«ΣΥΓΚΕΝΤΡΩΣΗ ΚΑΙ ΚΑΤΑΓΡΑΦΗ ΕΛΛΗΝΙΚΩΝ ΓΛΩΣΣΙΚΩΝ ΠΟΡΩΝ ΤΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ (ΠΑΔΑ)»

Αγγελική Μπαμνιώτη, *Βιβλιοθηκονόμος, MA, MSc, Εθνική Βιβλιοθήκη της Ελλάδος, Τμ. Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης, ΠαΔΑ*

Σαράντος Καπιδάκης, *Καθηγητής,* Τμ. Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης, ΠαΔΑ

Εισαγωγή

Οι γλωσσικοί πόροι είναι σύνολα δεδομένων που χαρακτηρίζονται από ποικίλες μορφές. Μπορεί να είναι δομημένοι ή αδόμητοι, πρωτογενείς, επεξεργασμένοι, σε οργανωμένη μορφή ή σχετιζόμενοι με εργαλεία γλωσσικών τεχνολογιών. Ορισμένα χαρακτηριστικά παραδείγματα είναι τα σώματα κειμένου, τα γλωσσάρια, τα μονόγλωσσα ή πολύγλωσσα λεξικά, οι θησαυροί κτλ. Αποτελούν σημαντική παρακαταθήκη της γλώσσας μας αλλά και εργαλεία για τη διαμόρφωση και εκπαίδευση έξυπνων διαδικασιών καθώς και την ανάπτυξη τεχνολογιών γλωσσικής επεξεργασίας, όπως είναι για παράδειγμα η γραμματική ή συντακτική ανάλυση, η αυτόματη μετάφραση, η κατανόηση κειμένου, η εξαγωγή περιλήψεων, κλπ. Η περιγραφή τους γίνεται με ιδιαίτερο τρόπο που τους επιτρέπει να είναι ανακτήσιμοι και εκμεταλλεύσιμοι σε υπολογιστικό περιβάλλον. Στόχος του παρόντος άρθρου είναι να διερευνήσει ποιοι είναι οι γλωσσικοί πόροι που έχουν παραχθεί μέσα στα πλαίσια της λειτουργίας και εκπαιδευτικού και ερευνητικού έργου του Πανεπιστημίου Δυτικής Αττικής, από τους διδάσκοντες ή ενδεχομένως και τους φοιτητές του, ανά τμήμα και σχολή. Θα διερευνηθεί ποια είναι η μορφή και το περιεχόμενο των σχετικών πόρων, καθώς και με ποιους όρους είναι διαθέσιμοι στο κοινό από τους δημιουργούς τους. Στη συνέχεια, θα καταγραφούν με κατάλληλο τρόπο και αναφορτωθούν (ή συνδεθούν) με το αντίστοιχο ιδρυματικό αποθετήριο του Πανεπιστημίου Δυτικής Αττικής στην ελληνική εκδοχή του Clarin. Το Clarin είναι μια ευρωπαϊκή διαδικτυακή υποδομή που συσσωρεύει γλωσσικούς πόρους, τεχνολογίες και υπηρεσίες, σε διάφορες γλώσσες, ώστε να τους διαθέσει προς την ερευνητική κοινότητα γενικότερα και την κοινότητα της γλωσσολογίας ειδικότερα αλλά και τον απλό ιδιώτη, προς προώθηση της γνώσης και επεξεργασία του υλικού μέσω διαφόρων γλωσσικών τεχνολογιών (https://www.clarin.eu/). Σημαντικός είναι ο αριθμός των πανεπιστήμιων και ερευνητικών κέντρων της Ελλάδας που διαθέτουν ήδη ψηφιακό αποθετήριο στο Clarin:el, το οποίο φιλοξενεί τους παραγόμενους γλωσσικούς τους πόρους. Στο συγκεκριμένο άρθρο επιχειρείται η περιγραφή της διαδικασίας της δημιουργίας και του εμπλουτισμού ανάλογου ψηφιακού αποθετηρίου, μέσα στην υποδομή του Clarin:el, για το Πανεπιστήμιο Δυτικής Αττικής (https://www.clarin.eu/). Η εργασία περιλαμβάνει και περιγράφει αναλυτικά όλα τα στάδια της διερεύνησης, καταγραφής και συγκέντρωσης του υλικού, καθώς και την επικοινωνία με τους σχετικούς δημιουργούς των γλωσσικών πόρων, προς διευκρίνιση, μεταξύ άλλων, των όρων διάθεσης και μεταφόρτωσης τους στη γλωσσική υποδομή του Clarin:el. Επίσης παραθέτει τη μορφή των συλλεγόμενων πόρων και των μεταδεδομένων τους.

Κίνητρο και σκοπός της έρευνας

Μεγάλος αριθμός πανεπιστημιακών ιδρυμάτων και ερευνητικών κέντρων της Ελλάδας διαθέτουν ιδρυματικό ψηφιακό αποθετήριο στη διαδικτυακή υποδομή του Clarin:el. Η συγκεκριμένη διασύνδεση τους επιτρέπει να προβάλλουν το έργο τους, να οργανώνουν τους γλωσσικούς τους πόρους σε ένα ψηφιακό, υπολογιστικό περιβάλλον, να διασυνδέονται με την επιστημονική κοινότητα ή να γίνονται μέλη μιας ευρύτερης κοινωνίας που επικοινωνεί και που υφίσταται σε διεθνές επίπεδο, με πολλά οφέλη και προνόμια στην προώθηση της γνώσης. Στο συγκεκριμένο άρθρο περιγράφεται η διασύνδεση του Πανεπιστημίου Δυτικής Αττικής με τη διεθνή γλωσσική υποδομή του Clarin, μέσω του ελληνικού της παραρτήματος, που είναι το Clarin:el.

Οι στόχοι της έρευνας περιλαμβάνουν τη συγκέντρωση των γλωσσικών πόρων που έχουν παραχθεί από το Πανεπιστήμιο Δυτικής Αττικής στα πλαίσια του ερευνητικού και εκπαιδευτικού του έργου, την καταγραφή τους, τη διασαφήνιση των όρων με τους οποίους παρέχονται και χρησιμοποιούνται από τους χρήστες, την οργάνωση και ταξινόμηση ώστε να είναι ανακτήσιμοι από την επιστημονική κοινότητα και το ενδιαφερόμενο κοινό, με τα ανάλογα δικαιώματα, τη διασύνδεση και τεκμηρίωση των πόρων μέσω του Clarin και εν τέλη τον εμπλουτισμό και οργάνωση του ψηφιακού αποθετηρίου του Πανεπιστημίου Δυτικής Αττικής στην υποδομή του Clarin:el.

Ορισμοί-Βιβλιογραφική επισκόπηση

Το *Clarin* είναι μια ευρωπαϊκή διαδικτυακή υποδομή η οποία συγκεντρώνει γλωσσικούς πόρους, τεχνολογίες και υπηρεσίες, σε διάφορες γλώσσες, με σκοπό να τους διαθέσει προς την ερευνητική κοινότητα ή και τον απλό ιδιώτη. Είναι οργανωμένο σε κατά τόπους και γλώσσα υποδομές του Clarin. Το υλικό που συσσωρεύει τεκμηριώνεται κατάλληλα και μπορεί να είναι επεξεργάσιμο μέσω γλωσσικών τεχνολογιών. Παρέχει επίσης εκπαίδευση αναφορικά με τις γλωσσικές τεχνολογίες και οργανώνει παγκοσμίως διάφορες δράσεις. Το υλικό από τα κατά τόπου παραρτήματα του, συλλέγεται και διασυνδέεται με την κεντρική υποδομή του Clarin (https://www.clarin.eu/).

Το *CLARIN:EL* είναι το ελληνικό παράρτημα της ευρωπαϊκής υποδομής Clarin. Περιέχει κεντρικό κατάλογο, όπου εμφανίζονται προς το κοινό όλοι οι γλωσσικοί πόροι που φιλοξενεί. Οι πόροι είναι οργανωμένοι ανά ψηφιακό αποθετήριο. Συνολικά περιέχει μέχρι στιγμής δεκατρία ψηφιακά αποθετήρια, οργανωμένα ανά τον ανάλογο φορέα που εκπροσωπούν. Περιλαμβάνει οκτώ ακαδημαϊκά αποθετήρια, τέσσερα αποθετήρια ερευνητικών κέντρων και ένα αποθετήριο φιλοξενούμενων πόρων, στο οποίο εντάσσονται μεμονωμένοι πόροι που δεν συσχετίζονται με κάποιο συγκεκριμένο ψηφιακό αποθετήριο από τα προϋπάρχοντα. Η ελληνική υποδομή του Clarin έχει περισσότερους από 1.350 εγγεγραμμένους χρήστες και φιλοξενεί συνολικά, μέχρι στιγμής, 799 γλωσσικούς πόρους (https://www.clarin.gr/).

Υπάρχουν και άλλες *παρεμφερείς υποδομές*, με διεθνή ή/και ελληνική διάσταση που συλλέγουν γλωσσικούς πόρους, ψηφιοποιημένο ή ψηφιακό υλικό αναφορικά με την παγκόσμια πολιτιστική κληρονομιά και προάγουν τις ανθρωπιστικές επιστήμες. Ενδεικτικά θα αναφέρουμε το Dariah.eu (https://www.dariah.eu/), με την ελληνική εθνική υποδομή Dariah-gr/ ΔΥΑΣ (https://dyas-net.gr/), την Europeana.eu που χρηματοδοτείται από την Ευρωπαϊκή Ένωση και συλλέγει και φιλοξενεί την ψηφιακή πολιτιστική κληρονομιά της Ευρώπης (https://www.europeana.eu/) ή την Απολλωνίς (https://apollonis-infrastructure.gr/). Οι συγκεκριμένες υποδομές προάγουν σαφώς την ανοιχτή πρόσβαση (Καπιδάκης, 2014).

Αποθετήρια	Αριθμός γλωσσικών πόρων
Ερευνητικό Κέντρο Αθηνά	340
Πανεπιστήμιο Δυτικής Αττικής	193
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης	52

Πανεπιστήμιο Αιγαίου	45
Κέντρο Ελληνικής Γλώσσας	39
Πανεπιστήμιο Κρήτης	31
Ιόνιο Πανεπιστήμιο	26
Αποθετήριο Φιλοξενούμενων Πόρων	18
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών	17
Εθνικό Κέντρο Κοινωνικών Ερευνών	13
Οικονομικό Πανεπιστήμιο Αθηνών	12
ΕΚΕΦΕ Δημόκριτος	8
Πάντειο Πανεπιστήμιο	5

Πίνακας 1. Περιεχόμενο των ψηφιακών αποθετηρίων του Clarin:EL

Ως *γλωσσικός / ορολογικός πόρος* ορίζεται οποιοδήποτε σύνολο δεδομένων, σε κάθε μορφή, δομημένο ή αδόμητο, που συνδέεται με τη γλώσσα. Μπορεί να έχει περιεχόμενο πρωτογενές, επεξεργασμένο ή οργανωμένη μορφή γνώσης (μονόγλωσσα ή πολύγλωσσα λεξικά, γλωσσάρια, λίστες λέξεων, θησαυροί κτλ.), καθώς και διάφορες εφαρμογές και εργαλεία γλωσσικής τεχνολογίας (εργαλεία λογισμικού σε κείμενα, εργαλεία εξόρυξης γνώσης, λημματοποίησης, παρουσίασης δεδομένων κτλ.) (https://www.clarin.gr/). Η συλλογή και καταγραφή γλωσσικών πόρων συμβάλει σημαντικά στην εξέλιξη των γλωσσικών τεχνολογιών, οι οποίες αναπτύσσουν εργαλεία και εφαρμογές γλωσσικής ανάλυσης και επεξεργασίας.

Οι *γλωσσικές τεχνολογίες* είναι διάφορα υπολογιστικά εργαλεία γλωσσικής ανάλυσης μέσω των οποίων μπορούν να πραγματοποιηθούν ενέργειες όπως η ανάλυση, επισημείωση, επεξεργασία και τροποποίηση των διαφόρων γλωσσικών / ορολογικών δεδομένων (https://www.clarin.gr/).

Αντίστοιχα, οι *υπηρεσίες γλωσσικής επεξεργασίας* επιτρέπουν τη χρήση των γλωσσικών πόρων και τεχνολογιών, όπως επίσης και των εφαρμογών αυτών στο διαδίκτυο (https://www.clarin.gr/).

Οι *ψηφιακές βιβλιοθήκες* είναι οντότητες που καθιστούν διαθέσιμους στο κοινό πόρους και ψηφιακές ή ψηφιοποιημένες συλλογές τεκμηρίων, οι οποίες έχουν επιλεγεί, οργανωθεί, τεκμηριωθεί, διανεμηθεί και συντηρηθεί από εξειδικευμένο προσωπικό. Στόχος των ψηφιακών βιβλιοθηκών είναι η παροχή πληροφοριακών υπηρεσιών σε ψηφιακό περιβάλλον, οι οποίες είναι εφάμιλλες με αυτές των συμβατικών βιβλιοθηκών (Κυριάκη – Μάνεση και Κουλούρης, 2015). Η σημαντικότητα των ψηφιακών βιβλιοθηκών έγκειται στο ότι οργανώνουν το ψηφιακό τους περιεχόμενο με τρόπο ανεξάρτητο από την προσβασιμότητα του (Καπιδάκης, 2014).

Επέκταση ή γενικά εξέλιξη των ψηφιακών βιβλιοθηκών αποτελούν τα *ψηφιακά αποθετήρια*. Στα αποθετήρια κατατίθεται ψηφιακό υλικό και δεδομένα για φύλαξη και διάχυση στο διαδίκτυο. Τα ακαδημαϊκά ιδρύματα της χώρας μας διαθέτουν ψηφιακά αποθετήρια, όπου οι ερευνητές, τα μέλη του

διδακτικού προσωπικού, οι φοιτητές και οι λοιποί εμπλεκόμενοι με το έκαστο πανεπιστήμιο, μπορούν ή οφείλουν να καταθέσουν τη συγγραφική και ερευνητική τους παραγωγή για φύλαξη, κρίση του έργου, τεκμηρίωση, παραγωγή μεταδεδομένων, διάχυση στο διαδίκτυο, προβολή, προαγωγή της έρευνας και χρήση, με άμεση σύνδεση με την ανοιχτή πρόσβαση (Shirley, 2005). Τα ψηφιακά αποθετήρια τα οποία έχουν δημιουργηθεί στο Clarin:el, βρίσκονται σε αναφορά και σύνδεση με το εκάστοτε ακαδημαϊκό ίδρυμα ή ερευνητικό κέντρο που διαμοιράζει το υλικό του και διαθέτουν χαρακτηριστικά που διέπουν ένα ιδρυματικό αποθετήριο.

Με την πρόοδο της τεχνολογίας, μια σημαντική εξέλιξη είναι οι *ψηφιακές ανθρωπιστικές επιστήμες* οι οποίες αξιοποιούν ψηφιακούς πόρους και εργαλεία προάγοντας την έρευνα στον τομέα των ανθρωπιστικών σπουδών (Γούτσος και Φραγκάκη, 2015). Οι ψηφιακές ανθρωπιστικές επιστήμες αφορούν το σύνολο των ανθρωπιστικών και κοινωνικών επιστημών, των επιστημών της τέχνης και της φιλολογίας (Schreibman, Siemens and Unsworth, 2004).

Η *υπολογιστική γλωσσολογία* αποτελεί διεπιστημονικό πεδίο της πληροφορικής και της γλωσσολογίας που ασχολείται με την επεξεργασία της φυσικής γλώσσας Τάντος κ.ά., 2015). Διάφορα πεδία της γλωσσολογίας, όπως είναι η μορφολογία, η σύνταξη, η φωνολογία κτλ, μπορούν να υποστούν επεξεργασία η οποία έχει ως σκοπό τη δημιουργία διαφόρων υπολογιστικών εφαρμογών μέσω των οποίων οι ηλεκτρονικοί υπολογιστές θα είναι σε θέση να αναγνωρίσουν, επεξεργαστούν και παράγουν τη φυσική γλώσσα (Γούτσος και Φραγκάκη, 2015). Στην περίπτωση των γλωσσικών πόρων που διαμοιράζονται μέσω του Clarin:el και της γλωσσικής επεξεργασίας που δύνανται να υποστούν, μπορούμε να διακρίνουμε καθαρά εφαρμογές που εμπίπτουν στον τομέα της υπολογιστικής γλωσσολογίας.

Η *τεχνητή νοημοσύνη* αναφέρεται στην ικανότητα των μηχανών, μέσω της τεχνολογίας, να αναπαράγουν διάφορες λειτουργίες που προσομοιάζουν αυτές που παράγονται από το ανθρώπινο νου. Η τεχνητή νοημοσύνη, μπορεί να έχει πρακτική εφαρμογή στις γλωσσικές υπηρεσίες μέσω της χρήσης και εκπαίδευσης των κατάλληλων, εξειδικευμένων αλγορίθμων (Γαβριηλίδου, Πιπερίδης, 2021). Το Clarin:el συλλέγει τους ανάλογους γλωσσικούς πόρους, οι οποίοι, μέσω γλωσσικών εργαλείων και επεξεργασίας συμβάλουν στην ανάπτυξη εφαρμογών που βασίζονται στην τεχνητή νοημοσύνη και τις εξελίξεις της.

Μεθοδολογία

Το εγχείρημα της συγκέντρωσης και καταγραφής γλωσσικών πόρων που παράχθηκαν στα πλαίσια της ακαδημαϊκής λειτουργίας του Πανεπιστήμιου Δυτικής Αττικής είναι αρκετά σύνθετο και περιλαμβάνει τόσο θεωρητικό πλαίσιο, το οποίο αφορά στη συγκέντρωση και διερεύνηση του υλικού, την οργάνωση της επικοινωνίας με τους δημιουργούς και την τεκμηρίωση και διάθεση μέσω του συστήματος, όσο και πρακτική εφαρμογή πάνω στην υποδομή του Clarin:el.

Ακολουθήθηκαν τα εξής στάδια:

- *Διασύνδεση με το Clarin:el*

Μέσω ακαδημαϊκού λογαριασμού συνδεθήκαμε στο αποθετήριο του Πανεπιστημίου Δυτικής Αττικής στην υποδομή του Clarin:el, το οποίο υπήρχε στο σύστημα αλλά ήταν άδειο και αδρανές. Στη συνέχεια θέσαμε τα στάδια και τις διαδικασίες-βήματα που θα έπρεπε να ακολουθηθούν ώστε να πραγματοποιηθεί η αναζήτηση των ανάλογων γλωσσικών πόρων και ο εμπλουτισμός του αποθετηρίου.

- *Αναζήτηση γλωσσικών πόρων*

Ασχοληθήκαμε με την έμπρακτη αναζήτηση γλωσσικών πόρων που παράχθηκαν από το ΠαΔΑ. Ήρθαμε σε επικοινωνία με τους πιθανούς δημιουργούς γλωσσικών πόρων μέσω ηλεκτρονικού ταχυδρομείου ή με συνάντηση μέσω teams ή δια ζώσης. Συντάξαμε ένα έγγραφο συχνών ερωταπαντήσεων για καλύτερη πληροφόρηση και ενημέρωση για το εγχείρημα που απευθυνόταν στους παραγωγούς γλωσσικών πόρων, από τους οποίους αναζητήσαμε υλικό. Επιμείναμε και ξαναήρθαμε περισσότερες από μια φορές σε επαφή μαζί τους. Καθορίσαμε ότι το υλικό θα διατίθεται με ανοιχτές άδειες χρήσης Creative Commons.

- *Υλοποίηση*

Έχοντας συγκεντρώσει το απαραίτητο υλικό, προχωρήσαμε στην έμπρακτη υλοποίηση του εγχειρήματος της καταγραφής και διασύνδεσης του με την εφαρμογή του Clarin:el. Το υλικό χρειάστηκε να προετοιμαστεί κατάλληλα. Συλλέξαμε μόνο κατάλληλο υλικό, σε μορφότυπους που υποστηρίζει το Clarin:el. Οι πόροι κατηγοριοποιήθηκαν κι οργανώθηκαν με συνάφεια και συμπιέστηκαν για τη μεταφόρτωση τους στην υποδομή, χρησιμοποιώντας τους κατάλληλους μορφότυπους. Η τεκμηρίωση τους πραγματοποιήθηκε χρησιμοποιώντας τα κατάλληλα μεταδεδομένα, ώστε να προβούμε στην ορθότερη περιγραφή των γλωσσικών πόρων.

Περιγραφή Υλοποίησης – Εφαρμογής

Το γλωσσικό υλικό που μπορεί να μεταφορτωθεί στο Clarin κατατάσσεται σε τέσσερις κατηγορίες, οι οποίες είναι οι εξής:

• Εργαλεία ή υπηρεσίες γλωσσικής επεξεργασίας

• Σώματα κειμένου

• Γλωσσικές περιγραφές και υπολογιστικά μοντέλα

•Λεξικά, γλωσσάρια, θησαυροί, οντολογίες ή λίστες λέξεων, φράσεων κτλ., που αποτελούν τους εννοιολογικούς πόρους

Το υλικό που συλλέχθηκε από την παραγωγή γλωσσικών πόρων του ΠαΔΑ ανταποκρίνεται μόνο σε δύο κατηγορίες.

Πρόκειται για τα σώματα κειμένου και τα λεξικά / γλωσσάρια. Η καθεμία από αυτές τις δύο κατηγορίες γλωσσικού υλικού έχει διαφορές στα μεταδεδομένα και τον τρόπο περιγραφής της. Ως σώματα κειμένου νοούνται οι συλλογές από πρωτογενή ή επεξεργασμένα δεδομένα σε διάφορα μέσα. Στην περίπτωση της παρούσας έρευνας, το υλικό που συλλέχτηκε αποτελείται σχεδόν αποκλειστικά από ψηφιακά και ψηφιοποιημένα κείμενα γραπτού λόγου, σε διάφορες μορφές: άρθρα, εργασίες, υλικό από μαθήματα του ΠαΔΑ, περιγραφές μαθημάτων, ακαδημαϊκά συγγράμματα, περιγραφές συνεδρίων, εισηγήσεις κτλ. (Clarin:el, 2020).

Τα λεξικά ή γλωσσάρια όρων και ορολογίας είναι δομημένα γλωσσικά δεδομένα, όπως οι λίστες και κατάλογοι λέξεων, οι θησαυροί, ή άλλοι εννοιολογικοί πόροι (Clarin:el, 2020).

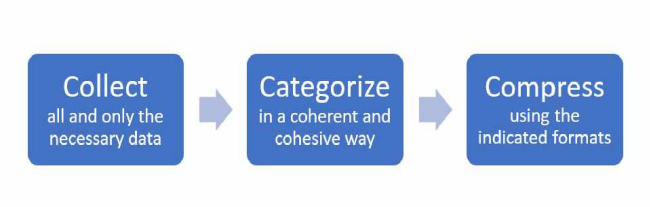
Όσον αφορά τις κατηγορίες του υλικού που συλλέξαμε, κανονικά οι πόροι θα πρέπει να είναι σε μορφή XML ή TXT, ενώ, η υποδομή σωρεύει και κείμενα σε PDF και MS-Word, αφού βρίσκεται σε φάση δρομολόγησης μιας υπηρεσίας που μετατρέπει τα αρχεία σε TXT (Clarin:el, 2020). Οπότε, στην παρούσα φάση, μόνο το υλικό στους συγκεκριμένους μορφότυπους είναι συμβατό με το Clarin για τις δύο

κατηγορίες πόρων που συλλέξαμε. Για το λόγο αυτό, αποφύγαμε να μεταφορτώσουμε πόρους αμιγώς σε άλλους μορφότυπους (πχ PPT, TIFF) που βρήκαμε. Αντίθετα, κάποιοι πόροι σε PPT μεταφορτώθηκαν όταν ήταν μέρος μικτού υλικού που περιλάμβανε και επεξεργάσιμο υλικό σε TXT, Word ή PDF. Ακόμα, σε κάποιες περιπτώσεις που δεν ήταν πολύ δύσκολο και που δεν υποβαθμιζόταν η ποιότητα του υλικού, πραγματοποιήσαμε μετατροπή του κειμένου σε TXT.

	Recommended	Acceptable
CLARIN:EL processable data	**Monolingual textual data**: plain text **Monolingual encoded data**: XCES-ILSP variant (XML based format compliant with the XCES model for corpora) **Bi-/Multilingual encoded data**: TMX (XML based format for aligned data), MOSES (text-based format for parallel data)	
Textual Data	**File Formats**: plain text **Formatted/Encoded**: ODT, DOCX, PDF/A, HTML, Latex, TeX, MOSES	PDF, SGML, Rich Text Format (.rtf), Microsoft Word (.doc, .docx), PostScript
Text Annotation	**File Formats**: XML, XMI, CSV, TSV, RDF (all serialisation formats RDF/XML, Turtle, Notation3, N-Triples, TriG, N-Quads, JSON-LD, HDT), JSON	SGML, Plain Text, Microsoft Excel (.xlsx, .xls), ELLOGON
	Models: XCES for corpora and structural annotation, TEI for structural and linguistic annotation, GrAF linguistic annotation, TMX for aligned, GATE linguistic annotation, CoNLL family (CoNLL-U, CoNLL-2000, CoNLL-2002, CoNLL-2003, CoNLL-2006, CoNLL-2008, CoNLL-2009, CoNLL-2012) for linguistic annotation, NIF linguistic annotation for RDF data, WARC for web crawled data	
Language Description	**ML Model**: H5, ProtoBuf, ONNX, PMML, Pickle, MLeap, YAML, JSON **N-gram model**: ARPA	
Lexical/Conceptual Resource	**File Formats**: XML, CSV, TSV, RDF (RDF/XML, Turtle, Notation3, N-Triples, TriG, N-Quads, JSON-LD, HDT), OWL	Microsoft Excel (.xlsx, .xls), Plain Text, SQL
	Models: LMF for lexica, OWL for ontologies, SKOS for thesauri, OntoLex-Lemon for lexica, TBX for terminological data	
Image data	**All images**: TIFF, SVG, JPEG 2000, PNG, GIF **Scanned images**: PDF/A	JPEG, BMP, Photoshop, NifTi, FlashPix, PDF
Audio data	WAV, AIFF, FLAC	MP3, MPEG, Windows Media Audio
Video data	AVI	MPEG-4, RealNetworks 'Real Video', Windows Media Video, Flash Video, QuickTime Video

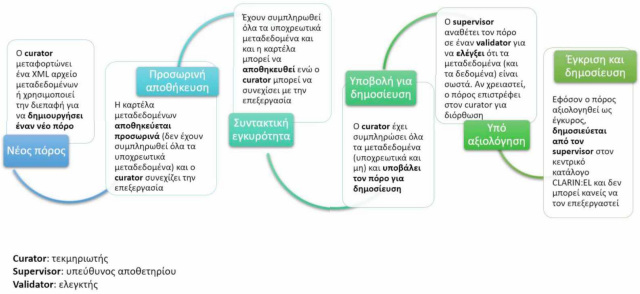
Εικόνα 1. Συνιστώμενοι μορφότυποι. Πηγή: https://www.clarin.gr/

Το υλικό δεν είναι δυνατόν να μεταφορτωθεί απευθείας στη μορφή που βρίσκεται. Τα αρχεία, ένα ή περισσότερα, είναι αναγκαίο να βρίσκονται σε έναν συμπιεσμένο φάκελο σε έναν από τους ακόλουθους μορφότυπους .zip, .tgz, .gz, .tar. Εμείς επιλέξαμε να τους συμπιέσουμε σε μορφότυπο .zip.



Εικόνα 2. Προετοιμασία του υλικού. Πηγή: https://www.clarin.gr/

Στη συνέχεια, πραγματοποιήθηκε η μεταφόρτωση και τεκμηρίωση στο Clarin των δύο προαναφερόμενων μορφών υλικού, δηλαδή των σωμάτων κειμένου και των γλωσσάριων που συλλέξαμε. Υπάρχουν κάποιοι ρόλοι, που καταλαμβάνουν διαφορετικά άτομα στην προετοιμασία, υποβολή και δημοσίευση ενός γλωσσικού πόρου. Πρόκειται για τον curator (τεκμηριωτή), metadata validator (ελεγκτή μεταδομένων), legal validator (νομικό ελεγκτή) και το supervisor (υπεύθυνο αποθετηρίου), οι οποίοι ελέγχουν τα διαφορετικά στάδια της πορείας του πόρου προς τη δημοσίευση του στο αποθετήριο.



Εικόνα 3. Στάδια δημοσίευσης των πόρων. Πηγή: https://www.clarin.gr/

Το Clarin:el και το σχήμα μεταδεδομένων που υποστηρίζει ανταποκρίνονται στις αρχές FAIR:

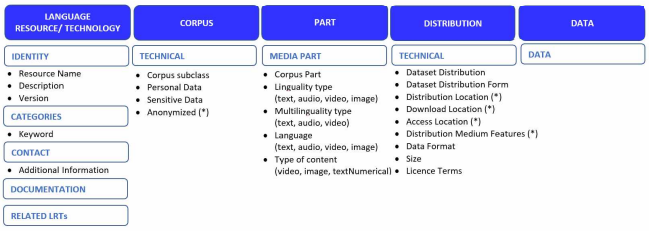
Findability (ευρεσιμότητα)

Accessibility (προσβασιμότητα)

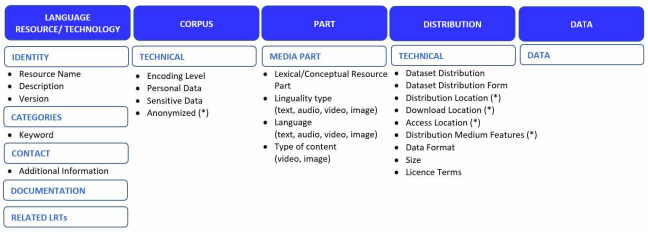
Interoperability(διαλειτουργικότητα)

Reuse (επαναχρησιμοποίηση

Υπάρχουν κάποια υποχρεωτικά μεταδεδομένα και άλλα προαιρετικά, ανάλογα με την κατηγορία του υλικού, με σκοπό την καλύτερη περιγραφή του. Τα κοινά υποχρεωτικά μεταδεδομένα που πρέπει να έχουν στην περιγραφή τους όλοι οι πόροι, όποια και αν είναι η μορφή τους, (Clarin:el, 2020) αφορούν το όνομα του γλωσσικού πόρου, την περιγραφή του περιεχόμενου του, το αναγνωριστικό του (LRT identifier), τον αριθμό εκδοχής (version), τις λέξεις κλειδιά (keywords), την επαφή όπου κάποιος μπορεί να λάβει περισσότερες πληροφορίες σχετικά με τον πόρο, τον τρόπο διανομής (distribution), δηλαδή τη μορφή του υλικού και τέλος τους όρους χρήσης (licence terms).



Εικόνα 4. Υποχρεωτικά μεταδεδομένα για την τεκμηρίωση γλωσσικών πόρων σε μορφή σώματος κειμένου. Πηγή: https://www.clarin.gr/



Εικόνα 5. Υποχρεωτικά μεταδεδομένα για την τεκμηρίωση γλωσσικών πόρων σε μορφή λεξικο-εννοιολογικών πόρων. Πηγή: https://www.clarin.gr/

Πέρα από τα υποχρεωτικά μεταδεδομένα, τα οποία χρησιμοποιήθηκαν στην τεκμηρίωση και περιγραφή των γλωσσικών πόρων του ΠαΔΑ στο Clarin, συμπληρώσαμε κάποια προαιρετικά μεταδεδομένα, με στόχο να γίνει πληρέστερη περιγραφή. Μερικά από τα πεδία αυτά αφορούν το δημιουργό του γλωσσικού πόρου, το τομέα στον οποίο αναφέρεται το υλικό, στον τύπο κειμένου και περιεχομένου, καθώς και τη γεωγραφική και χρονική κάλυψη του περιεχομένου του πόρου.

Στο διαμοιρασμό των πόρων του ΠαΔΑ στο Clarin:el ακολουθήσαμε την αρχική άδεια που είχε ο πόρος όταν πρωτοδημοσιεύτηκε. Πρόκειται για δύο άδειες Creative Commons, την CC-BY-NC-SA (επιτρέπει μη εμπορική χρήση) και την CC-BY-NC-ND (επιτρέπει μη εμπορική χρήση του πόρου, άλλα χωρίς να μπορούν να διανεμηθούν τα παράγωγα του) (https://creativecommons.org/).

Αποτελέσματα – Ευρήματα

Οι περισσότεροι από τους γλωσσικούς πόρους του ΠαΔΑ που συλλέχθηκαν σε μορφή σώματος κειμένου ή λεξικο εννοιολογικού πόρου, εκτός από ελάχιστους που είναι στα αγγλικά, είναι στην ελληνική γλώσσα. Κάποιοι λίγοι περιέχουν δίγλωσσο υλικό, κυρίως αποδόσεις ορολογίας στα αγγλικά. Τα σώματα κειμένου που συλλέξαμε αποτελούνται αποκλειστικά από πρωτογενές υλικό. Το επίπεδο κωδικοποίησης των λεξικο-εννοιολογικών πόρων είναι σημασιολογία κατά πλειοψηφία ή μορφολογία. Όλοι οι πόροι αποτελούνται από κείμενο (text part). Επιλέχθηκε το σύνολο του υλικού να είναι καταφορτώσιμο (downloadable) ώστε να μπορεί να μεταφορτώνεται από τους ενδιαφερόμενους. Διαλέξαμε κανένας πόρος να μην είναι επεξεργάσιμος (processable) για αποφυγή αλλοίωσης τους. Οι πόροι είναι ανοιχτοί στο κοινό, οπότε έχουμε no private distribution. Κανένας από τους πόρους δεν περιέχει προσωπικά ή ευαίσθητα προσωπικά δεδομένα, οπότε δεν χρειάστηκε να ανωνυμοποιηθούν. Η μορφή των δεδομένων (data format) είναι σε μορφότυπο TEXT, PDF, MS- WORD ή και PPT και Excel, τα τελευταία δύο μόνο ως μέρος μικτού πόρου που περιέχει και αρχείο στο συγκεκριμένο μορφότυπο μαζί με το υλικό σε αποδεκτούς από το Clarin τύπους. Οι άδειες χρήσης των γλωσσικών πόρων είναι ανοιχτής πρόσβασης Creative Commons.

Συνολικά έχουμε διασυνδέσει 193 γλωσσικούς πόρους με το Clarin:el. Πιο αναλυτικά, διαπιστώνουμε ότι το μεγαλύτερο μέρος του υλικού που έχουμε μεταφορτώσει στο Clarin:el προέρχεται από το ΤΕΙ Πειραιά (80 γλωσσικοί πόροι), έπειτα από το ΤΕΙ Αθηνών (68 γλωσσικοί πόροι) και τέλος από το Πανεπιστήμιο Δυτικής Αττικής (45 γλωσσικοί πόροι).

Προέλευση πόρου ανά ακαδημαϊκό ίδρυμα	Αριθμός γλωσσικών πόρων
ΤΕΙ Πειραιά	80
ΤΕΙ Αθηνών	75
Πανεπιστήμιο Δυτικής Αττικής	38

Πίνακας 2. Προέλευση πόρου ανά ακαδημαϊκό ίδρυμα

Το είδος του υλικού που ενσωματώσαμε στην υποδομή του Clarin:el αποτελείται κατά μεγάλη πλειοψηφία από Ανοιχτά Ακαδημαϊκά Μαθήματα του ΤΕΙ Αθηνών και του ΤΕΙ Πειραιά. Διαθέτουμε κάποιο αριθμό γλωσσικών πόρων οι οποίοι μας αποστάλθηκαν απευθείας από τους δημιουργούς τους,

που είναι διδάσκοντες στο Πανεπιστήμιο Δυτικής Αττικής, έπειτα από επικοινωνία μαζί τους. Πρόκειται για άρθρα ή διάφορες δημοσιεύσεις τους. Ενδεικτικά έχουμε ενσωματώσει οκτώ διπλωματικές και πτυχιακές εργασίες φοιτητών του τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης*, επτά συγγράμματα από τον Κάλλιππο με δημιουργούς καθηγητές του Τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης* και έξι περιγραφές μαθημάτων, που στάλθηκαν από τη διδάσκουσα καθηγήτρια τους. Φυσικά το Clarin:el θα μπορούσε να εμπλουτιστεί μελλοντικά με επιπλέον διπλωματικές και πτυχιακές εργασίες καλής ποιότητας και συγγράμματα του Κάλλιππου και από άλλες σχολές και τμήματα του Πανεπιστημίου Δυτικής Αττικής.

Είδος γλωσσικού πόρου	Αριθμός γλωσσικών πόρων
Ανοιχτά Ακαδημαϊκά Μαθήματα	148
Υλικό που στάλθηκε απευθείας από τους διδάσκοντες του ΠαΔΑ	24
Εργασίες φοιτητών του ΠαΔΑ	8
Κάλλιππος	7
Περιγραφές μαθημάτων	6

Πίνακας 3. Είδος γλωσσικού πόρου

Στη συνέχεια θα εξετάσουμε τους παρεχόμενους γλωσσικούς πόρους ανά σχολή και τμήμα, τόσο του Πανεπιστημίου Δυτικής Αττικής, όσο και των δύο Τεχνολογικών Ιδρυμάτων, των ΤΕΙ Αθηνών και Πειραιά δηλαδή. Αναφορικά με τα δύο ΤΕΙ, Αθηνών και Πειραιά, παρατηρούμε ότι όλο το υλικό που διαθέσαμε στο Clarin:El προέρχεται από τα Ανοιχτά Ακαδημαϊκά Μαθήματα που έχουν δημιουργηθεί και δημοσιευθεί με άδειες ανοιχτής πρόσβασης Creative Commons. Πιο συγκεκριμένα, διασυνδέσαμε στη γλωσσική υποδομή 80 μαθήματα από το ΤΕΙ Πειραιά και 68 από το ΤΕΙ Αθηνών.

Προέλευση Ανοιχτού Ακαδημαϊκού Μαθήματος	Αριθμός γλωσσικών πόρων
ΤΕΙ Αθηνών	68
ΤΕΙ Πειραιά	80

Πίνακας 4. Προέλευση Ανοιχτών Ακαδημαϊκών Μαθημάτων

Ταξινομώντας τους παρεχόμενους πόρους ανά σχολή του ΤΕΙ Αθηνών, διαπιστώνουμε ότι ο μεγαλύτερος αριθμός προέρχεται από τη σχολή Τεχνολογικών Εφαρμογών (36 πόροι), ακολουθούμενης από τη σχολή Καλλιτεχνικών Σπουδών (11 πόροι), τη σχολή Επαγγελμάτων Υγείας και Πρόνοιας (9 πόροι), τη σχολή Τεχνολογίας Τροφίμων και Διατροφής (7 πόροι) και τέλος τη σχολή Διοίκησης και Οικονομίας (5 πόροι).

Υλικό ανά σχολή του ΤΕΙ Αθηνών	Αριθμός γλωσσικών πόρων
Σχολή Διοίκησης και Οικονομίας	5
Σχολή Επαγγελμάτων Υγείας και Πρόνοιας	9
Σχολή Καλλιτεχνικών Σπουδών	11
Σχολή Τεχνολογίας Τροφίμων και Διατροφής	7
Σχολή Τεχνολογικών Εφαρμογών	36

Πίνακας 5. Υλικό ανά σχολή του ΤΕΙ Αθηνών

Εξετάζοντας τους γλωσσικούς πόρους του ΤΕΙ Αθηνών ανά τμήμα που τους παράγει, οι περισσότεροι προέρχονται από το τμήμα *Ναυπηγών Μηχανικών* (19 πόροι), ακολουθούμενο από τα τμήματα *Οινολογίας και Τεχνολογίας Ποτών*, *Μηχανικών Πληροφορικής* και *Φωτογραφίας και Οπτικοακουστικών Τεχνών* (από 6 πόρους το καθένα).

Υλικό ανά τμήμα του ΤΕΙ Αθηνών	Αριθμός γλωσσικών πόρων
Ναυπηγών Μηχανικών	19
Οινολογίας και Τεχνολογίας Ποτών	6
Μηχανικών Πληροφορικής	6
Φωτογραφίας και Οπτικοακουστικών Τεχνών	6
Πολιτικών Μηχανικών και Μηχανικών Τοπογραφίας και Γεωπληροφορικής	5
Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης	4
Μηχανικών Ενεργειακής Τεχνολογίας	4
Γραφιστικής	3
Αισθητικής και Κοσμετολογίας	2
Οπτικής και Οπτομετρίας	2
Ραδιολογίας και Ακτινολογίας	2
Μηχανικών Βιοιατρικής Τεχνολογίας	2
Συντήρησης Αρχαιοτήτων και Έργων Τέχνης	1
Δημόσιας Υγείας και Κοινοτικής Υγείας	1
Οδοντικής Τεχνολογίας	1
Ιατρικών Εργαστηρίων	1

Εσωτερικής Αρχιτεκτονικής, Διακόσμησης και Σχεδιασμού Αντικειμένων	1
Τεχνολογίας Τροφίμων	1
Διοίκησης Επιχειρήσεων - Τουριστικών Επιχειρήσεων και Επιχειρήσεων Φιλοξενίας	1

Πίνακας 6. Υλικό ανά τμήμα του ΤΕΙ Αθηνών

Αντίστοιχα, θα αναλύσουμε το παρεχόμενο υλικό ανά σχολή και τμήμα του ΤΕΙ Πειραιά. Το περισσότερο υλικό προέρχεται από τη σχολή Τεχνολογικών Εφαρμογών (57 πόροι). Ενώ η σχολή Διοίκησης και Οικονομίας διαθέτει 23 πόρους.

Υλικό ανά σχολή του ΤΕΙ Πειραιά	Αριθμός γλωσσικών πόρων
Τεχνολογικών Εφαρμογών	57
Διοίκησης και Οικονομίας	23

Πίνακας 7. Υλικό ανά σχολή του ΤΕΙ Πειραιά

Το τμήμα *Διοίκησης Επιχειρήσεων* είναι ο δημιουργός των περισσότερων παρεχόμενων πόρων στο Clarin:el (16 πόροι), ακολουθούμενο από τα τμήματα *Μηχανολόγων Μηχανικών* (14 πόροι), *Μηχανικών Αυτοματισμού* (13 πόροι) ή *Ηλεκτρονικών Μηχανικών* (12 πόροι).

Υλικό ανά τμήμα του ΤΕΙ Πειραιά	Αριθμός γλωσσικών πόρων
Διοίκηση Επιχειρήσεων	16
Μηχανολόγων Μηχανικών	14
Μηχανικών Αυτοματισμού	13
Ηλεκτρονικών Μηχανικών	12
Μηχανικών Ηλεκτρονικών Υπολογιστών Συστημάτων	8
Λογιστικής και χρηματοοικονομικής	7
Ηλεκτρολόγων Μηχανικών	6
Πολιτικών Μηχανικών	4

Πίνακας 8. Υλικό ανά τμήμα του ΤΕΙ Πειραιά

Το Πανεπιστήμιο Δυτικής Αττικής αποτελεί το συνεχιστή των δύο τεχνολογικών ιδρυμάτων, Αθηνών και Πειραιά. Οι προερχόμενοι από το ακαδημαϊκό ίδρυμα γλωσσικοί πόροι, οι οποίοι έχουν διασυνδεθεί με τη γλωσσική εφαρμογή αποτελούνται κατά πλειοψηφία από υλικό που δόθηκε απευθείας από τους διδάσκοντες του ΠαΔΑ, έπειτα από επικοινωνία μαζί τους (άρθρα, προϊόν έρευνας τους, δημοσιεύσεις κτλ.). Στη συνέχεια διαθέτουμε κάποια επιλεγμένα συγγράμματα από τον Κάλλιππο, εργασίες φοιτητών, καθώς και περιγραφές μαθημάτων.

Είδος υλικού από το ΠαΔΑ	Αριθμός γλωσσικών πόρων
Υλικό που δόθηκε απευθείας από τους διδάσκοντες του ΠαΔΑ	24
Συγγράμματα από τον Κάλλιππο	7
Διπλωματικές εργασίες	7
Περιγραφές μαθημάτων	6
Πτυχιακή εργασία	1

Πίνακας 9. Είδος γλωσσικού πόρου του ΠαΔΑ

Ταξινομώντας το υλικό που προέρχεται από το ΠαΔΑ ανά σχολή, διαπιστώνουμε ότι ο μεγαλύτερος αριθμός γλωσσικών πόρων προέρχεται από τη σχολή Διοικητικών Οικονομικών και Κοινωνικών Επιστημών. Έπειτα, το υλικό προέρχεται από τη σχολή Μηχανικών, ενώ διαθέτουμε και δύο γλωσσικούς πόρους από διδάσκοντες των σχολών Εφαρμοσμένων Τεχνών και Πολιτισμού και Επιστημών Υγείας και Πρόνοιας.

Υλικό ανά σχολή του ΠαΔΑ	Αριθμός γλωσσικών πόρων
Διοικητικών Οικονομικών και Κοινωνικών Επιστημών	28
Μηχανικών	8
Εφαρμοσμένων Τεχνών και Πολιτισμού	1
Επιστημών Υγείας και Πρόνοιας	1

Πίνακας 10. Υλικό ανά σχολή του ΠαΔΑ

Αναλύοντας το γλωσσικό υλικό που διασυνδέσαμε με το Clarin:el, το μεγαλύτερο μέρος, με 29 γλωσσικούς πόρους, προέρχεται από το τμήμα *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης*. Έπειτα, υπάρχουν 8 πόροι από το τμήμα *Μηχανικών Βιομηχανικής Σχεδίασης και Παραγωγής*, 6 από το τμήμα *Αγωγής και Φροντίδας στην Πρώιμη Παιδική Ηλικία* και από ένας πόρος από τα τμήματα *Γραφιστικής και Οπτικής Τεχνολογίας* και *Βιοιατρικών Επιστημών*.

Υλικό ανά τμήμα του ΠαΔΑ	Αριθμός γλωσσικών πόρων
Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης	29
Μηχανικών Βιομηχανικής Σχεδίασης και Παραγωγής	8
Αγωγής και Φροντίδας στην Πρώιμη Παιδική Ηλικία	6
Γραφιστικής και Οπτικής Τεχνολογίας	1
Βιοιατρικών Επιστημών	1

Πίνακας 11. Υλικό ανά τμήμα του ΠαΔΑ

Αναφορικά με τον τύπο του διαθέσιμου στο Clarin:el γλωσσικού πόρου, η μεγάλη πλειοψηφία είναι σώματα κειμένου (188 γλωσσικοί πόροι), ενώ υπάρχουν και 5 λεξικό – εννοιολογικοί πόροι. Υλικό από εργαλεία ή υπηρεσίες γλωσσικής επεξεργασίας καθώς και από γλωσσικές περιγραφές και υπολογιστικά μοντέλα δεν έχει συλλεχθεί καθόλου.

Τύπος γλωσσικού πόρου	Αριθμός γλωσσικών πόρων
Σώμα κειμένου	188
Λεξικό – εννοιολογικοί πόροι	5

Πίνακας 12. Τύπος γλωσσικού πόρου

Όλο το υλικό των 193 γλωσσικών πόρων που συλλέξαμε και διασυνδέσαμε με τη υποδομή του Clarin:el είναι επίσης σε μορφή κειμένου (text).

Μορφή πόρου	Αριθμός γλωσσικών πόρων
Text	193

Πίνακας 13. Μορφή των γλωσσικών πόρων

Αναφορικά με τη γλώσσα στην οποία είναι γραμμένοι οι γλωσσικοί πόροι που σωρευτήκαν στο Clarin:el, η ελληνική γλώσσα συναντάται 190 φορές, η αγγλική 6 φορές ενώ έχουμε και ένα γλωσσάρι σε διάλεκτο της Μεσσηνίας, με τη μετάφραση τους στα κοινά νέα ελληνικά.

Γλώσσα πόρων	Αριθμός γλωσσικών πόρων
Ελληνικά	190
Αγγλικά	6
Μεσσηνιακή διάλεκτος	1

Πίνακας 14. Γλώσσα των πόρων

Οι 189 γλωσσικοί πόροι είναι μονόγλωσσοι, στη συντριπτική τους πλειοψηφία, στην ελληνική γλώσσα με κάποιους λίγους, μεμονωμένους πόρους στα αγγλικά, οι οποίοι αποτελούν δημοσιεύσεις καθηγητών του ΠαΔΑ στη γλώσσα αυτή. Οι δίγλωσσοι πόροι είναι 4. Οι 3 στα ελληνικά και τα αγγλικά και ο ένας στα κοινά νέα ελληνικά και το γλωσσικό ιδίωμα της Μεσσηνίας.

Γλωσσικός τύπος	Αριθμός γλωσσικών πόρων
Μονόγλωσσοι	189
Δίγλωσσοι	4

Πίνακας 15. Γλωσσικός τύπος

Συμπεράσματα

Οι βασικοί έμπρακτοι στόχοι της έρευνας μας ήταν η διασύνδεση του Πανεπιστημίου Δυτικής Αττικής με την ψηφιακή υποδομή του Clarin:el. Το ΠαΔΑ έγινε μέλος του ελληνικού παραρτήματος μιας διεθνούς υποδομής και αποτελεί πλέον κομμάτι μιας διεθνούς κοινότητας που διαμοιράζεται γλωσσικούς πόρους και εργαλεία γλωσσικής τεχνολογίας, με στόχο την προώθηση της έρευνας στην Ελλάδα και το εξωτερικό.

Η ερευνητική προσέγγιση που χρησιμοποιήθηκε ήταν η έρευνα πεδίου με στόχο τη συλλογή του ζητούμενου υλικού. Πραγματοποιήθηκε δράση επικοινωνίας με τους δημιουργούς των γλωσσικών πόρων οι οποίοι είναι διδάσκοντες στο ΠαΔΑ, γραπτά, με μήνυμα ηλεκτρονικού ταχυδρομείου, με ενημερωτική συνάντηση μέσω της πλατφόρμας Teams ή και δια ζώσης. Η ανταπόκριση ήταν σχετικά μικρή αλλά παρόλα αυτά, συγκεντρώθηκε κάποιος ικανοποιητικός αριθμός γλωσσικού υλικού. Στη συνέχεια συλλέξαμε άλλους πόρους που είχαν ανοιχτές άδειες Creative Commons. Πρόκειται για συγγράμματα από τις ακαδημαϊκές ψηφιακές εκδόσεις του Κάλλιππου. Επίσης συλλέχτηκαν κείμενα που ήταν σε υποστηριζόμενους μορφότυπους από το Clarin:el, από τα *Ανοιχτά Ακαδημαϊκά Μαθήματα* του ΤΕΙ Αθηνών και του ΤΕΙ Πειραιά, τα οποία αποτελούν την προηγούμενη νομική μορφή που συγχωνευόμενα δημιούργησαν το Πανεπιστήμιο Δυτικής Αττικής. Τέλος, ενδεικτικά συμπεριλήφθηκαν στην έρευνα διπλωματικές εργασίες μικρού αριθμού φοιτητών του τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης*, έπειτα από ερώτηση αν θα επιθυμούσαν το υλικό τους να διασυνδεθεί με το Clarin:el. Στη συνέχεια, το υλικό, με την κατάλληλη επεξεργασία και τεκμηρίωση, διασυνδέθηκε με το Clarin:el. Οι γλωσσικοί πόροι σε σώματα κειμένου σε αμιγώς ελληνική γλώσσα υπερτερούν.

Οι βάσεις της διασύνδεσης του ΠαΔΑ με την υποδομή του Clarin:El έχουν πλέον τεθεί. Ένας συνεχής εμπλουτισμός του ψηφιακού αποθετηρίου με παραγωγή γλωσσικών πόρων στα πλαίσια της μελλοντικής ακαδημαϊκής λειτουργίας του ΠαΔΑ θα ήταν επιθυμητή.

Βιβλιογραφικές Αναφορές

Απολλωνίς. https://apollonis-infrastructure.gr/. [Ανάκτηση 4/10/2023].

Clarin. https://www.clarin.eu/. [Ανάκτηση 1/10/2023]

Clarin:el. https://www.clarin.gr/. [Ανάκτηση 1/10/2023]

Clarin:el. (2020). Ηλεκτρονικό εγχειρίδιο χρήσης Clarin:el. Ανάκτηση 20/10/2023, από https://clarin-platform-documentation.readthedocs.io/el/stable/.

Creative Commons. https://creativecommons.org /. [Ανάκτηση 4/10/2023].

Γαβριηλίδου, Μ., & Πιπερίδης, Σ., (2021). *Τεχνητή Νοημοσύνη, Γλωσσική Τεχνολογία και Γλωσσικές υποδομές.* Αθήνα: Περιοδικό «Οικονομικός Ταχυδρόμος»,21/10/2021. https://www.ot.gr/2021/10/19/academia/texniti-noimosyni-glossiki-texnologia-kai-glossikes-ypodomes/.[Ανάκτηση 29/9/2023].

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

Dariah.eu. https://www.dariah.eu/. [Ανάκτηση 4/10/2023].

Dariah-gr/ ΔΥΑΣ. https://dyas-net.gr/. [Ανάκτηση 6/10/2023].

Dipper, S. (2008). Theory-driven and corpus-driven computational linguistics, and the use of corpora. In A.Lüdeling & M. Kytö (eds) Corpus Linguistics: An International Handbook. Berlin: Walter deGruyter, 68-96.

Europeana.eu. https://www.europeana.eu/. [Ανάκτηση 4/10/2023].

Καπιδάκης, Σ. (2014). *Εισαγωγή στις ψηφιακές βιβλιοθήκες*. Αθήνα : Δίσιγμα Εκδόσεις.

Καπιδάκης, Σ., Λαζαρίνης, Φ., & Τοράκη, Κ. (2015). Θέματα βιβλιοθηκονομίας και επιστήμης των πληροφοριών [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. http://hdl.handle.net/11419/1674.

Κυριάκη-Μάνεση, Δ., & Κουλούρης, Α. (2015). *Διαχείριση ψηφιακού περιεχομένου* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. http://hdl.handle.net/11419/2496.

Ν. 4521/2018 (ΦΕΚ Α΄, 38/2-3-2018): «Ίδρυση Πανεπιστημίου Δυτικής Αττικής». Πανεπιστήμιο Δυτικής Αττικής. https://www.uniwa.gr. [Ανάκτηση 2/9/2022].

Πιπερίδης, Σ., Λαμπροπούλου, Π. & Γαβριηλίδου, Μ. (2015). Clarin:el: Δημιουργώ, επεξεργάζομαι, μοιράζομαι, 10ο Συνέδριο "Ελληνική Γλώσσα και Ορολογία". Αθήνα. [Conference paper].

Πιπερίδης, Σ., Λαμπροπούλου, Π. & Γαβριηλίδου, Μ. (2015). Clarin:el: μια υποδομή τεκμηρίωσης, διαμοιρασμού και επεξεργασίας γλωσσικών δεδομένων,*12ο Συνέδριο Ελληνικής Γλωσσολογίας*. Βερολίνο. [Conference paper].

Πουλή, Κ., Τσιούλη, Η. & Λαμπροπούλου, Π. (2017). Ορολογικοί πόροι ΟΡΟΣΗΜΟ: επιμέλεια, ταξινόμηση και αποτελέσματα,11ο Συνέδριο "Ελληνική Γλώσσα και Ορολογία". Αθήνα. [Conference paper].

Schreibman, S., Siemens R. & Unsworth J. (eds.). (2004). *A Companion to Digital Humanities*. Oxford: Blackwell, http://www.digitalhumanities.org/companion. [Ανάκτηση 9/9/2023].

Shirley, W. (2005). Leung, International conference on developing digital institutional repositories: experiences and challenges. *Library high tech news, no. 2,* pp. 14-15.

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