12 Λεξικό όρων Υποδομών Γλωσσικών Πόρων και Γλωσσικής Τεχνολογίας

Μαρία Γαβριηλίδου, Πένυ Λαμπροπούλου, Κανέλλα Πουλή, Ηρώ Τσιούλη

ΠΕΡΙΛΗΨΗ

Η ανακοίνωση αυτή περιγράφει τη δημιουργία ενός δίγλωσσου λεξικού (αγγλικά-ελληνικά) με όρους Γλωσσικής Τεχνολογίας και Γλωσσικών Πόρων, καθώς και όρους που περιγράφουν την υποδομή όπου οι πόροι τεκμηριώνονται και φιλοξενούνται, αλλά και τους όρους που χρησιμοποιούνται για τις υπηρεσίες διαχείρισης και επεξεργασίας τους. Το λεξικό δημιουργήθηκε με τη μορφή Συνδεδεμένων Δεδομένων (Linked Data), και στην παρούσα φάση περιλαμβάνει 575 εγγραφές, ενώ οι όροι είναι οργανωμένοι σε σύνολα λέξεων με όρους που έχουν χαρακτηριστεί ως προτιμητέοι και συνοδεύονται από όρους που είναι χαρακτηρισμένοι ως εναλλακτικοί. Σε επόμενο στάδιο προβλέπεται η δημιουργία ιεραρχιών και λοιπών σχέσεων μεταξύ των όρων. Τόσο η τρέχουσα μορφή του λεξικού όσο και η μελλοντική θα διατίθενται μέσω της υποδομής CLARIN:EL.

Dictionary of terms for Language Resources and Language Technology Infrastructures

Maria Gavriilidou, Penny Labropoulou, Kanella Pouli, Iro Tsiouli

ABSTRACT

In this paper we describe the construction of a bilingual terminological lexicon (English – Greek) that includes mainly terms of Language Technology and Language Resources, as well as terms found in the infrastructure where these are documented and hosted, and terms used for the related services of management and processing. This lexicon, available in the form of Linked Data, currently comprises 575 entries, organized on the basis of concepts, with preferred and alternative terms. At a later stage, the addition of hierarchical and other relations between the terms is envisaged. Both the current and future forms of the lexicon will be available through the CLARIN:EL infrastructure.

0 Εισαγωγή

Στην ανακοίνωση αυτή παρουσιάζεται ένα δίγλωσσο αγγλο-ελληνικό ηλεκτρονικό λεξικό, το οποίο περιλαμβάνει όρους δύο επιμέρους τομέων: αφενός όρους Γλωσσικής Τεχνολογίας και Γλωσσικών Πόρων και αφετέρου όρους σχετικούς με την τεκμηρίωση και διαχείριση των

πόρων και των υπηρεσιών επεξεργασίας τους σε υποδομή που υποστηρίζει τη διάθεσή τους στην ερευνητική κοινότητα. Παρουσιάζονται οι αρχές και η διαδικασία κατασκευής του λεξικού καθώς και η υλοποίησή του με τη μορφή Συνδεδεμένων Δεδομένων (Linked Data).

1 Στόχος

Το λεξικό αυτό επιχειρεί τη συστηματοποίηση και απόδοση στα ελληνικά των όρων Γλωσσικής Τεχνολογίας και Γλωσσικών Πόρων, όπως χρησιμοποιούνται στο πλαίσιο της ερευνητικής υποδομής CLARIN:EL[1] (Gavriilidou et al., υπό έκδοση·Καπιδάκης κ.ά., 2015· Πιπερίδης κ.ά, 2015·Πιπερίδης κ.ά., 2018).

Η υποδομή CLARIN:EL έχει ως αποστολή τη συλλογή, την τεκμηρίωση και τον διαμοιρασμό ψηφιακών Γλωσσικών Πόρων (σωμάτων πολυμεσικών και πολυτροπικών δεδομένων, λεξικών και εννοιολογικών πόρων, γλωσσικών μοντέλων, υπολογιστικών γραμματικών, κτλ.), καθώς και εργαλείων και διαδικτυακών υπηρεσιών Γλωσσικής Τεχνολογίας (λημματοποιητών, συντακτικών αναλυτών, εργαλείων εξαγωγής πληροφορίας κτλ.) για την επισημείωση και ανάλυση πόρων δεδομένων που είναι αναρτημένοι στην υποδομή ή πόρων δεδομένων που διαθέτουν οι ερευνητές. Απευθύνεται σε ερευνητές και στο ευρύ κοινό που δραστηριοποιούνται στη μελέτη της γλώσσας ή/και ανάλυση δεδομένων μέσω της γλώσσας σε ποικίλους επιστημονικούς τομείς και με διαφορετικό βαθμό εξοικείωσης με τις έννοιες και εργασίες της Γλωσσικής Τεχνολογίας. Για την υποστήριξη των χρηστών έχει αναπτυχθεί και αναρτηθεί στη διαδικτυακή πύλη της υποδομής ειδικό εκπαιδευτικό υλικό. Παράλληλα, οι εργασίες που επιτελούνται μέσω της υποδομής υποστηρίζονται από ένα δυναμικό περιβάλλον τεκμηρίωσης, διαχείρισης, αναζήτησης και χρήσης των πόρων και υπηρεσιών (εφεξής διεπαφή CLARIN:EL).

Η υποδομή CLARIN:EL αποτελεί το ελληνικό σκέλος της ευρωπαϊκής υποδομής CLARIN ERIC[2] και απευθύνεται τόσο σε Έλληνες ερευνητές όσο και σε ερευνητές άλλων χωρών που ενδιαφέρονται για τη μελέτη της ελληνικής γλώσσας καθώς και για τη συγκριτική μελέτη με άλλες γλώσσες, μέσω της χρήσης πόρων και υπηρεσιών που διατίθενται από άλλες υποδομές του δικτύου. Για τη σύνδεση με το ευρωπαϊκό δίκτυο και την εξασφάλιση της πρόσβασης στην υποδομή από το ευρωπαϊκό κοινό, η διεπαφή και ο κατάλογος των πόρων και εργαλείων[3] όπως και το περιεχόμενο της διαδικτυακής πύλης διατίθενται στην αγγλική

1 https://www.clarin.gr/

2 https://www.clarin.eu/

3 https://inventory.clarin.gr/

γλώσσα. Για τη διευκόλυνση της ελληνικής ερευνητικής κοινότητας, το περιεχόμενο της διαδικτυακής πύλης και τα υποστηρικτικά εγχειρίδια διατίθενται και στην ελληνική, ενώ ταυτόχρονα είναι σε εξέλιξη η μετάφραση της διεπαφής στην ελληνική.

Στο πλαίσιο αυτό, η ανάγκη συστηματοποίησης της εν λόγω ορολογίας προέκυψε ως απαίτηση για βελτίωση της λειτουργικότητας της υποδομής CLARIN:EL. Διαπιστώθηκε η ανάγκη εναρμόνισης των όρων, καθώς οι ίδιες έννοιες αποδίδονταν με διαφορετικούς όρους σε τέσσερα περιβάλλοντα: (α) στη διεπαφή, (β) στο μοντέλο τεκμηρίωσης που χρησιμοποιείται για την περιγραφή των πόρων[4], (γ) στα εγχειρίδια χρήσης της υποδομής και (δ) στο συνοδευτικό υλικό της διαδικτυακής πύλης. Η διαφορετική ορολογία ήταν αποτέλεσμα της συγγραφής των κειμένων σε διαφορετικά χρονικά διαστήματα, από διαφορετικά άτομα. Η απόπειρα συγκέντρωσης, συστηματοποίησης και κωδικοποίησης των όρων αυτών οδήγησε στον σχεδιασμό και την υλοποίηση του δίγλωσσου λεξικού.

2 Πηγές κατάρτισης λεξικού

Σημαντική πηγή για τη συλλογή των όρων του λεξικού αποτελεί το μοντέλο μεταδεδομένων που χρησιμοποιείται για την τεκμηρίωση των πόρων που διατίθενται μέσω της υποδομής και το οποίο φυσικά περιλαμβάνει όρους σχετικούς με τη Γλωσσική Τεχνολογία και τους Γλωσσικούς Πόρους. Το μοντέλο CLARIN-SHARE (Labropoulou et al., 2020) βασίζεται στο META-SHARE (Gavriilidou et al., 2012), το οποίο αναπτύχθηκε για να υποστηρίξει την τεκμηρίωση Γλωσσικών Πόρων και εργαλείων Γλωσσικής Τεχνολογίας από ομάδα ειδικών από διάφορες ευρωπαϊκές χώρες. Το μοντέλο META-SHARE υποστηρίζει την αναπαράσταση πληροφοριών για τις τεχνικές ιδιότητες και το περιεχόμενο των πόρων, την καταγραφή όλων των διεργασιών στις οποίες υπόκεινται από τη στιγμή της σχεδίασης μέχρι τη χρήση τους, τη σχέση τους με άλλους πόρους κτλ. Το CLARIN-SHARE χρησιμοποιεί ένα υποσύνολο των χαρακτηριστικών αυτών, προσαρμοσμένο ώστε να εξυπηρετήσει καλύτερα τις ανάγκες των χρηστών της υποδομής. Στην παρούσα φάση διατίθεται στην αγγλική γλώσσα, αλλά για τις ανάγκες της ελληνικής ερευνητικής κοινότητας κρίνεται σημαντική η απόδοση των όρων και των συνοδευτικών τους πληροφοριών (π.χ. ορισμών, παραδειγμάτων, σχολίων) στην ελληνική. Δεδομένου ότι το μοντέλο περιλαμβάνει 276 κλάσεις, 581 ιδιότητες, και 1.196 στιγμιότυπα, το έργο της μεταφραστικής απόδοσης είναι ιδιαίτερα απαιτητικό σε χρόνο και ανθρωποδύναμη. Για τον λόγο αυτό, αποφασίστηκε σε πρώτη φάση να δοθεί προτεραιότητα για την ένταξη στο λεξικό σε συγκεκριμένο ελεγχόμενο

[4] Για λόγους οικονομίας, εφεξής ο όρος «πόρος» χρησιμοποιείται τόσο για πόρους δεδομένων όσο και για εργαλεία και διαδικτυακές υπηρεσίες.

λεξιλόγιο που χρησιμοποιείται για την κωδικοποίηση της λειτουργίας που επιτελούν οι υπηρεσίες Γλωσσικής Τεχνολογίας.

Σε κάθε πόρο που διατίθεται από την υποδομή αντιστοιχεί μια εγγραφή μεταδεδομένων. Η δημιουργία των εγγραφών αυτών υποστηρίζεται από τη διεπαφή CLARIN:EL, με τη χρήση της οποίας μπορούν οι πάροχοι να περιγράψουν και να διαθέσουν τους πόρους τους μέσω της υποδομής. Οι εγγραφές αυτές προσφέρουν χρήσιμες πληροφορίες για τη χρήση του εκάστοτε πόρου (π.χ. άδεια χρήσης, μορφότυπο με το οποίο διατίθενται, πληροφορίες για τα περιεχόμενα κτλ,), ενώ, παράλληλα, υποσύνολο των μεταδεδομένων που περιέχονται στις εγγραφές αξιοποιούνται (με τη μορφή φίλτρων και λέξεων αναζήτησης) για τον εντοπισμό πόρων. Οι όροι που περιλαμβάνονται σε αυτό το περιβάλλον είναι τόσο όροι Γλωσσικής Τεχνολογίας όσο και όροι αρχειονομίας, τεκμηρίωσης και διαχείρισης πόρων αποθετηρίων και αποτέλεσαν τη δεύτερη πηγή όρων για το λεξικό.

Για την υποστήριξη των χρηστών καταρτίστηκε δίγλωσσο ηλεκτρονικό εγχειρίδιο χρήσης (Pouli et al., 2023), το οποίο περιγράφει αναλυτικά τις διαδικασίες τεκμηρίωσης και ανάρτησης, αναζήτησης, επιλογής και επεξεργασίας πόρων, ενώ προσφέρει και βασικές πληροφορίες για τα μεταδεδομένα που βάσει του μοντέλου είναι υποχρεωτικά για τη δημιουργία εγγραφής. Το εγχειρίδιο χρήσης συντάχθηκε στα ελληνικά προς διευκόλυνση του συνόλου των χρηστών της υποδομής, που είναι στην πλειοψηφία τους Έλληνες, αλλά και στα αγγλικά για να καλύψει τις ανάγκες των χρηστών της ευρωπαϊκής υποδομής και αξιοποιήθηκε ως τρίτη πηγή όρων για το λεξικό.

Τέλος, το σύνολο του υλικού που περιλαμβάνει η Διαδικτυακή Πύλη του CLARIN:EL, το οποίο αποτέλεσε την τέταρτη πηγή όρων για το λεξικό, είναι δίγλωσσο (ελληνικά και αγγλικά) και αποτελείται από γενικές πληροφορίες σχετικά με την υποδομή, τα μέλη του δικτύου που την απαρτίζουν, τις υπηρεσίες αναζήτησης, επεξεργασίας, τεκμηρίωσης και διαμοιρασμού που υποστηρίζει, καθώς και εκπαιδευτικό και υποστηρικτικό υλικό (π.χ. οδηγίες για συμμετοχή στο δίκτυο, απαντήσεις σε συχνές ερωτήσεις, νομικά θέματα και θέματα πολιτικής, επιστημονικές δημοσιεύσεις, παρουσιάσεις, κτλ.).

3 Διαδικασία κατάρτισης λεξικού

3.1 Συγκέντρωση όρων

Η άντληση των υποψήφιων όρων έγινε με διαφορετικές διαδικασίες, ανάλογα με τη μορφή διάθεσης της κάθε πηγής: η επιλογή όρων από το κειμενικό υλικό (κείμενα τεκμηρίωσης,

εγχειρίδιο χρήσης και υλικό διαδικτυακής πύλης) έγινε από ανθρώπους[5] ενώ η προσθήκη των επιλεχθέντων όρων του μοντέλου τεκμηρίωσης έγινε με αυτόματη εξαγωγή από το λεξιλόγιο OMTD-SHARE[6] (Labropoulou et al., 2022) των όρων που ανήκουν στην ταξινομία Operation, η οποία κωδικοποιεί τις λειτουργίες των υπηρεσιών επεξεργασίας γλώσσας.

Η διαδικασία συλλογής όρων από το κειμενικό υλικό οργανώθηκε στις ακόλουθες ενέργειες:

- Εντοπισμό υποψήφιου όρου στο κειμενικό υλικό στην ελληνική γλώσσα. Ως υποψήφιοι όροι επιλέγονται όσοι είναι ήδη μεταφρασμένοι μέσα στο κείμενο (π.χ. υ*πό συνθήκη υποχρεωτικό στοιχείο (mandatory upon condition element),* ή όσοι βρίσκονται σε μορφή υπερσυνδέσμου, ο οποίος οδηγεί στον ορισμό του όρου, τακτική που υποδεικνύει την σπουδαιότητα του συγκεκριμένου όρου για την υποδομή. Τέλος, εντοπίζονται όροι οι οποίοι, κατά την κρίση των ειδικών, είναι σημαντικό να καταγραφούν.
- Αντιστοίχιση στο μεταφραστικό ισοδύναμο κάθε όρου στην άλλη γλώσσα, όπως εμφανίζεται στο ίδιο το δίγλωσσο κειμενικό υλικό.
- Εύρεση ορισμού και στις δύο γλώσσες, είτε από το ίδιο το κειμενικό υλικό (εφόσον υπήρχε) είτε από άλλες πηγές, με αναγραφή της εκάστοτε πηγής.
- Σε περίπτωση ύπαρξης εναλλακτικών όρων, καταγραφή τους και στις δύο γλώσσες (εφόσον εντοπίζονταν σε κάποια από τις κειμενικές πηγές) ή προσθήκη των εναλλακτικών όρων (εφόσον εντοπίζονταν σε βιβλιογραφικές πηγές).
- Επιλογή και επισήμανση του προτιμητέου όρου σε κάθε γλώσσα.

Από το λεξιλόγιο OMTD-SHARE εξήχθησαν 291 όροι και οι συνοδευτικές τους πληροφορίες (αναγνωριστικό, προτιμητέος όρος, εναλλακτικοί όροι, ορισμός, σχόλιο, σχέσεις υπερωνυμίας και συσχέτισης).

Ως αποτέλεσμα των εργασιών αυτών συγκεντρώθηκαν 935 εγγραφές με τις εξής πληροφορίες: προτιμητέος όρος σε ελληνικά και αγγλικά, εναλλακτικοί όροι σε ελληνικά και αγγλικά (εφόσον υπάρχουν), ορισμός σε ελληνικά και αγγλικά, πηγή ελληνικού και αγγλικού όρου, πηγή ελληνικού και αγγλικού ορισμού (εφόσον ο ορισμός προέρχεται από λεξικό, εγκυκλοπαίδεια, εγχειρίδιο κτλ.) και σχόλια ειδικού.

---

[5] Πολύτιμη στο στάδιο αυτό υπήρξε η συμβολή των Δ. Γρηγορίου, Ε. Πετεινού και Β. Πετικά, φοιτητ(ρι)ών στο ΠΜΣ Ψηφιακές Μέθοδοι στης Ανθρωπιστικές Επιστήμες του ΟΠΑ (Ακαδ. Έτος 2021-22).

[6] http://w3id.org/meta-share/omtd-share

3.2 Έλεγχος καταλόγου υποψήφιων όρων

Το επόμενο στάδιο επεξεργασίας αφορούσε τον έλεγχο του καταλόγου υποψήφιων όρων, με στόχο τις παρακάτω βελτιώσεις:

- Εκκαθάριση διπλοεγγραφών: (α) απαλοιφή της μίας εγγραφής σε περίπτωση όρου προερχόμενου από δύο πηγές (β) ενοποίηση εγγραφών σε μία για όρους που προέρχονται από δύο ή περισσότερες πηγές που παρέχουν συμπληρωματική πληροφορία
- Ποιοτικό έλεγχο των σχολίων των ειδικών
- Εντοπισμό ελλείψεων ή λανθασμένων πληροφοριών (π.χ. απουσίας απόδοσης όρου ή ορισμού σε μία από τις δύο γλώσσες, απουσίας παράθεσης πηγής, λανθασμένου ορισμού) και διόρθωσή τους.

Ειδικά στην περίπτωση των όρων που αντλήθηκαν από το OMTD-SHARE, καθώς οι όροι αυτοί έχουν ήδη ελεγχθεί από ειδικούς στο πλαίσιο προηγούμενων ερευνητικών έργων, θεωρήθηκε ότι δεν χρειαζόταν περαιτέρω έλεγχος, οπότε η μόνη εργασία αφορούσε στη μεταφραστική απόδοση τους στην ελληνική (βλ. επόμενη ενότητα).

3.3 Εμπλουτισμός καταλόγου υποψήφιων όρων

Η διαδικασία εμπλουτισμού εστίασε στην απόδοση μεταφραστικών αντιστοίχων κυρίως για όρους χωρίς καμία μετάφραση. Πρόκειται για 240 όρους που προέρχονται από το μοντέλο τεκμηρίωσης και δεν περιλαμβάνονταν στο δίγλωσσο κειμενικό υλικό, ώστε να τους έχει ήδη αποδοθεί κάποιο μεταφραστικό αντίστοιχο. Καθώς αυτοί αποτελούν ένα μικρό υποσύνολο των όρων που περιλαμβάνονται στο μοντέλο και οι οποίοι θα πρέπει σύντομα να μεταφραστούν, αποφασίστηκε η διεξαγωγή ενός πειράματος με την αξιοποίηση μηχανών αυτόματης μετάφρασης και ελέγχου των αποτελεσμάτων τους. Η αυτόματη μετάφραση χρησιμοποιείται ευρέως για ρέον κείμενο με πολύ καλά αποτελέσματα, ωστόσο η μετάφραση μεμονωμένων λέξεων και φράσεων εκτός συγκειμένου θέτει δυσκολίες.

Για τον σκοπό αυτό χρησιμοποιήθηκαν δύο συστήματα μηχανικής μετάφρασης (DeepL translator[7] και Google translate[8]), οι μεταφραστικές προτάσεις των οποίων ελέγχθηκαν από δύο ειδικούς, οι οποίοι επέλεξαν την κατά τη γνώμη τους εγκυρότερη απόδοση ή, αν καμία

[7] https://www.deepl.com/en/translator

[8] https://translate.google.com/

από τις δύο δεν ήταν κατάλληλες, πρότειναν τρίτη απόδοση. Τέλος, οι επιλογές των ειδικών ελέγχθηκαν από επικυρωτή, ο οποίος έκανε την τελική επιλογή.

Είναι ενδιαφέρον ότι για σχεδόν τους μισούς όρους (124) οι δύο μηχανές πρότειναν το ίδιο μεταφραστικό αντίστοιχο, ενώ και σε πολλές άλλες περιπτώσεις οι διαφορές ήταν μικρές (π.χ. *σχολιασμός παραγλώσσας – παραγλωσσικός σχολιασμός, κανονικοποίση μετρήσεων – κανονικοποίηση των μετρήσεων*). Μεταξύ των δύο ειδικών υπήρξε απόλυτη συμφωνία σε ποσοστό 45% (για 108 όρους), αλλά και στους όρους στους οποίους διαφοροποιήθηκαν οι διαφορές ήταν μικρές, και αφορούσαν αποδόσεις που μπορούν να θεωρηθούν εναλλακτικές.

Στον παρακάτω πίνακα δίνονται στοιχεία για τις επιλογές που έκαναν

- οι δύο ειδικοί: αριθμός περιπτώσεων που επέλεξαν μία από τις δύο μηχανές, περιπτώσεων που επέλεξαν την κοινή (και από τις δύο μηχανές) απόδοση, περιπτώσεις απόδοσης του όρου με δική τους πρόταση
- ο επικυρωτής: φορές που επέλεξε την πρόταση μίας από τις δύο μηχανές, ή την κοινή τους πρόταση, φορές που επέλεξε την πρόταση του εκάστοτε ειδικού και φορές που πρότεινε νέα απόδοση, εφόσον δεν έκρινε καμία πρόταση ικανοποιητική.

Οι προτάσεις μεταφραστικής απόδοσης των δύο μηχανών έγιναν αποδεκτές για λιγότερους από τους μισούς όρους. Τα στοιχεία αυτά δεν είναι ιδιαίτερα ενθαρρυντικά ώστε να χρησιμοποιηθεί η αυτόματη μετάφραση για το σύνολο του μοντέλου τεκμηρίωσης, αλλά το δείγμα του πειράματος ήταν μικρό, και τα κριτήρια επιλογής αυστηρά, δεδομένου ότι μόνο ένα μεταφραστικό αντίστοιχο μπορούσε να επιλεγεί. Εάν συμπεριλάβουμε τις προτάσεις των δύο μεταφραστικών μηχανών ως εναλλακτικούς όρους, τότε τα αποτελέσματα της διαδικασίας κρίνονται ικανοποιητικά. Συνεπώς, σκοπεύουμε να χρησιμοποιήσουμε την ημι-αυτόματη διαδικασία για τη μετάφραση του συνόλου του μοντέλου, ωστόσο με την εξής διαφοροποίηση: (α) αποδοχή όλων των έγκυρων μεταφραστικών αντιστοίχων που προτείνουν οι μηχανές, (β) προσθήκη εναλλακτικών αποδόσεων και (γ) επιλογή του προτιμητέου όρου.

Ειδικός	**DeepL**	**Google**	**Κοινοί**	**Νέα απόδοση**	**Απόδοση Ειδικού 1**	**Απόδοση Ειδικού 2**
Ειδικός 1	52	49	93	46		
Ειδικός 2	27	33	45	135		
Επικυρωτής	19	40	51	10	11	109

Επιπρόσθετα, εμπλουτισμός του καταλόγου έγινε και με την προσθήκη εναλλακτικών μεταφραστικών αντιστοίχων από τις εξής εκδόσεις: Γούτσος & Φραγκάκη (2015), Καρασίμος κ.ά. (2015), Μικρός (2015), Παναγιωτακόπουλος κ.ά. (2022), Παναγιωτάκος (2023) και Τάντος κ.ά. (2015).

3.4 Έλεγχος της καταλληλότητας των όρων σε σχέση με τους στόχους του λεξικού

Ο στόχος δημιουργίας του λεξικού ήταν η συστηματοποίηση της ορολογίας στην υποδομή και η τεκμηρίωση των όρων προς διευκόλυνση παρόχων και χρηστών. Οι όροι αυτοί ανήκουν σε συγκεκριμένους επιστημονικούς τομείς (π.χ. πληροφορική, γλωσσολογία, νομική, αρχειονομία), αλλά συμπεριλαμβάνονται στην υποδομή είτε επειδή χρησιμοποιούνται σε διαδικασίες τεκμηρίωσης, ανάρτησης και επεξεργασίας πόρων, είτε επειδή ανήκουν στα επιστημονικά πεδία που θεραπεύει η συγκεκριμένη υποδομή (γλωσσολογία, γλωσσική τεχνολογία, πληροφορική). Ο έλεγχος επομένως σε αυτό το στάδιο είχε στόχο να επιλεγούν μεταξύ των υποψηφίων όρων εκείνοι που είναι απαραίτητο να λημματογραφηθούν και να οριστούν, να προστεθούν όροι που λείπουν και να διαγραφούν οι περιττοί. Οι διαγραφές όρων αφορούσαν λέξεις που ανήκουν στο γενικό λεξιλόγιο, των οποίων η απόδοση στα ελληνικά έχει καθιερωθεί και δεν χρήζουν επεξήγησης (π.χ. *user, process*). Οι προσθήκες αφορούσαν κυρίως εναλλακτικούς όρους (π.χ. προστέθηκε ο όρος *annotator* στον υπάρχοντα όρο *annotation tool*), ή προσθήκη πολυλεκτικών σύμπλοκων όρων εφόσον λημματογραφούνται τα συνθετικά τους (π.χ. προστέθηκαν οι όροι *multilingual corpus* και *multilingual data,* δεδομένου ότι υπήρχαν στον κατάλογο οι μονολεκτικοί όροι *multilingual, corpus* και *data).*

Τέλος, για κάθε όρο προστέθηκε η πληροφορία του πεδίου στο οποίο ανήκει (Γλωσσική Τεχνολογία ή Υποδομή).

3.5 Έλεγχος απόδοσης όρων

Η κατάρτιση του λεξικού ακολουθεί τις βασικές αρχές ορολογίας (Ελληνικός Οργανισμός Τυποποίησης [ΕΛΟΤ], 2010· International Organization for Standardization [ISO], 2022). Για την απόδοση των ελληνικών όρων, επιλέχθηκε η τυπική λεξικογραφική / ορογραφική πρακτική της καταγραφής του λήμματος σε ονομαστική πτώση ενικού αριθμού, εκτός αν ο όρος έχει καθιερωθεί στον πληθυντικό. Στην περίπτωση αυτή η απόδοση είναι στον πληθυντικό και για τα ελληνικά (π.χ. *data - δεδομένα, metadata - μεταδεδομένα*).

Στην περίπτωση των πολυλεκτικών σύμπλοκων όρων της δομής *ουσιαστικό + ουσιαστικό σε γενική*, σε ορισμένες περιπτώσεις η χρήση πληθυντικού φαίνεται να είναι καταλληλότερη για την περίπτωση του δεύτερου ουσιαστικού. Για παράδειγμα, ο όρος *information extraction* αποδίδεται ως *εξαγωγή πληροφορίας* και όχι *πληροφοριών*. Ωστόσο, σε περιπτώσεις όρων όπως *question answering* ή *answer extraction* η απόδοση με πληθυντικό στο εξαρτώμενο ουσιαστικό (*απάντηση ερωτήσεων* και *εξαγωγή απαντήσεων*, αντίστοιχα) είναι ορθότερη. Ο λόγος προτίμησης του πληθυντικού είναι ότι οι συγκεκριμένες διαδικασίες δεν αφορούν συγκεκριμένη πληροφορία (δηλ. απάντηση συγκεκριμένης ερώτησης ή εξαγωγή συγκεκριμένης απάντησης), αλλά τον γενικό στόχο μιας διαδικασίας γλωσσικής τεχνολογίας.

Συχνή ήταν η περίπτωση ενδογλωσσικού αναλογικού σχηματισμού, όπου παρατηρούμε σχηματισμό νέου όρου βάσει ήδη καθιερωμένου όρου - για παράδειγμα, βάσει του καθιερωμένου όρου *machine learning* που αποδίδεται ως *μηχανική μάθηση*, σχηματίζεται και ο όρος *μηχανική κατανόηση* ως απόδοση του *machine understanding* (και όχι *κατανόηση μηχανής* ή *από μηχανή*). Αυτή η τάση παρατηρείται ακόμη και σε περιπτώσεις που η αρχική απόδοση ενδέχεται να είναι προβληματική, π.χ. ο όρος *speech synthesis* έχει επικρατήσει να αποδίδεται ως *σύνθεση φωνής* (ενώ το ορθό θα ήταν *ομιλίας*). Κατ' αναλογία παρατηρείται ο όρος *επισημείωση φωνής (speech annotation),* ενώ είναι σαφές ότι αφορά την ανθρώπινη ομιλία, σε αντιδιαστολή με το *voice recognition* που όντως αφορά αναγνώριση φωνής. Στο λεξικό καταγράφηκαν και οι δύο όροι.

Οι εναλλακτικές γλωσσικές πραγματώσεις που αντιστοιχούν σε μια έννοια καταγράφηκαν ως σύνολο όρων (με επισήμανση του προτιμητέου) σε κάθε γλώσσα (Βαλεοντής & Μάντζαρη, 2006). Δεν επιλέχθηκε για την παρούσα φάση αντιστοίχιση των όρων ένα-προς-ένα από γλώσσα σε γλώσσα. Με αυτήν την έννοια, οι όροι {*annotation, labeling, labelling, tagging}* αποτελούν ένα σύνολο αγγλικών όρων με προτιμητέο τον όρο annotation, και οι όροι *{επισημείωση, επισήμανση, σχολιασμός, ετικετοποίηση}* αποτελούν το αντίστοιχο ελληνικό σύνολο, χωρίς όμως να αντιστοιχίζεται συγκεκριμένα το *annotation* με το *επισημείωση*, το *labeling* με το *επισήμανση* κτλ.

3.6 Τελικός κατάλογος όρων

Με την ολοκλήρωση του ελέγχου, από τις 935 εγγραφές του αρχικού καταλόγου προέκυψε ο τελικός κατάλογος 575 όρων και των συνοδευτικών τους πληροφοριών (εναλλακτικών όρων και στις δύο γλώσσες, ορισμού, πηγής, θεματικού πεδίου κτλ.).

4 Δημοσίευση του λεξικού

Για την καλύτερη αξιοποίησή του κρίνεται σκόπιμο το λεξικό να υλοποιηθεί και δημοσιευθεί σε τυπική μορφή αναπαράστασης σύμφωνα με τα πρότυπα του Σημασιολογικού Ιστού, ώστε να είναι άμεσα αξιοποιήσιμο από μηχανές, να διασυνδεθεί με άλλους πόρους και να χρησιμοποιηθούν υπάρχοντα περιβάλλοντα επεξεργασίας και πρόσβασης σε αυτό.

Για τον σκοπό αυτό, επιλέχθηκε το μοντέλο SKOS[9] που χρησιμοποιείται ευρέως για ελεγχόμενα λεξιλόγια, θησαυρούς, ορολογικά λεξικά κτλ. Σύμφωνα με το μοντέλο, οι εγγραφές (έννοιες) προσδιορίζονται με μοναδικό αναγνωριστικό, αποδίδονται γλωσσικά με λεξικές πραγματώσεις σε μία ή περισσότερες φυσικές γλώσσες, τεκμηριώνονται με διάφορα είδη σχολίων, συνδέονται σημασιολογικά μεταξύ τους σε ιεραρχίες και δίκτυα συσχέτισης και ομαδοποιούνται σε εννοιολογικά μοντέλα.

Για τη μετατροπή του λεξικού από την αρχική μορφή CSV (στην οποία έγιναν οι διαδικασίες ελέγχου και εμπλουτισμού) σε μορφή RDF σύμφωνη με το πρότυπο SKOS, χρησιμοποιήθηκε το λογισμικό VocBench[10] που παρέχει, μεταξύ άλλων, αυτή τη λειτουργία. Το συγκεκριμένο λογισμικό θα χρησιμοποιηθεί και για την επεξεργασία του λεξικού σε μελλοντικές φάσεις.

Τέλος, παρόλο που η οργάνωση του ορολογικού λεξικού με βάση τις έννοιες είναι η πλέον ενδεδειγμένη, για τον σκοπό της απόδοσης στα ελληνικά όλου του υλικού της υποδομής CLARIN:EL, είναι χρήσιμο να αναπαρασταθεί ή/και συνδεθεί με λεξικογραφικό πόρο, ώστε να αντιστοιχισθούν οι μεταφραστικές αποδόσεις σε επίπεδο λέξης/φράσης και όχι έννοιας, καθώς και να συμπληρωθούν με μορφολογική πληροφορία (π.χ. κλίσης) οι εγγραφές. Για τους σκοπούς αυτούς διερευνάται το μοντέλο Ontolex-Lemon[11], το οποίο είναι το πλέον διαδεδομένο μοντέλο για την απόδοση γλωσσικής πληροφορίας σε οντολογίες. Ιδιαίτερο ενδιαφέρον παρουσιάζει η επέκταση του μοντέλου αυτού με εστίαση στην ορολογία, που είναι τώρα υπό σχεδίαση.

5 Μελλοντικά βήματα

Το λεξικό που παρουσιάστηκε αποτελεί το πρώτο στάδιο συγκέντρωσης και συστηματοποίησης της ορολογίας της υποδομής CLARIN:EL. Στην τρέχουσα μορφή οι όροι είναι καταγεγραμμένοι και οργανωμένοι σε σύνολα λέξεων με χαρακτηρισμό των

[9] https://www.w3.org/2004/02/skos/

[10] https://vocbench.uniroma2.it/

[11] https://www.w3.org/2016/05/ontolex/

προτιμητέων και των εναλλακτικών όρων. Στα επόμενα στάδια προβλέπεται η προσθήκη ιεραρχικών και άλλων σχέσεων μεταξύ των όρων. Τόσο η τρέχουσα μορφή του λεξικού όσο και η μελλοντική θα διατίθενται μέσω της υποδομής CLARIN:EL.

Βιβλιογραφία

Gavriilidou, M., Piperidis, S., Galanis, D., Bakagianni, J., Labropoulou, P., Kolovou, A., Gkoumas, D., Deligiannis, M., Pouli, K., Tsiouli, I., Voukoutis, L., & Gkirtzou, K. (υπό έκδοση). The CLARIN:EL infrastructure. Στο *CLARIN Annual Conference 2023, Leuven.*

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, St., Papageorgiou, H., Monachini, M. Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., & Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. Στο *Eighth International Conference on Language Resources and Evaluation, Istanbul, 21-27 May.* http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf

International Organization for Standardization. (2022). *Terminology Work - Principles and Methods* (ISO 704:2022).

Labropoulou, P., Gavriilidou, M., Galanis, D., Bakagianni, J., Deligiannis, M., Pouli, K., Gkirtzou, K., Voukoutis, L., & Tsiouli, I. (2020). *CLARIN-SHARE metadata schema.* CLARIN META-SHARE Working Group. https://clarin-platform-documentation.readthedocs.io/en/stable/all/5_Metadata/Full.html#fullschema

Labropoulou, P., Nedellec, C., Galanis, D., Knoth, P., Aubin, S., Villegas, M., Giagkou, M., Gkirtzou, K. (2022). *OMTD-SHARE ontology. Pre-release 2.0.0.* http://w3id.org/meta-share/omtd-share/

Pouli, K., Bakagianni, J., Galanis, D., Labropoulou, P., Tsiouli, I., & Gavriilidou, M. (2023). *The CLARIN:EL User Manual, v.1.0.* https://clarin-platform-documentation.readthedocs.io/el/stable/

Βαλεοντής Κ., & Μάντζαρη Ε. (2006). Η γλωσσική διάσταση της Ορολογίας: Αρχές και μέθοδοι σχηματισμού των όρων. Στο *1st Athens International Conference on Translation and Interpretation. Translation: Between Art and Social Science, 13-14 October*. https://www.eleto.gr/download/BooksAndArticles/HAU-Conference2006-ValeontisMantzari_GR.pdf

Γούτσος, Δ., & Φραγκάκη, Γ. (2015). *Εισαγωγή στη γλωσσολογία σωμάτων κειμένων*. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. https://hdl.handle.net/11419/1932

Ελληνικός Οργανισμός Τυποποίησης. (2010). *Ορολογική εργασία – Αρχές και μέθοδοι* (ΕΛΟΤ 402:2010).

Καπιδάκης, Σ., Πιπερίδης, Σ., Λαμπροπούλου, Π., & Γαβριηλίδου, Μ. (2015). Ανοιχτά Γλωσσικά Δεδομένα: η Υποδομή Γλωσσικών Πόρων και Υπηρεσιών clarin:el. Στο *2ο Διεθνές Συνέδριο «Δημιουργική Γραφή», Κέρκυρα, 1-4 Οκτωβρίου: Πρακτικά, τ. Β'* (σ. 1073-1088). http://cwconference.web.uowm.gr/archives/2nd_cw_conference_volume_2.pdf

Καρασίμος, Α., Μαρκόπουλος, Γ., Σγαρμπάς, Κ., & Χριστοφίδου, Α. (2015). Ορολογία Υπολογ/ικής Γλωσσολογίας. *Δελτίο Επιστημονικής Ορολογίας & Νεολογισμών*, 13.

Μικρός, Γ. (2015). *Υπολογιστική υφολογία.* Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. https://hdl.handle.net/11419/4860

Παναγιωτακόπουλος, Χ., Τσαλίδης, Χ., Γάκης, Π., & Κόκκινος, Θ. (2022). *Υπολογιστική γλωσσολογία.* Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. https://dx.doi.org/10.57713/kallipos-127

Παναγιωτάκος, Δ. (2023). *Λεξικό Τεχνολογίας Πληροφοριών και Επικοινωνιών, Αγγλοελληνικό ορολογικό λεξικό* (τ. 1). Ελληνική Εταιρεία Ορολογίας (ΕΛΕΤΟ). https://www.eleto.gr/download/Bodies/Lexic%CE%BF-ICT_Volume-One.pdf

Πιπερίδης, Στ., Λαμπροπούλου, Π., & Γαβριηλίδου, Μ. (2015). CLARIN EL: Δημιουργώ, Επεξεργάζομαι, Μοιράζομαι. Στο *10ο Διεθνές Συνέδριο «Ελληνική Γλώσσα και Ορολογία»*, *Αθήνα, 12-14 Νοεμβρίου.* http://www.eleto.gr/download/Conferences/10th%20Conference/Papers-and-speakers/10th_25-22-10_SPiperidis-PLabropoulou-MGavriilidou_Paper_V03.pdf

Πιπερίδης, Στ., Λαμπροπούλου, Π., & Γαβριηλίδου, Μ. (2017). clarin:el Υποδομή Τεκμηρίωσης, Διαμοιρασμού και Επεξεργασίας Γλωσσικών Δεδομένων. Στο Τ. Georgakopoulos, T. Pavlidou, M. Pechlivanos, A. Alexiadou, J. Androutsopoulos, A. Kalokairinos, S. Skopeteas, & K. Stathi (Επιμ.), *12ο Διεθνές Συνέδριο Ελληνικής Γλωσσολογίας (ICGL12), 16-19 September 2015, Berlin: Πρακτικά*, τ. 2, (σ. 851–869). http://www.cemog.fu-berlin.de/en/icgl12/offprints/piperidis-lampropoulou-gavriilidou/icgl12_Piperidis-et-al.pdf

Τάντος, Α., Μαρκαντωνάτου, Σ., Αναστασιάδη-Συμεωνίδη, Ά., & Κυριακοπούλου, Π. (2015). *Υπολογιστική γλωσσολογία* Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. https://hdl.handle.net/11419/2205

Μαρία Γαβριηλίδου

ΕΛΕ Α’, Ινστιτούτο Επεξεργασίας του Λόγου

Ηλ-ταχ.: maria@athenarc.gr

Πένυ Λαμπροπούλου

ΕΛΕ Α’, Ινστιτούτο Επεξεργασίας του Λόγου

Ηλ-ταχ.: penny@athenarc.gr

Κανέλλα Πουλή

Συνεργαζόμενη Ερευνήτρια, Ινστιτούτο Επεξεργασίας του Λόγου

Ηλ-ταχ.: kanella@athenarc.gr

Ηρώ Τσιούλη

Συνεργαζόμενη Ερευνήτρια, Ινστιτούτο Επεξεργασίας του Λόγου

Ηλ-ταχ.: tsiouli@athenarc.gr