο τελικός χρήστης να εισαγάγει ως όρο αναζήτησης κάποια λέξη στον πληθυντικό αριθμό. Επειδή, όμως, δε χρησιμοποιήθηκε μορφολογικός επεξεργαστής, ο οποίος θα μπορούσε να επεξεργαστεί τους όρους αναζήτησης και να εξαγάγει τα εκάστοτε λήμματα, αποφασίσαμε να καταχωρούμε κάθε όρο στην Ονομαστική Ενικού και Πληθυντικού, προκειμένου να καλύπτονται όλες οι πιθανές επιλογές του τελικού χρήστη.

Επιπλέον, η κάθε εγγραφή περιέχει όλους τους συνώνυμους ή σχεδόν συνώνυμους όρους, οι οποίοι θεωρούνται ισότιμοι. Ειδικότερα, σε κάθε εγγραφή μπορούν να εμφανισθούν τα ακόλουθα:

- συνώνυμοι ή σχεδόν συνώνυμοι όροι π.χ. *"κολιέ – περιδέραιο", "πετεινός – κόκορας"*
- όροι στο συνήθη τους τύπο συνοδευόμενοι από εναλλακτικούς τύπους διαφορετικού γένους ή υποκοριστικούς τύπους π.χ. *"γάτα – γάτος – γατάκι – γατούλα" / "cat – kitten – kitty"* (= *γάτα*)
- όροι συνοδευόμενοι από τα αλλόμορφά τους π.χ. *"αστυνόμος – αστυνομικός", "αδερφός – αδελφός", "Kriminalfilm – Krimi"* (= *αστυνομική ταινία*)
- όροι που εμφανίζουν λεξικές παραλλαγές π.χ. *"Κοινοβούλιο – Βουλή"*
- ακρωνύμια συνοδευόμενα από την ανεπτυγμένη τους μορφή π.χ. *"Τ.V. – Τηλεόραση"*
- μεταγεγραμμένοι ξενόγλωσσοι όροι συνοδευόμενοι από τον αντίστοιχο όρο στην οικεία γλώσσα π.χ. "Μπάντμιντον – Πετοσφαίριση", "Lifestyle – Lebensstil", "Καρτούν(ς) – Κινούμενα σχέδια"

3. ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ ΘΗΣΑΥΡΟΥ ΜΜΜ

Ως βάση για τη δημιουργία του θησαυρού ΜΜΜ απετέλεσε η Γερμανική ορολογική βάση, η οποία αντανακλά μέχρι ενός ορισμένου σημείου το διεθνές ταξινομητικό πρότυπο IPTC (βλ. επόμενο κεφάλαιο), χωρίς όμως να υπάρχει πλήρης αντιστοίχηση με τους όρους του ή να ακολουθεί πιστά την ιεραρχία του.

Αυτή η ιεραρχία των τριών επιπέδων εμπλουτίσθηκε στη συνέχεια με τους Ιταλικούς και Γαλλικούς όρους βάσει γενικών σημασιολογικών αρχών [2] και αποδόθηκε στις 6 γλώσσες. Επιπλέον, επαυξήθηκε με περισσότερα επίπεδα, προκειμένου να καλυφθούν όλοι οι όροι και να αποδοθούν λεπτομερέστερα οι σχέσεις υπωνυμίας και μερωνυμίας.

4. ΣΥΜΒΑΤΟΤΗΤΑ ΜΕ ΔΙΕΘΝΗ ΠΡΟΤΥΠΑ

Κατά τη δημιουργία του θησαυρού ΜΜΜ προσπαθήσαμε να επιτύχουμε τη μέγιστη δυνατή συμβατότητα με διεθνή ταξινομητικά πρότυπα [3] ή πρότυπα προδιαγραφών για τη δημιουργία θησαυρών [1] , μονόγλωσσων [4, 6] ή πολύγλωσσων [5]. Στη συνέχεια αναλύουμε το πώς επιχειρήσαμε να ενσωματώσουμε τα πρότυπα αυτά στην κατασκευή του θησαυρού.

4.1 Το πρότυπο IPTC

Το πρότυπο IPTC [3] είναι ένα διεθνές ταξινομητικό πρότυπο, το οποίο χρησιμοποιείται από ειδησεογραφικά πρακτορεία για το σχολιασμό, ευρετηρίαση και αναζήτηση εικόνων και αρχείων κειμένου και το οποίο ανανεώνεται ανά σχετικά μικρά χρονικά διαστήματα. Είναι διαθέσιμο στην Αγγλική γλώσσα, ενώ υπάρχουν και εκδόσεις σε άλλες γλώσσες, που δεν καλύπτουν όμως όλους τους όρους. Οι όροι αυτού του προτύπου είναι ιεραρχικά οργανωμένοι (σε 3 επίπεδα) από το γενικότερο στον πιο συγκεκριμένο. Ακολουθεί ενδεικτικό απόσπασμα από το πρότυπο αυτό:

Subject = 11000000 Name = politics (= Πολιτική)
SubjectMatter = 11016000 Name = interior policy (= *Εσωτερική πολιτική*)
SubjectDetail = 11016001 Name = data protection (= *προστασία δεδομένων*)
SubjectDetail = 11016002 Name = housing and urban planning (= *οικοδομικός & αστικός σχεδιασμός*
SubjectDetail = 11016003 Name = pension and welfare (= *συντάξεις και πρόνοια*)
SubjectDetail = 11016004 Name = personal weapon control (= *έλεγχος οπλοφορίας πολιτών*)

Δεδομένου ότι συχνοί χρήστες των υπηρεσιών πώλησης ψηφιακών εικόνων είναι οι δημοσιογράφοι και γενικότερα τα ειδησεογραφικά πρακτορεία, που είναι εξοικειωμένοι με την ορολογία του IPTC, θεωρήσαμε απαραίτητο να καταστήσουμε συμβατό το θησαυρό ΜΜΜ με το εν λόγω πρότυπο, αντιστοιχίζοντας τους όρους του θησαυρού ΜΜΜ με τους όρους του IPTC με βάση την αριθμητική κωδικοποίηση του IPTC.

4.2 Το πρότυπο ISO 5964

Το πρότυπο ISO 5964 [5] παρέχει λεπτομερείς προδιαγραφές για τη δημιουργία πολύγλωσσων θησαυρών. Χειρίζεται θέματα καταχώρησης όρων, σε συμφωνία με το πρότυπο ISO 2788 [4] για την κατασκευή μονόγλωσσων θησαυρών, και ζητήματα αντιστοίχησης όρων από διαφορετικές γλώσσες. Ωστόσο, καταλήξαμε στο ότι αυτές οι προδιαγραφές ίσως είναι κατάλληλες για ευρετηρίαση, όχι όμως και για αναζήτηση όρων. Προσπαθήσαμε, λοιπόν, να συμβιβάσουμε τις οδηγίες του προτύπου με τις ανάγκες του τελικού χρήστη, που ήταν βασικός παράγοντας στη λήψη των όποιων αποφάσεων κατά την κατασκευή του θησαυρού.

Πιο συγκεκριμένα, το πρότυπο ISO 5964 προτείνει για τους συνώνυμους ή παρεμφερείς όρους τη διάκριση μεταξύ προτιμώμενου όρου και μη προτιμώμενων όρων. Δεδομένου ότι ο θησαυρός καλύπτει ποικίλες θεματικές περιοχές, δεν ήμαστε σε θέση να αποφασίσουμε για το ποιο όροι είναι προτιμητέοι και ποιοι όχι. Εξάλλου, μία τέτοια διάκριση δεν είναι λειτουργική, καθώς ο τελικός χρήστης μπορεί να χρησιμοποιήσει έναν από τους μη προτιμώμενους όρους ως όρους αναζήτησης. Συνεπώς, αναγκαζόμαστε να θεωρήσουμε όλους τους όρους ως ισότιμους και αποδεκτούς (π.χ. ανατολή ηλίου, αυγή, χάραμα).

Το ίδιο ισχύει και στην περίπτωση των ακρωνυμίων και της αναπτυγμένης τους μορφής (π.χ. Ευρωπαϊκή Ένωση – ΕΕ).

Παρόμοια τακτική ακολουθούμε και στην περίπτωση κύριου και σχετιζόμενου όρου. Οι δύο όροι είτε θεωρήθηκαν ισότιμοι (π.χ. πεύκο, πευκώνας) είτε εισήλθαν σε σχέση υπωνυμίας (π.χ. γαλακτοκομικά – γάλα) ή μερωνυμίας (π.χ. κήπος – συντριβάνι).

Τέλος, η διάκριση μεταξύ ευρύτερου όρου και στενότερου όρου αντανακλάται στις ιεραρχικές σχέσεις υπωνυμίας και μερωνυμίας του θησαυρού. Οι στενότεροι όροι, δηλαδή, αντιπροσωπεύονται ως υπώνυμα (π.χ. τροφή – φρούτο – σταφύλι) ή μερώνυμα (π.χ. ζώο – ουρά).

5. ΑΝΤΙΣΤΟΙΧΗΣΗ ΟΡΩΝ

Όσον αφορά στην αντιστοίχηση όρων από διαφορετικές γλώσσες, δύο ζητήματα προκύπτουν σχετικά:

α) η διαφορά μήκους (μονολεκτικοί και πολυλεκτικοί όροι)

β) η έλλειψη αντιστοιχίας 1:1 μεταξύ των όρων

Το πρώτο ζήτημα, δηλαδή η αναντιστοιχία μήκους μεταξύ των όρων του θησαυρού, δε μας απασχόλησε ιδιαίτερα, καθώς βρήκαμε τρόπους χειρισμού των πολυλεκτικών όρων από τη μηχανή αναζήτησης[2].

Όσον αφορά στο δεύτερο ζήτημα, την έλλειψη αντιστοιχίας 1:1 μεταξύ των όρων, προσπαθήσαμε να ξεπεράσουμε αυτή την αναντιστοιχία, εκλαμβάνοντας όλους τους όρους ως ισότιμους, ακόμα και εάν δεν είχαν ταυτόσημη δήλωση (denotation), π.χ. Science – Wissenschaft – Επιστήμη – Scienza – Science.

6. ΕΠΙΛΟΓΟΣ

Ο θησαυρός ΜΜΜ έχει ήδη ενσωματωθεί στο πιλοτικό πρωτότυπο του έργου και έχει λειτουργήσει πολύ ικανοποιητικά. Το θετικό του στοιχείο είναι πως είναι εύκολα επεκτάσιμος τόσο σε επίπεδο κάλυψης όρων άλλης γλώσσας όσο και σε επίπεδο επεκτασιμότητας με προσθήκη νέων όρων ή περισσότερων επιπέδων. Έχει ενσωματώσει με δημιουργικό τρόπο τις προδιαγραφές διεθνών προτύπων, το κυριότερο, όμως, πλεονέκτημά του είναι ότι καλύπτει πολλές και ποικίλες θεματικές περιοχές, γεγονός που τον καθιστά έναν πολύτιμο γλωσσικό πόρο, ο οποίος μπορεί άνετα και άμεσα να επαναχρησιμοποιηθεί.

Βιβλιογραφία

1. Aitchison J, A. Gilchrist & D. Bawden: *Thesaurus Construction and use: a practical manual*: Aslib 2000
2. Cruse D. A.: *Lexical Semantics*. Cambridge University Press: CUP, 1986
3. IPTC [http://www.IPTC.org/]
4. ISO Standard 2788: Documentation - Guidelines for the establishment and development of monolingual thesauri, 1986
5. ISO Standard 5964: Documentation - Guidelines for the establishment and development of multilingual thesauri, 1985

[2] Αναπτύξαμε μία σειρά προδιαγραφών / οδηγιών για τον εντοπισμό της κεφαλής του εκάστοτε πολυλεκτικού όρου. Συνεπώς, εάν η μηχανή αναζήτησης δεν εντοπίσει μέσα στο θησαυρό ΜΜΜ τον πολυλεκτικό όρο όπως έχει, θα προσπαθήσει να εντοπίσει την κεφαλή (π.χ. άγριο ποτάμι vs. ποτάμι)

6. National Information Standards Organisation: Guidelines for the Construction, Format, and Management of Monolingual Thesauri. Bethesda: NISO Press

Μαρίνα Βασιλείου (*Γλωσσολόγος*)
Δρ. Στέλλα Μαρκαντωνάτου (*Ερευνήτρια*)
Ινστιτούτο Επεξεργασίας του Λόγου (ΙΕΛ)
Αρτέμιδος 6 & Επιδαύρου
Παράδεισος Αμαρουσίου
151 25 Αθήνα
Τηλ.: *210 6875452*
Τηλεομ.: *210 6856794*
Ηλ. διεύθ.: *{mvas, marks}@ilsp.gr*