ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΣΧΟΛΗ ΑΝΘΡΩΠΙΣΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΚΑΙ ΠΟΛΙΤΙΣΜΙΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑ ΦΙΛΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΑΡΧΑΙΑ ΚΑΙ ΝΕΑ ΕΛΛΗΝΙΚΗ ΦΙΛΟΛΟΓΙΑ
ΚΑΤΕΥΘΥΝΣΗ: ΝΕΑ ΕΛΛΗΝΙΚΗ ΓΛΩΣΣΑ ΚΑΙ ΦΙΛΟΛΟΓΙΑ

«Εργαλεία για την μέτρηση της συντακτικής ευελιξίας των ρηματικών πολυλεκτικών στην γενική γλώσσα και στην γλώσσα της λογοτεχνίας: εφαρμογή στην περίπτωση Ταχτσή.»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Πέτρου Βασιλικού

Διπλωματούχου του Τμήματος Φιλολογίας του Πανεπιστημίου Πελοποννήσου, 2017

Επιβλέπουσα Καθηγήτρια: Στυλιανή Μαρκαντωνάτου**,** Ερευνήτρια Α' Ινστιτούτο Επεξεργασίας Λόγου

Συνεπιβλέποντες: Ανδρειωμένος Γεώργιος, Καθηγητής, Τμήμα Φιλολογίας, Πανεπιστήμιο Πελοποννήσου

Χατζηδάκη Ουρανία, Αναπληρώτρια Καθηγήτρια, Τμήμα Αεροπορικών Επιστημών Τομέας Ηγετικής/Διοικητικής, Ανθρωπιστικών Επιστημών και Φυσιολογίας Σχολή Ικάρων

Καλαμάτα, Αύγουστος 2019

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

Αφορμή για τη συγκεκριμένη έρευνα αποτέλεσε η παρακολούθηση του μαθήματος Digital Humanities κατά τη διάρκεια της φοίτησής μου για την απόκτηση του μεταπτυχιακού μου διπλώματος. Το ενδιαφέρον μού κέντρισε ιδιαίτερα η χρήση των ρηματικών πολυλεκτικών εκφράσεων στη λογοτεχνία όπως και στα γλώσσα γενικότερα. Σε συνδυασμό με την ενασχόλησή μου με την πληροφορική θεώρησα ότι ήταν μια έρευνα που θα μπορούσα να φέρω εις πέρας. Σκοπός της συγκεκριμένης έρευνας είναι η εκτίμηση της αποτελεσματικότητας των εργαλείων πληροφορικής που έχουμε στη διάθεσή μας σε σύγκριση με τη χειρωνακτική έρευνα για την επίτευξη του στόχου της έρευνας. Αξιοποιώντας λοιπόν έναν επιλεγμένο crawler δημιουργούνται συλλογές κειμένων από το διαδίκτυο με κριτήριο την παρουσία μιας συγκεκριμένης πολυλεκτικής έκφρασης. Συγκρίνουμε το μέγεθος αυτών των συλλογών με αντίστοιχες συλλογές που έχουν γίνει χειρωνακτικά. Στην συνέχεια, δημιουργούμε αντίστοιχες συλλογές κειμένων για τις 4 πιο υψίσυχνες πολυλεκτικές εκφράσεις στο *Τρίτο Στεφάνι* του Κ. Ταχτσή. Στη συνέχεια, επεξεργαζόμαστε δύο από αυτές τις συλλογές με το MWE Toolkit με στόχο να αποκτήσουμε μια πρώτη ιδέα για τα θέματα που θέτει η χρήση αυτής της σουίτας εργαλείων για την συγκέντρωση υλικού για την μελέτη της συντακτικής ευελιξίας των ΠΛΕ. . Οι δυσκολίες που καλούμαστε να αντιμετωπίσουμε σε αυτήν την έρευνα αφορούν κυρίως την επιλογή των κατάλληλων εργαλείων για την δημιουργία συλλογών κειμένων..

Θα ήθελα σε αυτό το σημείο να ευχαριστήσω θερμά την κα Στυλιανή Μαρκαντωνάτου για την συνεχή αρωγή της στην έρευνα και την καθοδήγησή της. Επίσης θα ήθελα να ευχαριστήσω τον Κύριο Βασίλη Παπαβασιλείου ο οποίος δαπάνησε πολύ χρόνο για να προσαρμόσει το βασικό πρόγραμμα που χρησιμοποίησα σε αυτή την έρευνα σύμφωνα με τις ανάγκες μας και για την καθοδήγηση του για την λειτουργία του προγράμματος.

ΚΕΦΑΛΑΙΟ 1

Η ΓΛΩΣΣΑ ΤΟΥ ΤΑΧΤΣΗ

Η έρευνά μας αφορά την χρήση εργαλείων πληροφορικής για την μελέτη της συντακτικής ευελιξίας των ρηματικών πολυλεκτικών εκφράσεων, Αυτού του τύπου η έρευνα απαιτεί μεγάλες συλλογές κειμένων όπου κανείς μπορεί να εντοπίσει πολλά φαινόμενα και να εκτιμήσει την συχνότητα εμφάνισής τους, η οποία μπορεί να εξαρτάται και από το είδος του λόγου ή, στην περίπτωση της λογοτεχνίας, και από την γλώσσα του συγγραφέα. Επιλέξαμε ως αναφορά *Το Τρίτο Στεφάνι* του Κ. Ταχτσή γιατί, όπως συζητάμε στην συνέχεια αυτού του κεφαλαίου, είναι γραμμένο σε γλώσσα που έχει πολλά κοινά με τον προφορικό λόγο και ένα σημαντικό χαρακτηριστικό του προφορικού λόγου είναι η ιδιωματικότητα. Οι (ρηματικές) πολυλεκτικές εκφράσεις είναι χαρακτηριστικοί δείκτες της ιδιωματικότητας.

Η γλώσσα αυτού του βιβλίου έχει συζητηθεί περισσότερο από οποιαδήποτε άλλη πλευρά του και σύμφωνα με τον Roderick Beaton αξίζει να θεωρήσουμε Το Τρίτο Στεφάνι σαν ένα κωμικό αριστούργημα, ανάλογο με τα επιτεύγματα του δέκατου ένατου αιώνα στην κωμωδία. (Beaton, 1996, σσ. 424-425)

Ο Μάριο Βίττι γράφει για Το Τρίτο Στεφάνι: «Μια απομάκρυνση από τα άμεσα θέματα του πολέμου έχουμε στο μυθιστόρημα του Κώστα Ταχτσή (1927-1988) Το τρίτο στεφάνι (1962), όπου μια μικροαστή και η πεθερά της αναθυμούνται την προπολεμική ζωή με τις καθημερινές της εκπλήξεις.» (Vitti, 1926, σσ. 528-529)

«Κατά τη μεταπολεμική περίοδο, στομφώδεις χαρακτήρες τείνουν να χρησιμοποιούν στοιχεία της καθαρεύουσας στην ομιλία τους, ενώ η γλώσσα των εφημερίδων, καθώς και η επίσημη γλώσσα της γραφειοκρατίας, παρατίθεται ή ακόμη και παρωδείται.» (Beaton, 1996, σσ. 424-425) Στο Τρίτο Στεφάνι του Κώστα Ταχτσή παρατηρούμε τα αποτελέσματα των αλλαγών αυτών που επήλθαν σταδιακά. Ο Ταχτσής προσπάθησε να δημιουργήσει μια υποθετικά προφορική αφήγηση (δυο μικροαστές γυναίκες λένε τις ιστορίες τους). Όσον αφορά όμως τη γλώσσα, τα στοιχεία

του προφορικού λόγου των μυθιστορηματικών γυναικών αποτελούν το μέρος του κόσμου που κατανοούν οι αφηγήτριες και της ζωντανής προφορικής ανταπόκρισής τους προς αυτόν. Το μυθιστόρημα αυτό απέδειξε ότι η πραγματική ζωή δεν μπορούσε να δεσμεύεται από τις τεχνητές κατηγορίες της διγλωσσίας και ότι η αφηγηματική πεζογραφία δε χρειάζεται να ανήκει ούτε στη μία ούτε στην άλλη μορφή της γλώσσας. (Beaton, 1996, pp. 424-425) Πρώτη φορά στον Ταχτσή, εμφανίζεται μια γλώσσα απαλλαγμένη από τους περιορισμούς της δημοτικής με τη χρήση της πρωτοπρόσωπης αφήγησης σε συνδυασμό με την ζωντανή προφορικότητα. Αυτό ήταν κάτι το οποίο δεν είχε επιτύχει κανένας από τους σύγχρονούς του. Μέσω αυτής της γλώσσας μεταφέρεται στον αναγνώστη το βίωμα των συμβάντων από τις αφηγήτριες του έργου μέσα από την δική τους οπτική. (Vitti, 1926, σσ. 528-529)

«Το μυθιστόρημα έχει τη μορφή παρλάτας. Πρόκειται για μια διπλή πρωτοπρόσωπη αφήγηση, ενώ το σκηνικό είναι αποκλειστικά αστικό.» Μέσα από την γλωσσική ιδιοτυπία της παρλάτας[1] της κεντρικής ηρωίδας του έργου, με την ανάμειξη της καθαρεύουσας και της δημοτικής φαίνεται και η βασική αρετή του βιβλίου όπως επισήμαναν άλλωστε αρκετοί Έλληνες κριτικοί. (ΒΙΚΙΠΑΙΔΕΙΑ, 2018)

Σε κάθε γλώσσα σύμφωνα με τον Καζάζη, (Kazazis, 1979, σ. 17) υπάρχουν οι λεγόμενοι λογιοτατισμοί (στα αγγλικά : learnedisms), κάτι σαν αυτό που οι Γάλλοι ονομάζουν «mots savants» . Στην ελληνική γλώσσα τέτοιου τύπου λέξεις οφείλονται στην επιρροή της καθαρεύουσας. Σε γενικές γραμμές αναφέρει ότι είναι κατανοητή αυτή η αποφυγή των λογιοτατισμών αν αναλογιστούμε ότι η ελληνική λογοτεχνική γλώσσα καθιερώθηκε μετά από αγώνα ως ένα μέσο εκμάθησης. Μάλιστα ο Καζάζης αναφέρει ότι ένας έμπειρος αναγνώστης με γλωσσολογικές γνώσεις μπορεί να καταλάβει το τεχνητό κομμάτι μιας γλώσσας απαλλαγμένης από λογιοτατισμούς όπως αυτή παρουσιάζεται π.χ στον Νίκο Καζαντζάκη, τον Βενέζη και πιο πρόσφατα τον Ευάγγελο Αβέρωφ Τοσίτσα.

Ακόμα ο Καζάζης τονίζει ότι υπήρχε σίγουρα αντίδραση στην παγίωση της δημοτικής όπως φαίνεται στα κείμενα του Ανδρέα Εμπειρίκου και του Κώστα Ταχτσή. Στο σημείο αυτό αξίζει να τονίσουμε την φράση του Γεώργιου Σαββίδη το 1973 αναφερόμενος στον Ταχτσή όπως την παραθέτει ο Καζάζης

[1] παρλάτα 1) μονόλογος ενός ηθοποιού 2) (μεταφορικά) μονότονος μονόλογος κάποιου
Π.χ : (πέρασε ένας πωλητής, μας πούλησε την παρλάτα του αλλά έφυγε άπραγος!) (Βικιλεξικό, 2013)

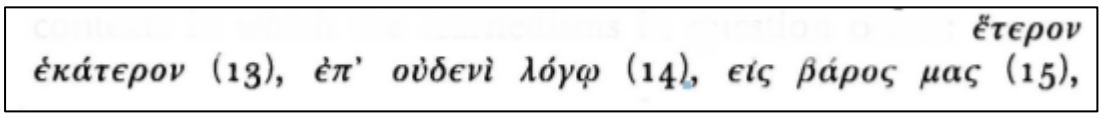
(Kazazis, 1979, σ. 18) : « Ελευθέρωσε την ελληνική γλώσσα από την τυραννία της δημοτικής». Παρατηρούμε π.χ εσωτερικές επαυξήσεις «απεφάσισα» αλλά και «αποφάσισα». Ακόμα εμφανίζονται λέξεις όπως «λεπτά-λεφτά», «νυχθημερόν αντί μέρα-νύχτα» καθώς και άλλα σύνολα λέξεων όπως τα εξής : «έκτρωμα της φύσεως, ιδίοις όμμασι» κ.α

Όσον αφορά τον Ταχτσή ως προς το σύνολο του έργου του φαίνεται να παραμένει σταθερός στη χρήση συγκεκριμένων τύπων λέξεων που μπορούν να γραφτούν με δυο τρόπους. Για παράδειγμα αναφέρει ότι ο χρησιμοποιεί τον τύπο «λεπτά» για να δηλώσει το χρόνο αντί για το «λεφτά» όπως και το «αδερφός, αδερφή» αντί για το «αδελφός, αδελφή». Σε αυτό το σημείο ο Καζάζης επισημαίνει βέβαια ότι γίνεται χρήση του «αδελφή» για να δηλωθεί η έννοια της νοσοκόμας όπως είναι άλλωστε και το θεμιτό. Αν και οι υπέρμαχοι της δημοτικής γλώσσας όπως αναφέρει θεώρησαν αυτές τις ταλαντώσεις στη γλώσσα απροσεξίες, κατά τη γνώμη του Καζάζη τέτοιες ταλαντώσεις στον Ταχτσή δεν αποδίδονται σε απροσεξία. Αντίθετα τις αποδίδει στην ίδια τη μοντέρνα ελληνική ομιλία ακόμα και όταν μιλάμε για ένα άτομο εννοώντας ότι κάποιος μπορεί να χρησιμοποιήσει και το « εν των μεταξύ» και το «στο μεταξύ».

Ο Ταχτσής παρεμπιπτόντως, κάνει μεγάλη χρήση της καθαρεύουσας όταν αναφέρει τι λέγεται στο δικαστήριο (165 κ.ε.), όταν έχει μέλη του νομικού (86), διευθυντές σχολείων που απευθύνονται στους γονείς των μαθητών τους (98) και φυσικά όταν κάποιος είναι σαρκαστικός. (Kazazis, 1979, pp. 17-21) Ένα παράδειγμα αυτής της χρήσης της καθαρεύουσας είναι όταν η Κυρά Εκάβη, ένας από τους πρωταγωνιστές στο Τρίτο Στεφάνι , θυμωμένη με τη σύζυγο του ιδιοκτήτη του σπιτιού της ανυψώνει τη φόρεμα της και πέρδεται προς την κατεύθυνση της γυναίκας του ιδιοκτήτη λέγοντας : «Ιδού η απάντησίς μου, Κυρία Μαργαρίτη μου, και εις την μητρικήν σας γλώσσα!» (Kazazis, 1979, p. 21) Αν και τα συγκεκριμένα παραδείγματα όμως αποτελούν ειδικές περιπτώσεις, ωστόσο ο Ταχτσής χρησιμοποιεί λογιοτατισμούς και σε χωρία που δεν φαίνεται να έχεις ως σκοπό να τους δώσει ένα ειδικό εφέ. Για παράδειγμα ο Καζάζης παραθέτει τα εξής παραδείγματα : «Ο Ταχτσής όπως αναφέρει χαρακτηριστικά χρησιμοποιεί την ονομαστική και αιτιατική σε (–αί και –άς) για τα ισοσύλλαβα αρσενικά ουσιαστικά που λήγουν σε (–ής) όπως : να μην πετάω άδικα λεφτά για προγυμναστάς, λησταί , παραθερισταί , οι δικασταί , μήπως λίγους εραστάς είχες και εσύ; , οι κομμουνισταί. Μόνο σε ένα χωρίο φαίνεται να χρησιμοποιεί την κατάληξη (συναγωνιστ –ές) μέσα σε εισαγωγικά : *έκανε νόημα στους "συναγωνιστές"* για να

δείξει ότι αυτός ήταν ο τρόπος με τον οποίο επικοινωνούσαν οι επαναστάτες κομμουνιστές την περίοδο της ανόδου των κομμουνιστών στην Αθήνα το Δεκέμβρη του 1944.» (Kazazis, 1979, σ. 22)

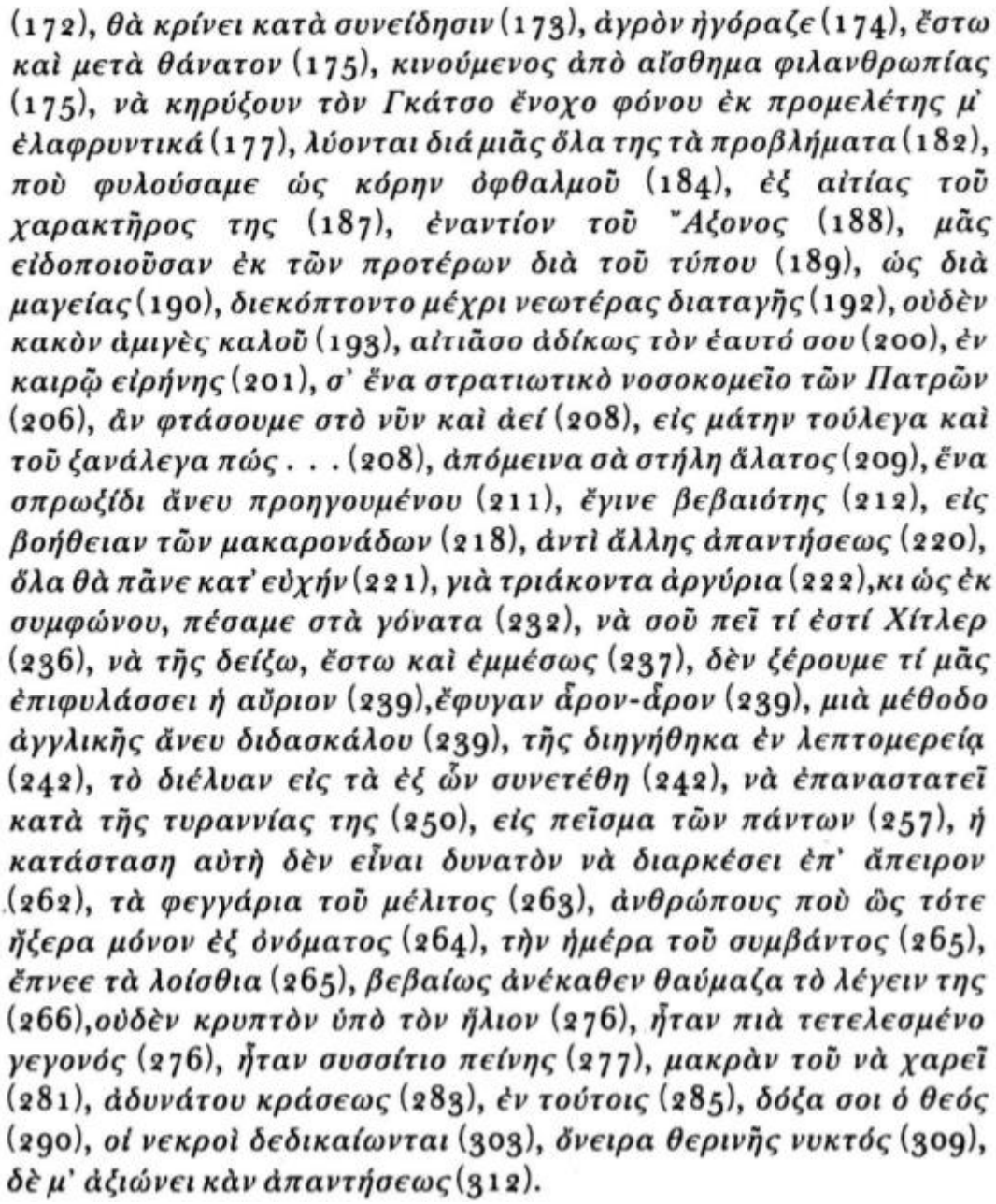
Κάτι που φαίνεται να μένει αναπάντητο σύμφωνα με τον Καζάζη είναι το κατά πόσο ο Ταχτσής χρησιμοποίησε τυχαία λογιοτατισμούς στο έργο του ή ο ίδιος φανταζόταν στο μυαλό του τους χαρακτήρες των έργων του με αυτόν τον τρόπο. «Ο ίδιος πιστεύει ότι αυτό μπορεί και να ήταν ένα προσωπικό στοίχημα του συγγραφέα με τον εαυτό του για τη χρήση των λογιοτατισμών στο έργο του.» (Kazazis, 1979, σ. 22) Παρακάτω παρατίθενται τα υψίσυχνα ευρήματα λογιοτατισμών που βρίσκει ο Καζάζης στο έργο του Ταχτσή και εντός παρενθέσεων φαίνεται η πρώτη εμφάνισή τους μέσα στο κείμενο.



Εικόνα 1 Λογιοτατισμοί 1

προσωποποίηση τοῦ διαβόλου ἐπὶ τῆς γῆς (16), ἔχοντας πικράν πείρα (16), καί οὕτω καθ' ἑξῆς (17), πρῶτον . . . , δεύτερον . . . (21), ἀφ' ἑνός . . . , ἀφ' ἑτέρου . . . (23), ἀπεποιήθη τὴν προσφορά (27), τό ἀπολωλὸς πρόβατο (30), μέγα μυστήριο! (31), πρὸς στιγμήν (35), μέχρις ἐσχάτων (36), ἐν θριάμβῳ (44), οὐδ' ἐπὶ στιγμήν (45), ἐγὼ δὲν κατέρχομαι βεβαίως στὸ ἐπίπεδό της (52), εἰρήσθω ἐν παρόδῳ (55), παντὶ τρόπῳ (55), θοῦ Κύριε φυλακὴν τῷ στόματί μου (56), ψυχῇ τε καὶ σώματι (57), ὡς ἐπὶ τὸ πλεῖστον (58), μέχρι ἀηδίας (62), κακὴν κακῶς (63), βρὲ ζῶον (64), δωρεὰ ἐν τῇ ζωῇ (65), ἐν ἀποστρατείᾳ (66), ἐπ' ἀνδραγαθίᾳ (66), ἀνελάμβανε τὴν ὑλοτόμηση μοναστηριακῶν δασῶν κατ' ἀποκοπήν (71), τοῖς μετρητοῖς (72), τόφεραν βαρέως (72), ὅπου γῆς καὶ πατρίς (73), τὸν κώδωνα τοῦ κινδύνου (73), ἕνα καὶ τὸ αὐτό (76), ὅταν ὁ ἀλέκτωρ ἐφώνησε τρίς (73), πίστευε καὶ μὴ ἐρεύνα (78), ἐκώφευσα (78), ὑπὲρ τῆς ἀμοιβαίας κατανοήσεως (78), συμβούλια ἐπὶ συμβουλίων (79), ἐν ἀνάγκῃ (80), μέρος προδιαγεγραμμένου σχεδίου (81), ἀγωγὴ διαζυγίου ἐπὶ ἐγκαταλείψει τῆς συζυγικῆς στέγης καὶ ἀγνώστῳ διαμονῇ (85), κατὰ προτροπήν του (86), ἐναντίον μιᾶς τόσον καταφώρου ἀδικίας (87), ἦταν τῶν ἀδυνάτων ἀδύνατον (88), ἦταν ὑπεράνω τῶν δυνάμεών μου (88), πρὸς χάριν τῶν παιδιῶν μου (88), ἀπὸ προσώπου τῆς γῆς (90), ὀλίγου δεῖ καὶ θὰ τὸν τουφέκιζαν (90), κατόπιν ἐντολῆς μου (91), πρὸ πολλοῦ (91), δυὸ μέρες πρὸ τῆς δίκης (91), μόλις καὶ μετὰ βίας (94), ἐπ' αὐτοφόρῳ (95, *sic* for ἐπ' αὐτοφώρῳ), τοῦ ζητοῦσε συγνώμην (97; the fully learned form is, of course, συγγνώμην), τὸ μὲν πνεῦμα πρόθυμο, ἀλλ' ἡ σάρξ ἀσθενής (99), ἔγινα πῦρ καὶ μανία (100), διάταγμα περὶ ἐθελουσίας ἐξόδου τῶν παλαιῶν ὑπαλλήλων (102), δουλειὰ ἀθλητικοῦ συντάκτου (105), οἱ σχέσεις τους ἦταν ὑπέρ ποτε καλές (105), περὶ τίνος ἐπρόκειτο (105), ἐν ὀλίγοις (108), ἐπεδείξατο μετάνοιαν καὶ ἀρίστην διαγωγήν (109), μιὰ ὡραία πρωία (114), ἐπὶ τόπου (118), πρὸς μεγάλην μου ἔκπληξιν (119), σάν τὸ πῦρ τῆς κολάσεως (122), μέσῳ ἐμοῦ (129), νὰ σκεφτοῦμε μαζὶ περὶ τοῦ πρακτέου (129), ἐξ ἐνστίκτου (130), ἐκ πείρας (136), αὐτὸ πιὰ εἶναι ἄνω ποταμῶν (136), πρὸς τὸ παρόν (140), στὰ χαρτιὰ ἐξακολουθοῦσε νάναι ἡ νόμιμος χήρα του (150), ὑπὸ τὰ ὄμματα τοῦ καταστηματάρχη (153), ὑπὸ τὸν ὅρον . . . ὅτι . . . (155), ἐν τῇ ἀφελείᾳ μου (155), ὑπὸ τύπον δανείου (156), θὰ σὲ στείλω συνοδείᾳ (160), εἶχε ἐκ θεοῦ τὸ χάρισμα νά . . . (163), διὰ τοῦ ὑπνωτισμοῦ (163), ἔπνεαν μένεα ἐναντίον του (165), μήνυση ἐπὶ μοιχείᾳ (167), κεκλεισμένων τῶν θυρῶν (170), εἰς ἔνδειξιν ὑπερτάτης ἀδυναμίας

Εικόνα 2 Λογιοτατισμοί 2



(172), θὰ κρίνει κατὰ συνείδησιν (173), ἀγρὸν ἠγόραζε (174), ἔστω καὶ μετὰ θάνατον (175), κινούμενος ἀπὸ αἴσθημα φιλανθρωπίας (175), νὰ κηρύξουν τὸν Γκάτσο ἔνοχο φόνου ἐκ προμελέτης μ' ἐλαφρυντικά (177), λύονται διά μιᾶς ὅλα της τὰ προβλήματα (182), ποὺ φυλούσαμε ὡς κόρην ὀφθαλμοῦ (184), ἐξ αἰτίας τοῦ χαρακτῆρος της (187), ἐναντίον τοῦ Ἄξονος (188), μᾶς εἰδοποιοῦσαν ἐκ τῶν προτέρων διὰ τοῦ τύπου (189), ὡς διὰ μαγείας (190), διεκόπτοντο μέχρι νεωτέρας διαταγῆς (192), οὐδὲν κακὸν ἀμιγὲς καλοῦ (193), αἰτιᾶσο ἀδίκως τὸν ἑαυτό σου (200), ἐν καιρῷ εἰρήνης (201), σ' ἕνα στρατιωτικὸ νοσοκομεῖο τῶν Πατρῶν (206), ἂν φτάσουμε στὸ νῦν καὶ ἀεί (208), εἰς μάτην τούλεγα καὶ τοῦ ξανάλεγα πώς . . . (208), ἀπόμεινα σὰ στήλη ἅλατος (209), ἕνα σπρωξίδι ἄνευ προηγουμένου (211), ἔγινε βεβαιότης (212), εἰς βοήθειαν τῶν μακαρονάδων (218), ἀντὶ ἄλλης ἀπαντήσεως (220), ὅλα θὰ πᾶνε κατ' εὐχήν (221), γιὰ τριάκοντα ἀργύρια (222), κι ὡς ἐκ συμφώνου, πέσαμε στὰ γόνατα (232), νὰ σοῦ πεῖ τί ἐστί Χίτλερ (236), νὰ τῆς δείξω, ἔστω καὶ ἐμμέσως (237), δὲν ξέρουμε τί μᾶς ἐπιφυλάσσει ἡ αὔριον (239), ἔφυγαν ἅρον-ἅρον (239), μιὰ μέθοδο ἀγγλικῆς ἄνευ διδασκάλου (239), τῆς διηγήθηκα ἐν λεπτομερείᾳ (242), τὸ διέλυαν εἰς τὰ ἐξ ὧν συνετέθη (242), νὰ ἐπαναστατεῖ κατὰ τῆς τυραννίας της (250), εἰς πεῖσμα τῶν πάντων (257), ἡ κατάσταση αὐτὴ δὲν εἶναι δυνατὸν νὰ διαρκέσει ἐπ' ἄπειρον (262), τὰ φεγγάρια τοῦ μέλιτος (263), ἀνθρώπους ποὺ ὡς τότε ἤξερα μόνον ἐξ ὀνόματος (264), τὴν ἡμέρα τοῦ συμβάντος (265), ἔπνεε τὰ λοίσθια (265), βεβαίως ἀνέκαθεν θαύμαζα τὸ λέγειν της (266), οὐδὲν κρυπτὸν ὑπὸ τὸν ἥλιον (276), ἦταν πιὰ τετελεσμένο γεγονός (276), ἦταν συσσίτιο πείνης (277), μακρὰν τοῦ νὰ χαρεῖ (281), ἀδυνάτου κράσεως (283), ἐν τούτοις (285), δόξα σοι ὁ θεός (290), οἱ νεκροὶ δεδικαίωνται (303), ὄνειρα θερινῆς νυκτός (309), δὲ μ' ἀξιώνει κὰν ἀπαντήσεως (312).

Εικόνα 3 Λογιοτατισμοί 3

Στη συνέχεια αξίζει να τονίσουμε ότι και οι δύο γυναίκες ανήκουν στη μεσαία τάξη και δεν φαίνεται να έχουν λάβει σημαντική παιδεία ώστε να συνδέσουμε τη χρήση της καθαρεύουσας με μια τέτοια γνώση. Παρατηρούμε ωστόσο ότι χρησιμοποιούν περίτεχνα και τους λογιοτατισμούς όταν εξάπτονται αλλά και την καθομιλουμένη σε απλές καταστάσεις. (Kazazis, 1979, σσ. 24-25)

Ωστόσο γύρω από την περίοδο συγγραφής του έργου του Ταχτσή παρατηρείται ένα φαινόμενο κατά το οποίο οι ημιμαθείς ή ακόμα και οι αμαθείς προσπαθούν να δείξουν

με το λόγο τους ότι έχουν μια στοιχειώδη παιδεία και θέλουν να δείχνουν άνθρωποι ανώτερης τάξης. Αυτό προσπαθούν να το επιτύχουν χρησιμοποιώντας σχεδόν αυθαίρετα τύπους λέξεων που έχουν συγκρατήσει χωρίς να γνωρίζουν τους κανόνες της γραμματικής για αυτές τις λέξεις. Παρ' όλα αυτά ο Ταχτσής δεν φαίνεται να χρησιμοποιεί τη γλώσσα στο έργο του με τέτοιο τρόπο ώστε να εντάξει κοινωνικά τους χαρακτήρες του έργου του αλλά ούτε και για να ευχαριστήσει τους αναγνώστες του. Κατά τον Καζάζη ακόμα και αν υπερβάλλει στατιστικά ελαφρώς ο Ταχτσής στη χρήση των λογιοτατισμών, με αποτέλεσμα να δυσανασχετήσουν οι ακραίοι δημοτικιστές, ωστόσο δεν μπορούμε να αρνηθούμε το γεγονός ότι λίγοι θα αμφισβητούσαν το ρεαλισμό του μυθιστορήματός του. (Kazazis, 1979, σσ. 25-27)

ΚΕΦΑΛΑΙΟ 2

ΠΟΛΥΛΕΚΤΙΚΑ ΣΥΝΟΛΑ-ΠΑΓΙΩΜΕΝΕΣ ΕΚΦΡΑΣΕΙΣ

2.1 Πολυλεκτικές εκφράσεις

Οι γλώσσες αποτελούνται από λέξεις που συνδυάζονται μορφοσυντακτικά μέσα σε προτάσεις και φράσεις για να αποδώσουν νόημα. «Πολυλεκτική έκφραση είναι ένα λέξημα αποτελούμενο από μια ακολουθία δύο ή περισσότερων λεξημάτων που έχουν ιδιότητες που δεν είναι προβλέψιμες από τις ιδιότητες του καθενός λεξήματος ή τον κανονικό τρόπο συνδυασμού τους. Οι Π.Λ.Ε έχουν σημαντικό ρόλο στις εφαρμογές της επεξεργασίας της φυσικής γλώσσας και στην υπολογιστική γλωσσολογία. Αποτελούν σοβαρό πρόβλημα στην επεξεργασία της γλώσσας λόγω της ιδιοσυγκρασιακής φύσης, της σημασιολογικής ποικιλομορφίας αλλά και λόγω των λεξικών, συντακτικών, πραγματολογικών και στατιστικών ιδιοτήτων τους.» (T.E. Jisha, 2015) Παραδείγματα ΠΛΕ είναι η «τινάζω τα πέταλα» (πεθαίνω), «παιδική χαρά» (τόπος παιχνιδιού των παιδιών), «βάζω μπρος» (ξεκινώ). Παρατηρούμε ότι η σημασία δεν προκύπτει από τις επιμέρους λέξεις και τις συντακτικές τους σχέσεις.

Τα τελευταία χρόνια οι πολυλεκτικές εκφράσεις έχουν επιτύχει να εγείρουν σε μεγάλο βαθμό την προσοχή της υπολογιστικής γλωσσολογίας και εφαρμογές επεξεργασίας φυσικής γλώσσας, αναγνώρισης ονομάτων (NER)[2], παραγωγής και κατανόησης φυσικής γλώσσας, αναγνώρισης οπτικών χαρακτήρων κ.α. Οι πολυλεκτικές εκφράσεις είναι εκείνες των οποίων η δομή και η σημασία δεν μπορούν να εξαχθούν από τις συνιστώσες λέξεις τους γιατί αυτές εμφανίζονται ως ανεξάρτητες.

[2] «Η αναγνώριση ονοματικών οντοτήτων (NER) είναι μια υποδιαίρεση της τεχνολογίας εξαγωγής πληροφοριών που επιδιώκει να εντοπίσει και να ταξινομήσει αναφορές ονομάτων οντοτήτων σε αδόμητο κείμενο σε προκαθορισμένες κατηγορίες, όπως τα ονόματα προσώπων, οργανισμούς, ιατρικούς κώδικες, χρονικές εκφράσεις, ποσότητες, νομισματικές αξίες, ποσοστά κ.λπ.

Παράδειγμα αναγνώρισης:

Ο Jim αγόρασε 300 μετοχές της Acme Corp. το 2006. -> [Jim] Το πρόσωπο αγόρασε 300 μετοχές της [Acme Corp.] Οργάνωσης το [2006] Time.» (Wikipedia t. f., 2019)

Οι βασικές εργασίες της Ε.Φ.Γ που σχετίζονται με τις Π.Λ.Ε είναι: (1) η εξακρίβωση και η εξαγωγή των Π.Λ.Ε από δεδομένα σωμάτων κειμένου και αποσαφήνιση της εσωτερικής σύνταξής τους και (2) η ερμηνεία των Π.Λ.Ε. Όλο και περισσότερο αυτές οι εργασίες συνδέονται με τη μηχανική μετάφραση. Γίνεται δηλαδή προσπάθεια για τον προσδιορισμό μεμονωμένων εμφανίσεων Π.Λ.Ε στο τρέχον κείμενο. Στην αναγνώριση των Π.Λ.Ε, βασικό πρόβλημα αποτελεί η διάκριση ανάμεσα σε Π.Λ.Ε και κυριολεκτικές χρήσεις για συνδυασμούς λέξεων όπως είναι το «γίνομαι άνθρωπος» που χρησιμοποιείται πολύ στο «Τρίτο Στεφάνι». Ενδεικτικά, πολυλεκτική χρήση είναι η ακόλουθη: «Μετά από αγώνα και σκληρή πάλη με τη ζωή έγινε επιτέλους άνθρωπος» ενώ μη πολυλεκτική χρήση η εξής: «Ο Θεός έγινε άνθρωπος για να σώσει τον άνθρωπο».

Παρακάτω παραθέτουμε μερικά παραδείγματα ως προς τη συχνότητα εμφάνισης των Π.Λ.Ε στο λόγο, μέσα σε 7 σελίδες από το έργο του Ταχτσή:

«Όχι, δεν είμαι όμορφη! Μα ξέρω να ζήσω. Ποια γυναίκα στην ηλικία μου θα βαστιόταν τόσο καλά, όσο βαστιέμαι εγώ; Όλες μου οι φίλες κι όλες μου οι συμμαθήτριες απ΄ τ΄ Αρσάκειο έχουν γεράσει. Τις βλέπω στο δρόμο και τρομάζω. Είναι κιόλας γιαγιάδες!… Όχι επειδή έχουν εγγόνια – η Ιουλία δεν έχει εγγόνια – αλλ΄ επειδή παραμέλησαν τον εαυτό τους. Αφέθηκαν και γέρασαν. Το σώμα δε γερνάει, αν δε γεράσει πρώτα η καρδιά. «Ας κάνουν οι κόρες μου λούσα!» σου λέει. «Ας πάνε τα παιδιά μου στους χορούς και στις διασκεδάσεις! Εγώ τά φαγα πια τα ψωμιά μου!» Μα το λένε, γιατί έχουν παιδιά που αξίζουν κάθε θυσία. Δεν έχουν τη Μαρία! Δεν ξέρουν τι θα πει νάχεις κόρη τη Μαρία, και γι΄ αυτό δεν τις αδικώ που με μέμφονται ότι ξαναπαντρεύτηκα, αντί να κοιτάξω να την παντρέψω. Δεν ξέρουν ότι την εποχή που απεφάσισα να κάνω το σάλτο να πάρω το Θόδωρο, ζύγισα όλα τα υπέρ και τα κατά. Η Μαρία, είπα με το νου μου, είναι σαν το ναυαγό που πνίγεται… Αν κάνω πως πάω κοντά της να τη σώσω, θα με παρασύρει και μένα στον πάτο. Ας σωθώ εγώ τουλάχιστον, για να της δώσω καιρό να μεγαλώσει λιγάκι, να ωριμάσει κάπως. «Πάντρεψέ την» μούλεγαν όλες, «και να δεις πως θα γίνει αγνώριστη». Να την παντρέψω εγώ; Ανημποριά έχει να βρει μόνη της γαμπρό; Εμένα στην ηλικία της με φλερτάριζαν δέκα άντρες συγχρόνως. Όπου πήγαινα κρεμόντουσαν απ΄ τη φούστα μου. Σ΄ όποιον νάλεγα «σε παίρνω», θάτρεχε με τα τέσσερα!.. Πώς έκανα –θα μού πείτε– τη στραβωμάρα να πάω να πέσω απάνω στο Φώτη, αυτό είν΄ άλλη ιστορία. Προτιμώ να μην το θυμάμαι, γιατί συγχύζομαι περισσότερο. Ίσως –λέω καμιά φορά με

το νου μου– νάταν γραφτό απ΄ το θεό να τον πάρω για να τραβήξω όσα τράβηξα. Γραφτό να γεννήσω αυτή τη Μέδουσα!… Άλλες πάλι φορές σκέπτομαι πως δε φταίει ούτ΄ ο θεός, ούτ΄ η μοίρα. Εγώ φταίω, κανείς άλλος! Ήμουνα πεισματάρα και πάτησα πόδι. Είπα «θα τον πάρω» και τον πήρα. Από ένα πείσμα. Ακριβώς επειδή δεν τον ήθελε κανένας απ΄ τους δικούς μου. Ούτε κι αυτός ακόμα ο συχωρεμένος ο μπαμπάς, που ήταν πάντα τόσο επιφυλακτικός στις κρίσεις του. Δεν εννοούσα να τους αφήσω ν΄ ανακατευτούν ακόμα μια φορά στις υποθέσεις μου και στη ζωή μου, όπως είχαν κάνει στο παρελθόν. Αρκετό κακό μούχαν κάνει με τις ανακατωσούρες τους στην περίπτωση του Αργύρη. Δεν ήμουνα πια δεκαοχτώ ετών, όπως άλλοτε. Ήμουνα εικοσιεφτά. Ήμουνα αυτεξούσια κι αποφασισμένη να κάνω του κεφαλιού μου, κι έβγαλα τα μάτια μου!… (Ταχτσής, 1987, σσ. 9-16)

2.1.2 Κατηγοριοποίηση των ΠΛΕ

Στο κεφάλαιο αυτό θα εξετάσουμε τις κατηγορίες των πολυλεκτικών εκφράσεων, θα τις διαχωρίσουμε μεταξύ τους και θα δώσουμε τους ορισμούς της κάθε επιμέρους περίπτωσης.

Α. Χωρισμός ως προς τη γραμματική μορφή

i. Ονοματικές Πολυλεκτικές Εκφράσεις

Οι ονοματικές Π.Λ.Ε εμφανίζονται πιο συχνά στις διάφορες γλώσσες. Είναι ακολουθίες που σχηματίζονται από ένα ουσιαστικό (κεφαλή) και άλλα στοιχεία που είναι προσαρτημένα σε αυτό, όπως άλλα ουσιαστικά, επίθετα και επιθετικούς προσδιορισμούς που εισάγονται από προθέσεις. Ο απλούστερος τύπος για μια ονοματική M.W.E είναι η ένωση 2 ουσιαστικών όπως «φακοί επαφής». Στην ίδια κατηγορία εντάσσονται και κύρια ονόματα όπως *Άγιος Στέφανος*. (Ramisch, 2015, σ. 42)

ii. Ρηματικές Πολυλεκτικές Εκφράσεις

Οι ρηματικές Π.Λ.Ε είναι αυτές στις οποίες κεφαλή είναι κάποιο ρήμα και χωρίζονται σύμφωνα με τους Baldwin, Timothy and Su Nam Kim (Baldwin, 2010) με τον εξής τρόπο:

- Δομή ελαφρού/υποστηρικτικού ρήματος (light/support-verb construction): ρήμα με αποδυναμωμένη σημασία + κατηγορικό όνομα άμεσα ή έμμεσα εξαρτώμενο

  - παίρνω απόφαση

  - to take a shower

- Δομή ρήματος-επιρρήματος (verb-particle construction): ρήμα + επίρρημα ή πρόθεση

  - βάζω μπρος, μπαίνω μέσα

  - to give up (Ramisch, 2015, σσ. 42-44)

Β. Χωρισμός ως προς τη συντακτική μορφή

i. Πολυλεκτικές εκφράσεις με σταθερή δομή (παγιωμένες εκφράσεις): Αυτές οι Π.Λ.Ε δεν αλλάζουν μορφοσυντακτική δομή ούτε επιδέχονται εσωτερικές αλλαγές. Π.χ «άσε μας τώρα» ή παροιμίες όπως «Ό,τι χορταράκι κοροϊδέψεις στην πόρτα σου θα φυτρώσει».

- Πολυλεκτικές εκφράσεις με σχετικά σταθερή δομή: Αυτές οι Π.Λ.Ε έχουν αυστηρούς κανόνες για την σειρά των λέξεων και τη σύνταξη αλλά επιτρέπουν κάποιο βαθμό αλλαγής όπως για παράδειγμα στην κλίση ή στις αυτοπαθείς αντωνυμίες που περιέχουν. Π.χ «εν συνεχεία» : Εν συνεχεία θα δούμε τις εξελίξεις στα οικονομικά.
- Πολυλεκτικές εκφράσεις με συντακτική ευελιξία: Αυτές οι Π.Λ.Ε παρουσιάζουν μια μεγάλη ευελιξία ως προς τη δομή τους. Π.χ «τη γλιτώνω παρά τρίχα» : Ουφ, παρά τρίχα τη γλίτωσα. Διαβάστε πως τη γλίτωσαν κυριολεκτικά στο παρά τρίχα 20 χολιγουντιανοί πρωταγωνιστές πασίγνωστων και δημοφιλών ταινιών! (Baldwin, 2010)

Γ. Χωρισμός ως προς τη σημασία

• Αμφίσημες: κυριολεκτική και μεταφορική σημασία

– αφήνω μισό κάποιον ή κάτι

• Πολύσημες:

– βγάζω τα σωθικά μου

(Markantonatou, Panagiotis, George, Vassiliki, & Maria, 2019) (Μαρκαντωνάτου, 2017, σσ. 11-13)

Δ. Χωρισμός ως προς την παγίωση

Οι ΠΛΕ είναι παγιωμένες εκφράσεις, για την ακρίβεια οι δύο όροι χρησιμοποιούνται εναλλακτικά. Στην ενότητα αυτή θα δούμε τις ΠΛΕ από την άποψη της παγίωσης. «Η παγίωση είναι η διαδικασία με την οποία μια ομάδα λέξεων της οποίας τα στοιχεία είναι ελεύθερα γίνονται μία έκφραση της οποίας τα στοιχεία δεν διαχωρίζονται.» Η ενότητα των λέξεων που δημιουργείται δηλαδή είναι αυτόνομη και έχει ολοκληρωμένο νόημα ανεξάρτητα από τη σημασία των εκάστοτε λέξεων που την αποτελούν. Οι παγιωμένες εκφράσεις βρίσκονται στο νοητικό λεξικό του ομιλητή και αναπαράγονται στη χρήση του λόγου. Επειδή όμως οι παγιωμένες εκφράσεις δεν αναγνωρίζονται από την εξωτερική τους μορφή ως λεξικοποιημένες ενότητες (δεν έχουν μορφοσυντακτικές αποκλίσεις) δεν είναι εύκολος ο εντοπισμός τους. Τέλος κάποιες παγιωμένες εκφράσεις μπορεί να λειτουργούν και ως ελεύθεροι συνδυασμοί ανάλογα με τα συμφραζόμενα» (Χιώτη, 2010, σ. 13)

Η εξέλιξη ελεύθερων συνδυασμών λέξεων σε παγιωμένες εκφράσεις είναι μια διαδικασία με μεταβατικά στάδια, χωρίς αυτό να σημαίνει απαραίτητα ότι μια ελεύθερη ακολουθία θα γίνει σταδιακά παγιωμένη. (Χιώτη, 2010, σ. 14) Όταν μια έκφραση περνάει από την κατηγορία των ελεύθερων συνδυασμών λέξεων και εγκαθίσταται στο νοητικό λεξικό ως παγιωμένη, συνεπάγεται ότι ο μέσος χρήστης της έκφρασης την αναγνωρίζει στις περισσότερες περιπτώσεις χωρίς όμως να γνωρίζει αναγκαστικά τη σημασία της σε όλες τις περιπτώσεις. (Συμεωνίδης, 2000, σ. 50)

Πρόκειται επομένως για συγκεκριμένο συνδυασμό λέξεων που δεν κατασκευάζεται από τους ομιλητές μιας φυσικής γλώσσας αλλά λειτουργούν ως ένα ενιαίο σύνολο στο λόγο. Ο όρος «παγιωμένη» δηλώνει το γεγονός ότι η έκφραση απομνημονεύεται από τους ομιλητές της γλώσσας οι οποίοι γνωρίζουν ότι οι λέξεις εμφανίζονται μαζί σε αυτή ή σε κάποια άλλη σύνταξη και η χρήση της έκφρασης είναι συμβατικά καθιερωμένη και μοιράζεται από τους ομιλητές της συγκεκριμένης γλώσσας. Οι παγιωμένες εκφράσεις δεν δημιουργούνται κάθε φορά ως στοιχεία της ομιλίας ή του κειμένου αλλά χρησιμοποιούνται από τους ομιλητές ως έτοιμες μονάδες. Έτσι ο ρόλος του ομιλητή μηδαμινός και δεν υπάρχει σχετική ελευθερία στην ευελιξία του ως προς αυτές. (Χιώτη, 2010, σσ. 14-15) Πρόβλημα αποτελεί επίσης για τους ξένους ομιλητές της γλώσσας η κατανόηση των παγιωμένων εκφράσεων καθώς συνηθίζεται η λέξη προς λέξη μετάφραση κάτι το οποίο δεν είναι γόνιμο στην περίπτωση αυτή.

Χωρίς αμφιβολία πρέπει να τονίσουμε ότι οι παγιωμένες εκφράσεις αντανακλούν πράξεις και αξίες της καθημερινής ζωής και έτσι θεωρούνται σπουδαία πηγή για τη μελέτη του συλλογικού βίου από διάφορους ερευνητές. Ωστόσο οι παγιωμένες εκφράσεις είναι και ένα κομμάτι φευγαλέο· πολλές είναι εφήμερες, άλλες συγχέονται με τις παροιμίες και στα λεξικά δεν είναι πάντοτε εύκολος ο εντοπισμός τους. (Σαραντάκος, 1997, σ. 7)

Ένα από τα ζητήματα που απασχόλησαν τους ερευνητές είναι ο ορισμός των παγιωμένων εκφράσεων δεδομένου του γεγονότος ότι δεν παρουσιάζονται ως ένα ενιαίο σύνολο με κοινά χαρακτηριστικά. Η Cacciari αναφέρει ότι ένας από τους λόγους που ο ορισμός είναι δύσκολος είναι επειδή η παγίωση είναι μια διαδικασία. Μια δεδομένη σύνταξη δεν είναι παγιωμένη μια για πάντα, αλλά αποκτά την παγίωση σταδιακά. Έτσι οι παγιωμένες εκφράσεις δεν είναι όλες ισοδύναμες, καθώς υπάρχουν οι μερικώς παγιωμένες αλλά και οι εντελώς παγιωμένες εκφράσεις. Επιπλέον ενώ η συνολική σημασία μιας παγιωμένης έκφρασης δεν προκύπτει από τη σημασία των λέξεων που την αποτελούν, οι ομιλητές έχουν για τις περισσότερες παγιωμένες εκφράσεις έντονη διαίσθηση όσον αφορά τις σχέσεις ανάμεσα στη σημασία των λέξεων και τη σημασία της έκφρασης π.χ «Μαύρα μάτια κάναμε να σε δούμε». Εδώ τα μαύρα μάτια έχουν από μόνα τους μα σημασία αλλά σαν έκφραση σημαίνει ότι έχουμε να δούμε πολύ καιρό κάποιον. (Cacciari, 1993, pp. 27-28)

2.2.1 Κατηγοριοποίηση των παγιωμένων εκφράσεων

Οι παγιωμένες εκφράσεις όσον αφορά το περιεχόμενό τους παρουσιάζουν την εξής εικόνα :

1. Μέλη/όργανα του σώματος

2. Περιβάλλον

3. Κοινωνική ζωή: χρήματα/μέτρα/όπλα/παιχνίδια

4. Βίος/χρόνος: ζωή - άνθρωπος - παιδί - μέρα - νύχτα - ώρα

5. Θρησκευτικές έννοιες : θεός - άγγελος

6. Ήθη, έθιμα, συναισθήματα: δίκαιος, χαρά, καλός

7. Βασικές δραστηριότητες: έχω, έρχομαι, βγάζω, είμαι, κάνω.

«Πολλές είναι και οι ιστορικές παγιωμένες εκφράσεις, δηλαδή φράσεις γεννημένες από ιστορικά ή τοπικά περιστατικά, ή που ειπώθηκαν από ιστορικά πρόσωπα, π.χ. διέβη το Ρουβίκωνα.» (Χιώτη, 2010, σ. 166) Επιπλέον έχουμε και πιο πρόσφατες όπως : «του έκανε τη ζωή πατίνι». Ακόμα έχουμε φράσεις από τη λογοτεχνία π.χ. «περασμένα μεγαλεία και διηγώντας τα να κλαις» από τον *Ύμνο εις την Ελευθερίαν* του Σολωμού.

Όσον αφορά τη χρονική προέλευση των παγιωμένων εκφράσεων διακρίνουμε εκφράσεις από τη 1) νέα ελληνική γλώσσα αλλά και από την 2) αρχαία π.χ «κτῆμα ἐς αἰεί» ή «οὐκ ἂν λάβοις παρὰ τοῦ μὴ ἔχοντος». Τέτοιες εκφράσεις έχουν διατηρήσει την αρχική τους μορφή στη νέα ελληνική. Εκτός όμως από την ελληνική γλώσσα έχουμε 3) παγιωμένες εκφράσεις λατινικής προέλευσης που διατηρούνται από την περίοδο της επικράτησης των Ρωμαίων μέχρι σήμερα. Έτσι έχουμε εκφράσεις όπως (ab ovo, o tempora o mores, alter ego, lapsus linguae, mutatis mutandis). Ειδικά αυτές που πέρασαν στον ελληνικό λόγο με τη μετάφρασή τους χρήζουν μελέτης π.χ (alea iacta est/ο κύβος ερρίφθη). (Χιώτη, 2010, σσ. 168-171) Ακόμα έχουμε 4) παγιωμένες εκφράσεις που απαντούν και σε βυζαντινά κείμενα, λαϊκά ή λόγια όπως : «ταῦτα εἰς νοῦν βαλόμενος Ἀλέξιος (= βάζει στο νου του, Άννης Κομνηνής, Ἀλεξιάς, 1,152,21), ἀπὸ τὰ νύχια ἕως τὴν κορυφὴν ἦτον ἀρματωμένος (= από την κορυφή ως τα νύχια, Ἀχιλληΐς, 138), καὶ τόπος οὐ χωρεῖ σε (= δεν τον χωράει ο τόπος, Μιχαήλ Γλυκάς, στ.

156)» Άλλη μία σημαντική πηγή παγιωμένων εκφράσεων αποτελεί 5) η γλώσσα της εκκλησίας κυρίως με φράσεις από την Παλαιά και την Καινή Διαθήκη π.χ. (κατ' εικόνα και καθ' ομοίωσιν, ελήλυθεν η ώρα) κ.α. Ακόμα έχουμε 6) παγιωμένες εκφράσεις που προέρχονται από την καθαρεύουσα π.χ (εν λευκώ, εφ' όλης της ύλης, επ' αυτοφώρω).

Επιπλέον αξίζει να τονίσουμε ότι έχουμε παγιωμένες εκφράσεις 7) από τα Τουρκικά και γενικότερα από τις βαλκανικές γλώσσες κάτι που είναι φυσικό αφού οι δύο γλώσσες συνυπήρξαν για πολλά χρόνια. Το γεγονός των δανείων παγιωμένων εκφράσεων ιδιαίτερα στις βαλκανικές χώρες μπορεί να οφείλεται και σε παράλληλες εξελίξεις καθώς και σε κοινές ανάγκες. Ο Συμεωνίδης παραθέτει πολλά παραδείγματα εκφράσεων που υπάρχουν π.χ σε τρεις γλώσσες (τουρκικά, ελληνικά, βουλγαρικά,) : ekmeğini çıkarmak = βγάζω το ψωμί μου = vadja si hljaba. (Χιώτη, 2010, σσ. 172-179)

2.2.2 Διαφορές Παγιωμένων Εκφράσεων-Παροιμιών

Πολλές φορές έχει παρατηρηθεί η σύγχυση των παγιωμένων εκφράσεων με τις παροιμίες λόγω των κοινών χαρακτηριστικών τους όμως όπως προκύπτει διαφέρουν μεταξύ τους. Η παροιμία (αγγλικά: proverb) είναι λαϊκή φράση η οποία επιγραμματικά εκφράζει συνήθως με τρόπο αλληγορικό, μεταφορικό ή σκωπτικό, μια αλήθεια για τη ζωή, μια γνώμη που πηγάζει από τη μακρόχρονη κοινή πείρα. (Μπαμπινιώτης, 1998) «Μπορεί να είναι έμμετρη ή πεζή και λέγεται για να παραδειγματίσει, να διδάξει ή να σχολιάσει μια κατάσταση. Αναφέρεται σε γενικά θέματα για τη φύση της ύπαρξής μας, τη φύση των ανθρώπων και των καταστάσεων που αντιμετωπίζουμε και το ρόλο μας στον κόσμο. Είναι λαϊκό δημιούργημα, όπως και το δημοτικό τραγούδι, και με το λαό συνδέεται συχνά στο λόγο από τους ομιλητές («ο λαός λέει μια παροιμία…»).» (Χιώτη, 2010)

Οι παροιμίες διαφέρουν από τις παγιωμένες εκφράσεις 1) για τη γενική τους αξία ως συμβουλές ή διαχρονικές αλήθειες για την ανθρώπινη ζωή αλλά και 2) στον τρόπο που λειτουργούν στο λόγο. Ως προς τη γενική αξία τους ανήκουν στις γενικευτικές προτάσεις οι οποίες εκφράζουν μια σχέση που ισχύει ανεξάρτητα από ειδικές καταστάσεις, δηλαδή δεν υπάρχει άμεση αντιστοιχία με τον τωρινό κόσμο. Για το λόγο αυτό υπάρχουν παροιμίες με αντικρουόμενο περιεχόμενο, π.χ. *το γοργό και χάριν έχει – όποιος βιάζεται σκοντάφτει*. (Άννα Αναστασιάδη - Συμεωνίδη, 2006, σ. 68). Όσον αφορά τον τρόπο που λειτουργούν στο λόγο, απαντώνται όπως τα παραθέματα, αφού

ο ομιλητής στην πραγματικότητα δεν δημιουργεί ο ίδιος τη γενικευτική πρόταση, η οποία προϋπάρχει της συγκεκριμένης περίστασης, απλά την εκφωνεί. (Άννα Αναστασιάδη - Συμεωνίδη, 2006, σ. 69) Μια παροιμία λοιπόν μπορεί να σταθεί από μόνη της αυτοτελώς στο λόγο χωρίς την ανάγκη ύπαρξης προηγούμενης ή επόμενης έκφρασης για να δημιουργηθεί μια ολοκληρωμένη δήλωση. «Με άλλα λόγια οι παροιμίες έχουν σημασιολογική και πραγματολογική αυτονομία, αφού δεν συνδέονται χωρο-χρονικά με το υπόλοιπο κείμενο, καθώς και κειμενική αυτονομία.» (Χιώτη, 2010, σ. 38)

«Αντίθετα οι παγιωμένες εκφράσεις παραπέμπουν σε ιδιαίτερη περίσταση χωρο-χρονικά προσδιορισμένη. Αυτό αποδεικνύεται και από το ότι οι παγιωμένες εκφράσεις εγγράφονται στο χρονικό σύστημα και εξαρτώνται από τα συμφραζόμενα της πρότασης.» (Χιώτη, 2010, σ. 37)

Οι παροιμίες όπως αναφέραμε και παραπάνω δεν είναι δημιούργημα του ομιλητή και για αυτό το λόγο πρέπει να μαθαίνονται απέξω. Μεταφέρονται από γενιά σε γενιά κυρίως μέσω του προφορικού λόγου και μέσω της πιστής μετάδοσης μέσα στο χρόνο προκύπτει η παγιωμένη μορφή τους. Αν και μπορεί να παρουσιάζουν σχετική ποικιλομορφία, θεωρούνται ήδη ολοκληρωμένες φράσεις με στατιστικά σπάνιες παραλλαγές. «Οι παροιμίες χαρακτηρίζονται από μη δυνατότητα σύνθεσης και αδυναμία πραγμάτωσης μετασχηματισμών. Για το λόγο αυτό είναι ο τελευταίος βαθμός παγίωσης.» (Χιώτη, 2010, σ. 38)

ΚΕΦΑΛΑΙΟ 3

ΔΙΑΘΕΣΙΜΑ ΕΡΓΑΛΕΙΑ

Στις μέρες μας ο Παγκόσμιος Ιστός θεωρείται μια αστείρευτη πηγή δεδομένων γλώσσας σε συνδυασμό με την ευκολία πρόσβασης που μας παρέχει. Αν και όταν μιλάμε για διαδίκτυο πρέπει να έχουμε στο μυαλό μας και κάποια μειονεκτήματα όπως είναι η έλλειψη ελέγχου και η ανεπαρκής τεκμηρίωση, παρ' όλα αυτά το διαδίκτυο συνιστά κατάλληλη πηγή σωμάτων κειμένων που έχουν δημιουργηθεί για συγκεκριμένους σκοπούς (π.χ. εργασία μετάφρασης ή διερμηνείας, σύνταξη μιας ορολογικής βάσης δεδομένων, εξειδικευμένες λειτουργίες μηχανικής μάθησης κ.α). Αυτά τα σώματα κειμένων αξιοποιούνται σε μεγάλο βαθμό από τους διάφορους ερευνητές της γλώσσας για να πετύχουν τον στόχο τους με τον πιο γρήγορο και αποτελεσματικό τρόπο.

Οι επιστήμονες και επαγγελματίες της γλώσσας στρέφονται ολοένα και περισσότερο στον ιστό ως πηγή δεδομένων γλώσσας λόγω του όγκου δεδομένων που μπορεί να παρέχει, επειδή είναι η μόνη διαθέσιμη πηγή για τον τύπο της γλώσσας που τους ενδιαφέρει ή απλά επειδή είναι δωρεάν. Γίνεται λοιπόν χρήση του Web για την κατασκευή διαφόρων τύπων συλλογών κειμένων, συμπεριλαμβανομένων μονόγλωσσων, συγκρίσιμων, παράλληλων και συλλογών κειμένων συγκεκριμένου τομέα. Τέτοιοι πόροι μπορούν να χρησιμοποιηθούν από τους γλωσσολόγους που μελετούν τη χρήση της γλώσσας και ταυτόχρονα μπορούν να αξιοποιηθούν σε τομείς εφαρμοσμένης έρευνας όπως η μηχανική μετάφραση και η εξαγωγή πολυγλωσσικών πληροφοριών. Επιπλέον, αυτές οι συλλογές ακατέργαστων δεδομένων μπορούν να επισημανθούν αυτόματα μέσω της χρήσης προγραμμάτων πληροφορικής και να χρησιμοποιηθούν για την παραγωγή διάφορων χρήσιμων εργαλείων όπως : λεξικά.

Ενώ είναι δυνατή η κατασκευή ενός web-based corpus μέσω χειροκίνητων ερωτημάτων και λήψεων, π.χ με αναζήτηση Google, αυτή η διαδικασία είναι εξαιρετικά χρονοβόρα σε βαθμό που να θεωρείται μη προτιμητέα και αποτελεσματική ειδικά εάν το τελικό αποτέλεσμα προορίζεται να είναι ένα σώμα κειμένων μιας χρήσεως.» (Bootcat, 2018).

Σε αυτήν την εργασία συγκρίνουμε την αποτελεσματικότητα δύο μεθόδων συγκέντρωσης κειμένων από το διαδίκτυο με συγκεκριμένο στόχο: την χειροκίνητη

συγκέντρωση και την χρήση εξειδικευμένου θεματικού crawler. Ο στόχος είναι η δημιουργία συλλογών κειμένων πλούσιων σε εκ των προτέρων ορισμένων ΠΛΕ με στόχο να μελετηθούν οι σημασιοσυντακτικές ιδιότητές τους.

3.1 ILSP Focused Crawler

«Ο ILSP-FC Focused Crawler (ILSP-FC) είναι ένα ερευνητικό πρωτότυπο για την απόκτηση μονόγλωσσων και δίγλωσσων σωμάτων κειμένων συγκεκριμένου τομέα. Ο ILSP-FC αναπτύχθηκε από ερευνητές του ILSP/Athena RIC και χρησιμοποιείται για τον συντονισμό των ευρωπαϊκών γλωσσικών πόρων.» (ILSP Focused Crawler, 2019) Ένα παράδειγμα ερευνητικής χρήση του crawler αποτελεί η υποβολή του από το Ινστιτούτο Επεξεργασίας του Λόγου / Κέντρο Έρευνας και Καινοτομίας Αθηνά (ILSP / ARC) για την κοινή εργασία φιλτραρίσματος WMT (Workshop on Statistical Machine Translation) 2018 Parallel Corpus. Εξετάστηκαν διάφορες ιδιότητες των προτάσεων και των ζευγών προτάσεων που συγκεντρώνει το σύστημα στο πλαίσιο της εργασίας με σκοπό τη συσσώρευση ζευγών προτάσεων ανάλογα με την καταλληλότητά τους στην εκπαίδευση των συστημάτων μηχανικής μετάφρασης. (The ILSP/ARC submission to the WMT 2018 Parallel Corpus Filtering Shared Task, 2018, σσ. 928-933)

Μια πρώτη έκδοση του crawler δημιουργήθηκε κατά τη διάρκεια ενός έργου που χρηματοδοτήθηκε από την ΕΕ για την παραγωγή γλωσσικών πόρων. Στη συνέχεια επεκτάθηκε με σκοπό την υπέρβαση των φραγμών ποιότητας στη μηχανική μετάφραση, την ανθρώπινη μετάφραση και στις γλωσσικές τεχνολογίες. Επίσης επεκτάθηκε αργότερα και για την ενίσχυση της συνεργασίας βιομηχανίας-ακαδημαϊκών κύκλων για την υιοθέτηση τεχνολογιών μηχανικής μετάφρασης. (ILSP Focused Crawler, 2019)

3.1.1 Τρόπος Λειτουργίας

Η διερεύνηση του ιστού για την κατασκευή μονόγλωσσων και / ή παράλληλων δεδομένων που σχετίζονται με συγκεκριμένο τομέα περιλαμβάνει διάφορες εργασίες (π.χ. ταξινόμηση δεσμών, καθαρισμός, ταξινόμηση κειμένου, αφαίρεση σχεδόν διπλότυπων). Έχουν προταθεί αρκετές μέθοδοι για κάθε μία από αυτές τις εργασίες. Ο ILSP Focused Crawler ξεκινάει από (τον κατάλογο των σελίδων προς επίσκεψη), δηλαδή από μια λίστα διευθύνσεων σπόρων (seeds) που παρέχει ο χρήστης, ταξινομεί τις ανακτημένες σελίδες ως σχετικές με τον στοχευόμενο τομέα, αποσπά τις συνδέσεις από τις ανεβασμένες ιστοσελίδες και τις προσθέτει στη λίστα των σελίδων που πρόκειται να επισκεφτεί.

Προκειμένου να διασφαλιστεί η δυνατότητα επέκτασης και κλιμάκωσης-τροποποίησης του crawler για μελλοντικές έρευνες, για την κατασκευή του χρησιμοποιήθηκε το Bixo[3], ένα εργαλείο ανοιχτού κώδικα που επιτρέπει την εύκολη διαμόρφωση ροών εργασίας και λειτουργεί πάνω από το πλαίσιο Hadoop7 για κατανεμημένη επεξεργασία δεδομένων. (Prokopis Prokopidis, 2019)

Σύμφωνα με τους κατασκευαστές του, το πρώτο τμήμα του crawler αφορά τη λήψη σελίδας. Έχει υιοθετηθεί η εφαρμογή πολυνηματικού crawling για να εξασφαλιστεί ταυτόχρονη επίσκεψη σε πολλές ιστοσελίδες. Οι χρήστες μπορούν να διαμορφώσουν πολλές ρυθμίσεις που καθορίζουν τη διαδικασία ανάκτησης, συμπεριλαμβανομένου του αριθμού των ταυτόχρονων συλλεκτών και το φιλτράρισμα ειδικών τύπων εγγράφων. Ο ανιχνευτής μπορεί να επηρεαστεί από τη χρήση ρυθμίσεων ανά χρονικά διαστήματα για την αναθεώρηση διευθύνσεων URL από τον ίδιο τον ιστότοπο, τον μέγιστο αριθμό διευθύνσεων URL από συγκεκριμένο κεντρικό υπολογιστή ανά επανάληψη, τον μέγιστο αριθμό προσπαθειών για να φέρει μια ιστοσελίδα κ.λπ.

Ακόμη και όταν μια ιστοσελίδα δεν αποθηκεύεται (γιατί θεωρήθηκε άσχετη με τον τομέα ή την στοχευμένη γλώσσα), οι δεσμοί της εξάγονται και προστίθενται στη λίστα των συνδέσμων που έχουν προγραμματιστεί για επίσκεψη. Οι σύνδεσμοι πρέπει να ταξινομηθούν και οι πιο ελπιδοφόρες συνδέσεις (δηλαδή οι σύνδεσμοι που οδηγούν σε

[3] « Το Bixo είναι ένα εργαλείο ανοιχτού κώδικα για την εξόρυξη ιστού που συμπληρωματικά με το Hadoop μπορεί να δημιουργήσει γρήγορα εξειδικευμένες εφαρμογές εξόρυξης ιστού που είναι βελτιστοποιημένες για μια συγκεκριμένη περίπτωση χρήσης.» (Bixo, 2019)

"ιστοσελίδες" ή (υποψήφιες μεταφράσεις) θα πρέπει να ακολουθηθούν πρώτα. Για το σκοπό αυτό υιοθετείται η χρήση βαθμολογίας συνάφειας.

Η υπομονάδα κανονικοποίησης χρησιμοποιεί το πακέτο εργαλείων Apache Tika 8[4] για να αναλύσει τη δομή κάθε ληφθείσας ιστοσελίδας και να εξαγάγει τα μεταδεδομένα της. Αν τα μεταδεδομένα θεωρηθούν σχετικά εξάγονται. Η κωδικοποίηση κειμένου της ιστοσελίδας ανιχνεύεται με βάση την κεφαλίδα περιεχομένου-κωδικοποίησης HTTP[5] και αν χρειαστεί το τμήμα περιεχομένων μετατρέπεται σε UTF-8.[6]

Εκτός από το κείμενο, μια ιστοσελίδα περιέχει επίσης το λεγόμενο boilerplate, δηλαδή στοιχεία όπως κεφαλίδες πλοήγησης, διαφημίσεις, αποποιήσεις ευθυνών κλπ., οι οποίες δεν είναι χρήσιμες για την παραγωγή γλωσσικών πόρων. Για την αφαίρεση του boileplate, ο crawler χρησιμοποιεί μια τροποποιημένη έκδοση του Boilerpipe 9 (Kohlschuetter et al, 2010) που εξάγει επίσης διαρθρωτικές πληροφορίες όπως τίτλο, επικεφαλίδα και στοιχεία καταλόγου. Σε αυτό το στάδιο, το κείμενο είναι κατακερματισμένο σε παραγράφους με βάση συγκεκριμένες ετικέτες HTML όπως <p>,
 και <li>. Οι παράγραφοι που κρίνονται ότι ανήκουν στο boillerplate και / ή ανιχνεύονται ως τίτλοι κ.λπ. είναι κατάλληλα υποσημειωμένες ώστε να αφαιρεθούν από το χρήστη αν χρειαστεί.

Η επόμενη ενότητα επεξεργασίας (module) ασχολείται με τη γλωσσική ταυτοποίηση. Χρησιμοποιείται μια βιβλιοθήκη γλωσσικής αναγνώρισης για την αναγνώριση γλώσσας. Εάν μια ιστοσελίδα δεν βρίσκεται στη στοχευμένη γλώσσα, η μόνη περαιτέρω χρήση της είναι η εξόρυξη νέων δεσμών. Το αναγνωριστικό γλώσσας εφαρμόζεται σε κάθε παράγραφο και επισημαίνει τα κείμενα ως σωστά ή λανθασμένα.

Στη συνέχεια άλλη ενότητα του crawler προσδιορίζει εάν μια σελίδα που είναι ομαλοποιημένη και στην στοχευμένη γλώσσα περιέχει δεδομένα σχετικά με τον στοχευόμενο τομέα. Για το σκοπό αυτό συγκρίνεται το περιεχόμενο της σελίδας για έναν ορισμένο τομέα που παρέχει ο χρήστης. Αυτό επιτυγχάνεται με μια μέθοδο

[4] Το σετ εργαλείων Apache Tika ™ εντοπίζει και εξάγει τα μεταδεδομένα και το κείμενο από διαφορετικούς τύπους αρχείων (όπως PPT, XLS και PDF). Όλοι αυτοί οι τύποι αρχείων μπορούν να αναλυθούν μέσω μιας ενιαίας διεπαφής, καθιστώντας την Tika χρήσιμη για την ευρετηρίαση των μηχανών αναζήτησης, την ανάλυση περιεχομένου, τη μετάφραση και άλλα. (Foundation, 2019)

[5] Το πρωτόκολλο μεταφοράς υπερκειμένου (HTTP) είναι ένα πρωτόκολλο που χρησιμοποιείται από τα πληροφοριακά συστήματα και είναι η βάση της επικοινωνίας δεδομένων για τον Παγκόσμιο Ιστό, όπου τα έγγραφα υπερκειμένου περιλαμβάνουν υπερσυνδέσεις με άλλους πόρους στους οποίους ο χρήστης μπορεί να μεταβεί με ένα κλίκ.

[6] Το UTF-8 είναι ένα πρότυπο κωδικοποίησης χαρακτήρων

αντιστοίχησης συμβολοσειρών που χρησιμοποιεί τριπλέτες όρων (τριάδες) με σχετικό βάρος που περιγράφουν έναν τομέα και, προαιρετικά, υποκατηγορίες αυτού του τομέα για να προσδιορίσει τη σχετικότητα του περιεχομένου.

Έπειτα η υπομονάδα (Εξαγωγέας) δημιουργεί ένα αρχείο XML για κάθε αποθηκευμένο έγγραφο ιστού. Κάθε αρχείο περιέχει μεταδεδομένα (π.χ. γλώσσα, τομέας, διεύθυνση URL κ.λπ.) σχετικά με το αντίστοιχο έγγραφο μέσα σε ένα στοιχείο κεφαλίδας (header). Επιπλέον, περιέχει ένα <body> στοιχείο και το περιεχόμενο του εγγράφου χωρίζεται σε παραγράφους. Εκτός από το (κανονικοποιημένο) κείμενο, κάθε στοιχείο <p> παραγράφου εμπλουτίζεται με χαρακτηριστικά που παρέχουν περισσότερες πληροφορίες σχετικά με το αποτέλεσμα της διαδικασίας.

Το κείμενο έχει υποστεί ένα είδος επεξεργασίας κατά την οποία αποκόπτονται οι ‘καταλήξεις’ των λέξεων και μένουν τα ‘θέματα’. Δεν πρόκειται για μορφολογική ανάλυση με βάση κάποιες αρχές, απλά το εργαλείο ‘μαθαίνει’ κάποιες καταλήξεις τις οποίες αποκόπτει (δείγμα αυτής της λειτουργίας του εργαλείου δίνεται στο παράρτημα με το σχετικό log file). Ο μηχανισμός αυτός, όμως, αντιμετωπίζει σε ένα βαθμό το θέμα της μορφολογίας της Ελληνικής και έτσι δεν είναι αναγκαίο να εισάγει ο χρήστης όλο το μορφολογικό παράδειγμα για ομαλά κλινόμενες λέξεις, π.χ αρκεί το «κόκκινος σαν αστακός» για να καλυφθεί όλο το μορφολογικό παράδειγμα (το επίθετο «κόκκινος» και το ουσιαστικό «αστακός» κλίνονται ομαλά).

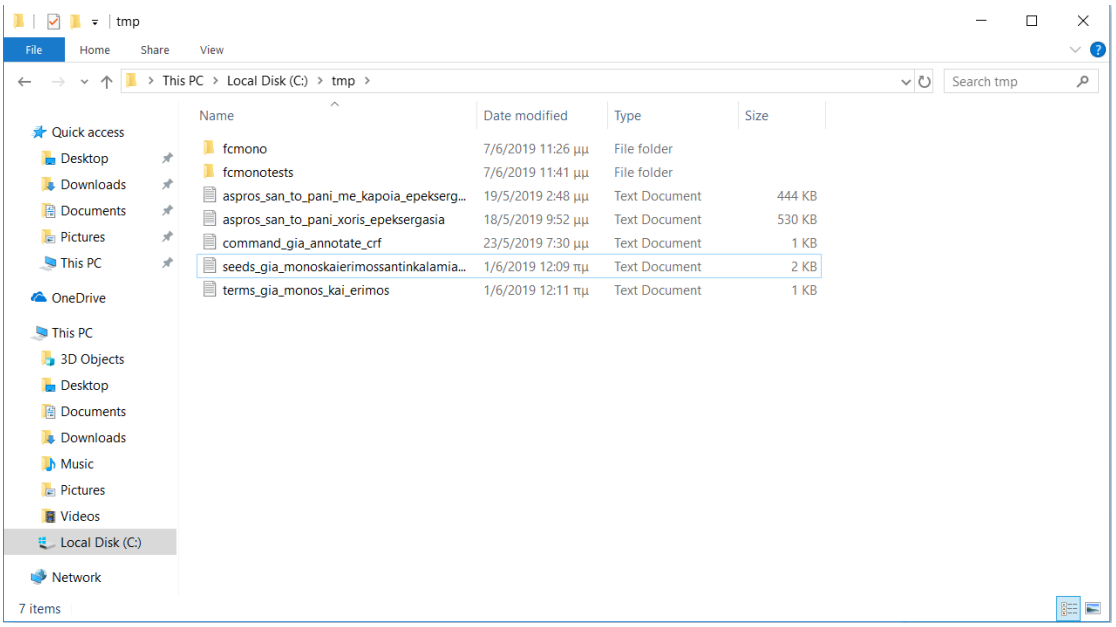
Τέλος, πράγμα που είναι εξαιρετικά σημαντικό για την έρευνά μας, το να αγνοηθεί το γεγονός ότι ο ιστός περιέχει πολλά σχεδόν διπλότυπα έγγραφα θα μπορούσε να έχει αρνητικό αποτέλεσμα στη δημιουργία ενός αντιπροσωπευτικού σώματος κειμένων. Έτσι το πρόγραμμα ανίχνευσης περιλαμβάνει μια ενότητα που ξεχωρίζει τα διπλότυπα κείμενα. Αφού φορτωθούν οι σελίδες εντός του τομέα, η ενότητα ανίχνευσης ζεύγους χρησιμοποιεί δύο συμπληρωματικές μεθόδους για τον προσδιορισμό ζευγών σελίδων που θα μπορούσαν να θεωρηθούν παράλληλες. Η πρώτη μέθοδος βασίζεται σε συσχέτιση, σε δύο έγγραφα, εικόνων με το ίδιο όνομα αρχείου, ενώ το δεύτερο λαμβάνει υπόψη δομική ομοιότητα. (Vassilis Papavassiliou, 2012, σσ. 1-9)

3.1.2 Γραφικό περιβάλλον και εντολές Crawler

Για να τρέξουμε το πρόγραμμα του crawler πρέπει να γνωρίζουμε επακριβώς τη θέση του αρχείου μέσα στον υπολογιστή όπως και τις θέσεις των υπόλοιπων αρχείων

που θα χρειαστεί να του δώσουμε ως βάση για την εργασία του δηλαδή το αρχεία με τους όρους (terms) και το αρχείο με τις λέξεις-σπόρους (seed urls).

Για παράδειγμα, στη δική μας περίπτωση κατασκευάσαμε έναν τέτοιο φάκελο στη διαδρομή : « C:\tmp». Επίσης δημιουργήσαμε ένα φάκελο «fcmonotests» για τα αποτελέσματα του crawler και ένα φάκελο «fcmono» για τα seed urls και τα terms.

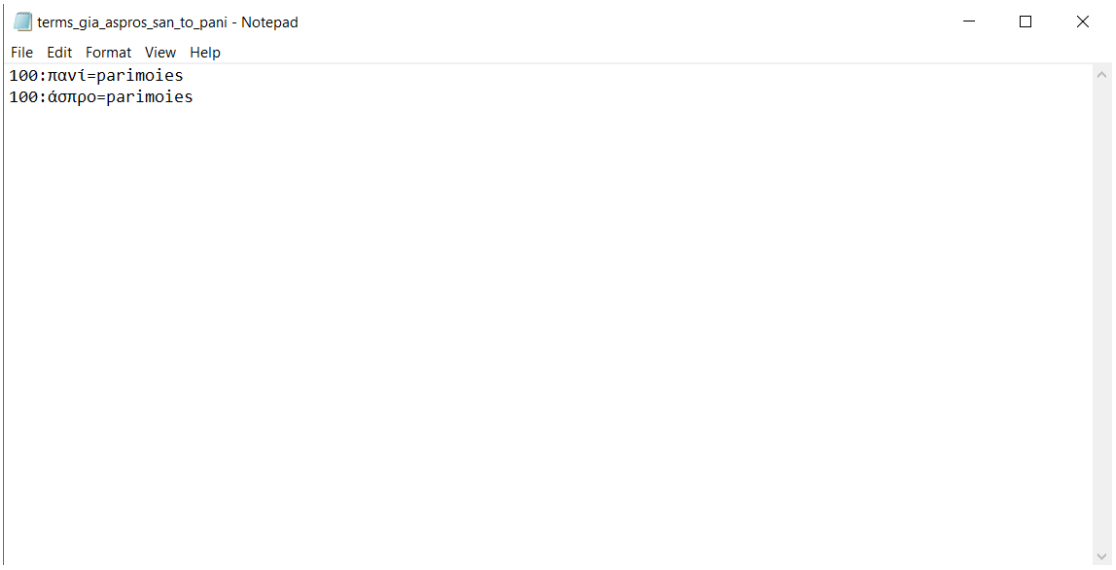


Εικόνα 4 Παράδειγμα προγράμματος μέσα σε φάκελο

Τέλος πρέπει να δημιουργήσουμε τα 2 αρχεία «.txt» για τα (seed urls) και τα (terms) της αναζήτησής μας. Για το αρχείο με τα seeds αρκεί να κάνουμε μια αρχική αναζήτηση σε μια μηχανή αναζήτησης και να καταγράψουμε χωρισμένα με «enter» τα links όπως παρακάτω.

Εικόνα 5 Αρχείο με παράδειγμα από Seeds

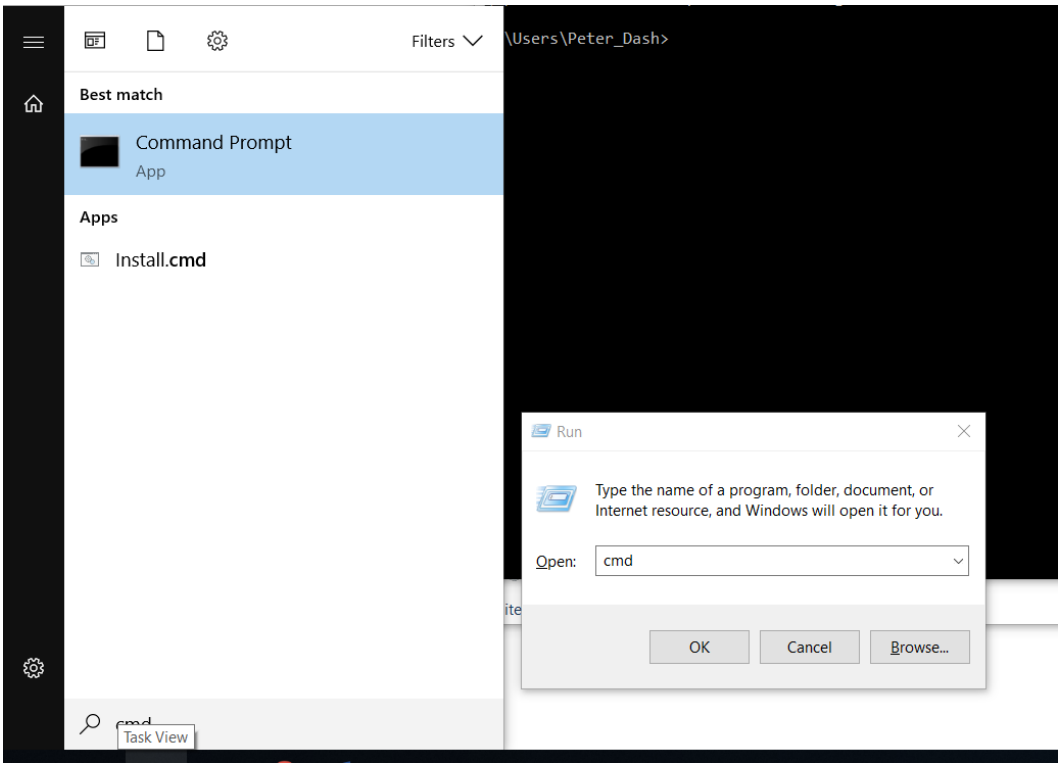
Όσον αφορά τα (terms) (όρους αναζήτησης) δημιουργούμε ένα αρχείο που θα περιέχει τους όρους που θα αναζητήσει ο crawler π.χ για την πολυλεκτική έκφραση : *άσπρος σαν το πανί* θα χρησιμοποιούσαμε τις λέξεις (άσπρος, πανί). Έπειτα προσθέτουμε το βάρος που έχουν οι λέξεις δηλαδή αν η μία είναι πιο σημαντική από την άλλη όσον αφορά την αναζήτησή μας και ορίζουμε ένα θέμα για τους όρους μας όπως παρακάτω.



Εικόνα 6 Παράδειγμα από αρχείο με terms

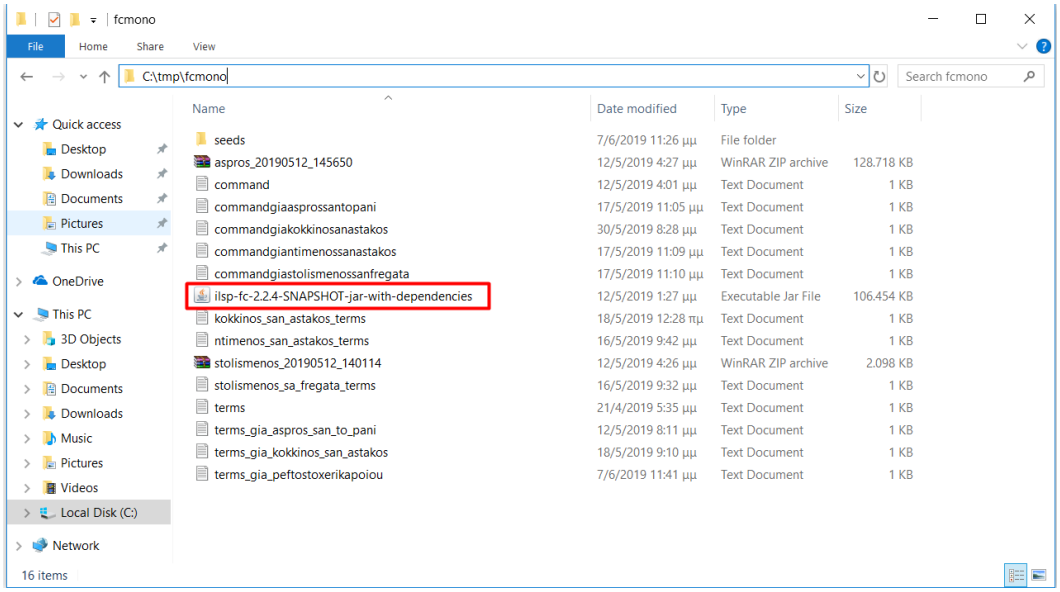
Για να τρέξουμε λοιπόν τον Crawler πρέπει πρώτα να ανοίξουμε τη γραμμή εντολών (cmd) των windows. Αυτό γίνεται είτε πηγαίνοντας στη έναρξη,

πληκτρολογώντας cmd και έπειτα πατώντας «enter» είτε με συνδυασμός πλήκτρων (win key +R) όπως εμφανίζει η εικόνα παρακάτω.



Εικόνα 7 Βήμα 1ο χρήσης Crawler

Έπειτα τρέχουμε τον crawler μέσω μιας εντολής java ως εξής : Πληκτρολογούμε τη διαδρομή που βρίσκεται το αρχείο του crawler με την εντολή : [(java –jar (κενό) (διαδρομή αρχείου)] .



Εικόνα 8 Βήμα 2ο χρήσης Crawler

Για την περίπτωσή μας πατάμε : java -jar "C:\tmp\fcmono\ilsp-fc-2.2.4-SNAPSHOT-jar-with-dependencies.jar"

Σε αυτό το σημείο θα ανοίξει το πρόγραμμα και θα εμφανίσει τη λίστα των εντολών – λειτουργιών για να μπορέσει ο χρήστης να δώσει μια πλήρη εντολή στον crawler ανάλογα με τις ανάγκες της έρευνας που θέλει να κάνει.

Εικόνα 9 Δείγμα λίστας εντολών

Παραθέτουμε ενδεικτικά τις εντολές όπως δίνονται από το πρόγραμμα.

```
-a,--agentname <arg>                 Agent name to identify the person or the organization
                                     responsible for the crawl
-align,--align_sentences <arg>       Sentence align document pairs using this aligner (default is
                                     maligna)
-bs,--basename <arg>                 Basename to be used in generating all files for easier
                                     content navigation
-c,--crawlduration <arg>             Maximum crawl duration in minutes
-cc,--creative_commons               Force the alignment process to generate a merged TMX with
                                     sentence alignments only from document pairs for which an
                                     open content license has been detected.
-cfg,--config <arg>                  Path to the XML configuration file
-crawl,--crawl                       Start a crawl
-d,--stay_in_webdomain               Force the monolingual crawler to stay in a specific web
                                     domain
-dbg,--debug                         Use debug level for logging
-dedup,--deduplicate                 Deduplicate and discard (near) duplicate documents
-del,--delete_redundant_files        Delete redundant crawled documents that have not been
                                     detected as members of a document pair
-dest,--destination <arg>            Path to a directory where the acquired/generated resources
                                     will be stored
-pdm,--pairDetectMethods <arg>       When creating a merged TMX file, only use sentence alignments
                                     from document pairs that have been identified by specific
                                     methods, e.g. auidh. See the pdm option.
-dom,--domain <arg>                  A descriptive title for the targeted domain
-export,--export                     Export crawled documents to cesDoc XML files
-f,--force                           Force a new crawl. Caution: This will remove any previously
                                     crawled data
-filter,--fetchfilter <arg>          Use this regex to force the crawler to crawl only in specific
                                     sub webdomains. Webpages with urls that do not match this
                                     regex will not be fetched.
-h,--help                            This message
-i,--inputdir <arg>                  Input directory for deduplication, pairdetection, or
                                     alignment
-ifp,--image_urls                    Full image URLs (and not only their basenames) will be used
                                     in pair detection with common images
-k,--keepboiler                      Keep and annotate boilerplate content in parsed text
-l,--loggingAppender <arg>           Logging appender (console, DRFA) to use
-lang,--languages <arg>              Two or three letter ISO code(s) of target language(s), e.g.
                                     el (for a monolingual crawl for Greek content) or eng;el (for
                                     a bilingual crawl)
-len,--length <arg>                  Minimum number of tokens per text block. Shorter text blocks
                                     will be annoteted as "ooi-length"
-mtlen,--minlength <arg>             Minimum number of tokens in crawled documents (after
                                     boilerplate detection). Shorter documents will be discarded.
-n,--numloops <arg>                  Maximum number of fetch/update loops
-oxslt,--offline_xslt                Apply an xsl transformation to generate html files during
                                     exporting.
-p_r,--path_replacements <arg>       Put the strings to be replaced, separated by ';'. This might
                                     be useful for crawling via the web service
-pairdetect,--pair_detection         Detect document pairs in crawled documents
-pdm,--pair_detection_methods <arg>  A string forcing the crawler to detect pairs using one or
                                     more specific methods: a (links between documents), u
                                     (patterns in urls), p (common images and similar digit
                                     sequences),i (common images), d (similar digit sequences), h, or m, or l
                                     (high/medium/low similarity of html structure)
-segtypes,--segtypes <arg>           When creating a merged TMX file, only use sentence alignments
                                     of specific types, ie. 1:1
-storefilter,--storefilter <arg>     Use this regex to force the crawler to store only webpages
                                     with urls that match this regex.
-t,--threads <arg>                   Maximum number of fetcher threads to use
-tc,--topic <arg>                    Path to a file with the topic definition
-tmxmerge,--tmxmerge                 Merge aligned segments from each document pair into one tmx
                                     file
-type,--type <arg>                   Crawl type: m (monolingual) or  p (parallel)
-u,--urls <arg>                      File with seed urls used to initialize the crawl
-u_r,--url_replacements <arg>        A string to be replaced, separated by ';'.
```

Εικόνα 10 Σύνολο Εντολών Crawler

Οι εντολές που θα χρησιμοποιήσουμε εμείς είναι οι εξής :

-crawl : Η βασική εντολή για «crawling»

-export : Ζητάμε από το πρόγραμμα να εξάγει τα αποτελέσματα \

-dedup : Ζητάμε από το πρόγραμμα να αφαιρέσει τα διπλότυπα και τα (σχεδόν) διπλότυπα.

-monomerge : Ζητάμε από το πρόγραμμα να κατασκευάσει μια μονόγλωσση συλλογή με τη συγχώνευση των ήδη εξαγόμενων αρχείων cesDoc

-a : Δίνουμε όνομα χρήστη για τον προσδιορισμό του ατόμου ή του οργανισμού που είναι υπεύθυνος για την ανίχνευση π.χ για τη δική μας διευκόλυνση δίνομαι την πολυλεκτική έκφραση που αναζητάμε μέσα σε εισαγωγικά.

-f : Ζητάμε κατά βούληση να κάνει ένα καινούργιο «crawl» και να διαγράψει τυχόν παλιά αρχεία.

-cdl : Ορίζουμε πόσους κύκλους αναζήτησης θέλουμε το πρόγραμμα να τρέξει π.χ (–cdl 200) για 200 κύκλους.

-type m ή p: Ορίζουμε στο πρόγραμμα αν θέλουμε η έρευνα να είναι σε μία γλώσσα ή σε παράλληλες γλώσσες.

-u : Δίνουμε τη διαδρομή του αρχείου με urls σπόρους που χρησιμοποιούνται για την έναρξη της ανίχνευσης π.χ "C:\tmp\fcmono\seeds\aspros_san_to_pani.1.txt"

-lang "el" Ορίζουμε τη γλώσσα. Για ελληνικά (el), για αγγλικά (eng) ή για 2 γλώσσες (eng;el)

-k -dest Ορίζουμε τη διαδρομή σε έναν φάκελο στον οποίο θα αποθηκεύονται οι πόροι που έχουν αποκτηθεί / δημιουργηθεί π.χ "C:/tmp/fcmonotests"

-bs Ορίζουμε τη Βασική ονομασία που χρησιμοποιείται για τη δημιουργία όλων των αρχείων για ευκολότερη πλοήγηση περιεχομένου π. χ"C:/tmp/fcmonotests/output" όπου το output θα περιλαμβάνεται σε κάθε αρχείο και θα ακολουθεί το πλήρες όνομα.

-tc Ορίζουμε τη διαδρομή σε αρχείο με τον ορισμό του θέματος "C:/tmp/fcmono/terms_gia_aspros_san_to_pani.txt"

-dom : Δίνουμε έναν περιγραφικό τίτλο για τον στοχευόμενο τομέα

Παράδειγμα μια πλήρους εντολής :

java -jar "C:\tmp\fcmono\ilsp-fc-2.2.4-SNAPSHOT-jar-with-dependencies.jar" -crawl -export -dedup -monomerge -a "asprossantopani" -f -cdl 200 -type m -u "C:\tmp\fcmono\seeds\aspros_san_to_pani.1.txt" -lang "el" -k -dest "C:/tmp/fcmonotests" -bs "C:/tmp/fcmonotests/output" -tc "C:/tmp/fcmono/terms_gia_aspros_san_to_pani.txt" -dom "paroimies"

Εικόνα 11 Παράδειγμα από εντολή που τρέχει ο crawler.

3.1.3 Προσαρμογές Crawler

Με τη βοήθεια του κυρίου Βασίλη Παπαβασιλείου προσαρμόσαμε τον crawler στις ανάγκες της έρευνας. Το πρώτο πρόβλημα που αντιμετωπίσαμε ήταν ότι οι εκδόσεις του crawler που ήταν αναρτημένες και ελεύθερες προς τη χρήση ήταν ξεπερασμένες χρονικά οπότε παρουσίαζαν προβλήματα συμβατότητας με τους τρόπους λειτουργίας τους στο διαδίκτυο. Για να ξεπεράσουμε αυτό το πρόβλημα χρειάστηκε αρκετό χρονικό διάστημα και μετά από πολλές προσπάθειες δοκιμών και σφαλμάτων δημιουργήσαμε μια καινούργια και ενημερωμένη έκδοση του crawler.

Στη συνέχεια, ένα δεύτερο πρόβλημα που αναδύθηκε αμέσως μετά τη δημιουργία της καινούργιας έκδοσης, ήταν η συμβατότητα με το περιβάλλον των windows. Ο Crawler λοιπόν αρχικά ήταν προγραμματισμένος να λειτουργεί σε περιβάλλον linux[7]. Για να παρακάμψουμε αυτό το εμπόδιο χρειάστηκε να αξιοποιήσουμε αρχικά το πρόγραμμα Cygwin[8]

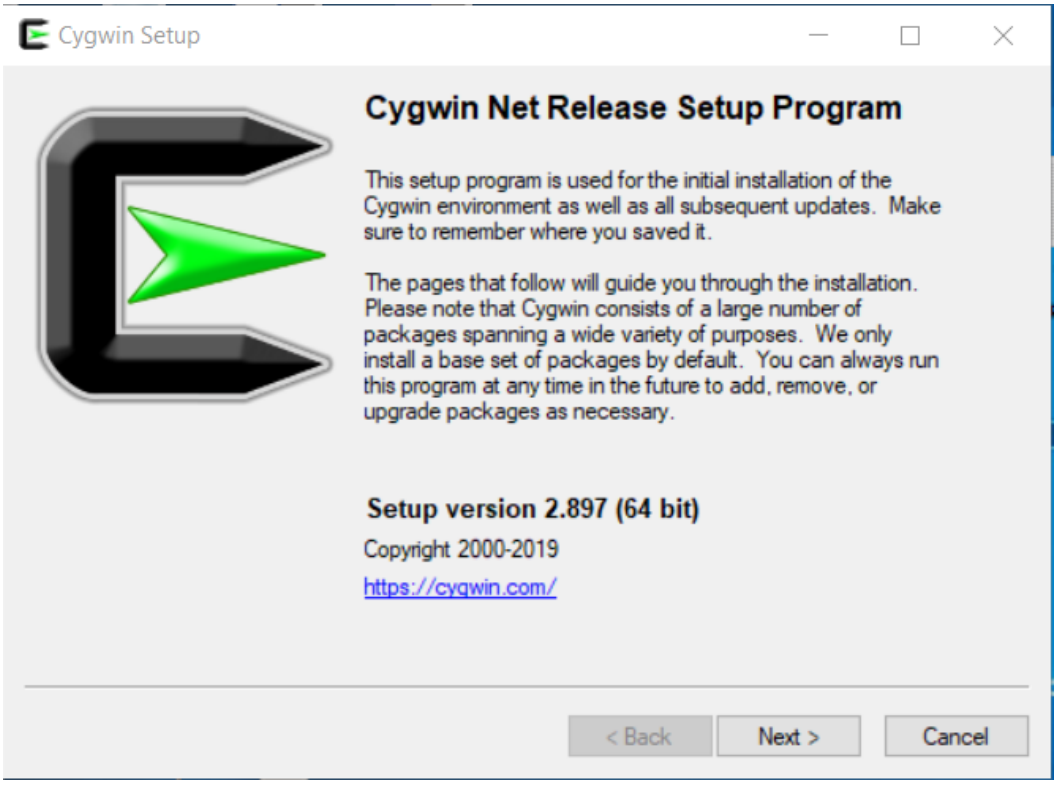
Αφού ξεπεράσαμε τις αρχικές δυσκολίες όσον αφορά την εγκατάσταση και την παραμετροποίηση του Cygwin στον υπολογιστή κατά τη λειτουργία του crawler ανέκυψαν διάφορα σφάλματα που αφορούσαν το τεχνικό κομμάτι του προγράμματος και τη συλλογή δεδομένων. Και αυτά τα επιλύσαμε σε συνεργασία με τον κύριο Παπαβασιλείου ο οποίος ανέλαβε να επαναπρογραμματίσει τον crawler σύμφωνα με την δική μας ανατροφοδότηση με βάση τα προβλήματα που ανέκυψαν.

Εγκατάσταση και χρήση Cygwin

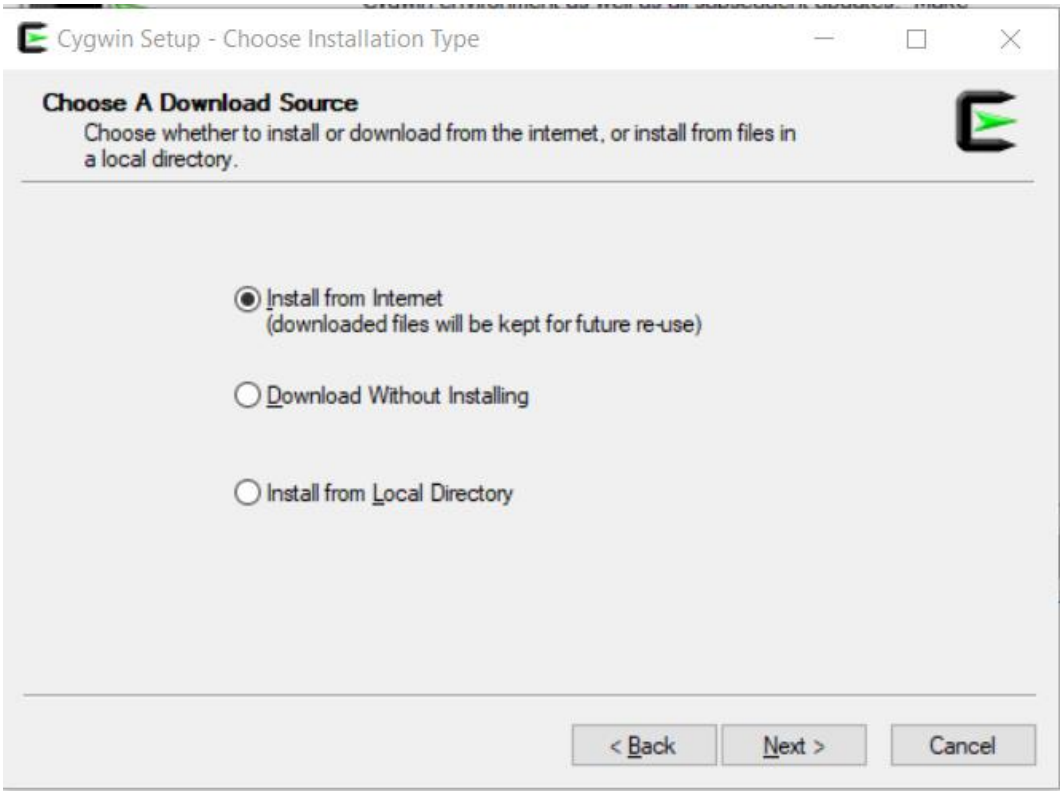
Αρχικά πρέπει να επισκεφτούμε το εξής url: https://cygwin.com/install.html. Στη συνέχεια πατάμε «setup-x86_64.exe» και πατάμε run.

[7] Το Linux είναι ένα λειτουργικό σύστημα που αποτελείται από ελεύθερο λογισμικό σε αντιδιαστολή με τα Windows.

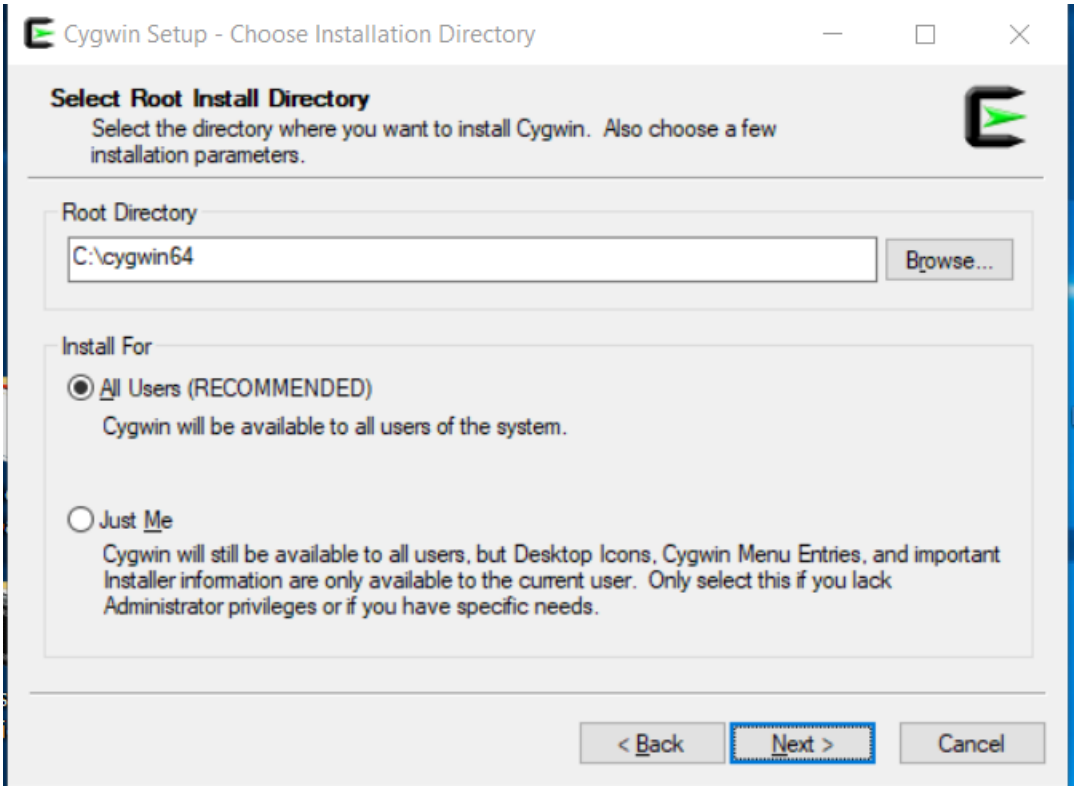
[8] Το Cygwin αποτελεί μια μεγάλη συλλογή εργαλείων δωρεάν που παρέχουν λειτουργικότητα παρόμοια με τα Linux στα Windows. (Cygwin, 2019)



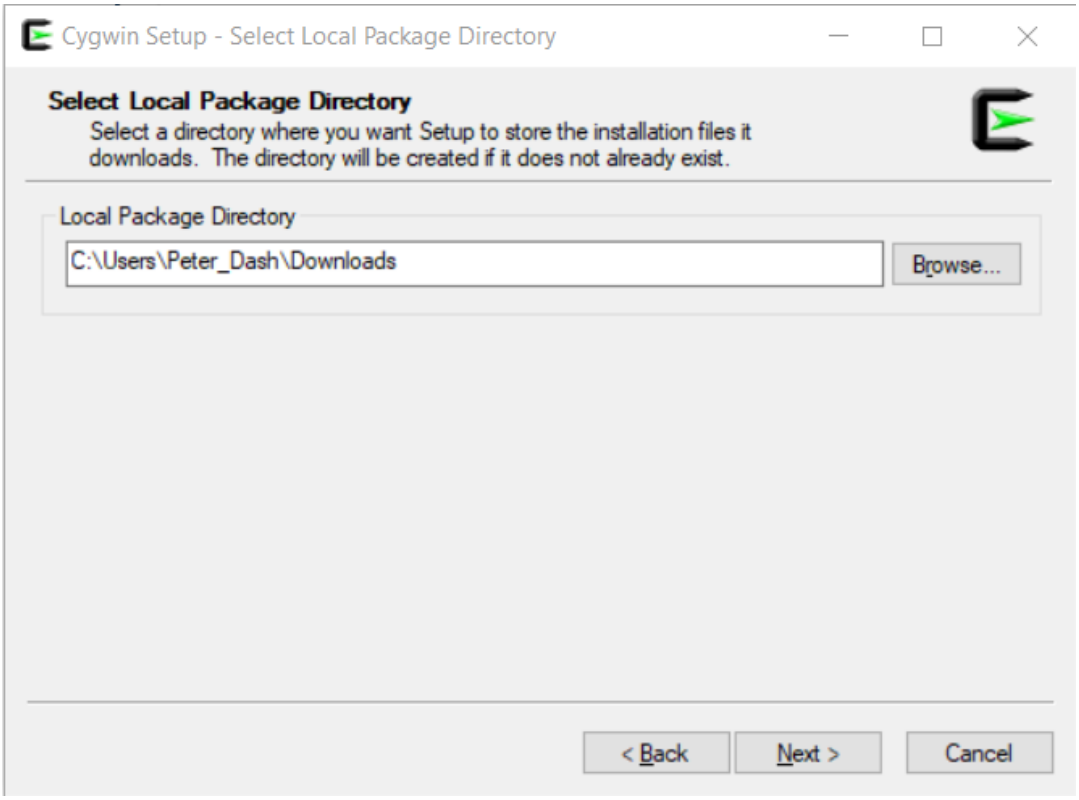
Εικόνα 12 Cygwin Πρώτο Βήμα



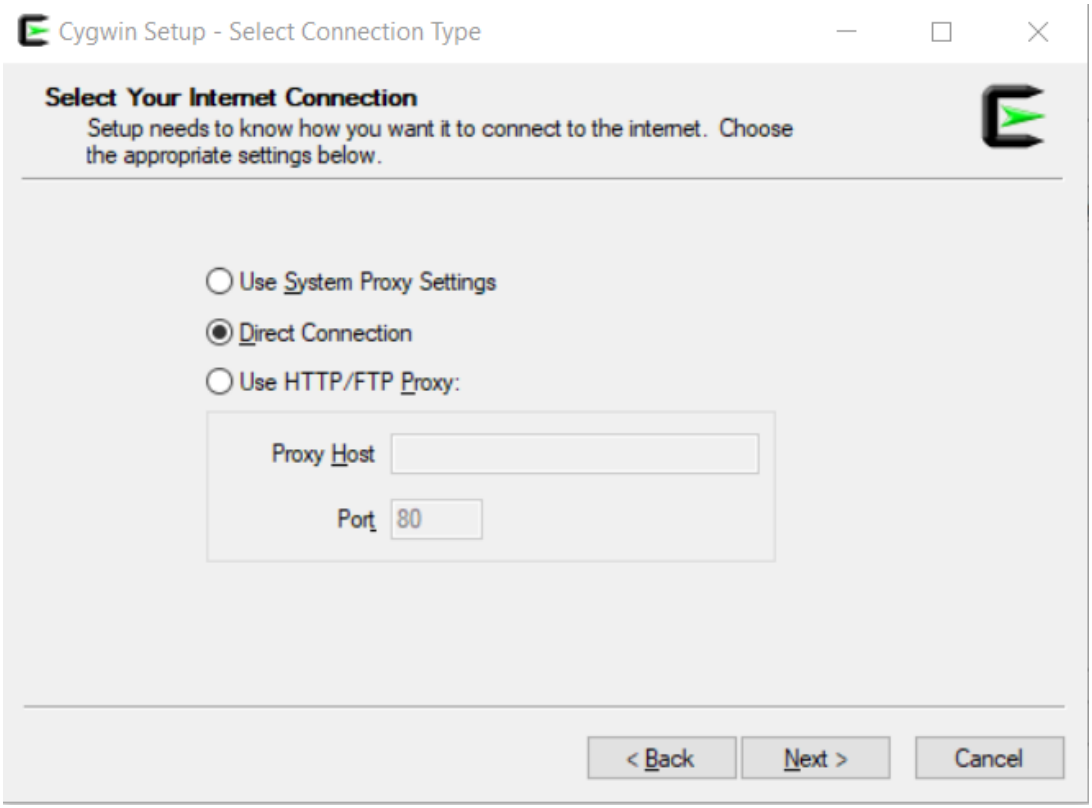
Εικόνα 13 Cygwin Δεύτερο Βήμα



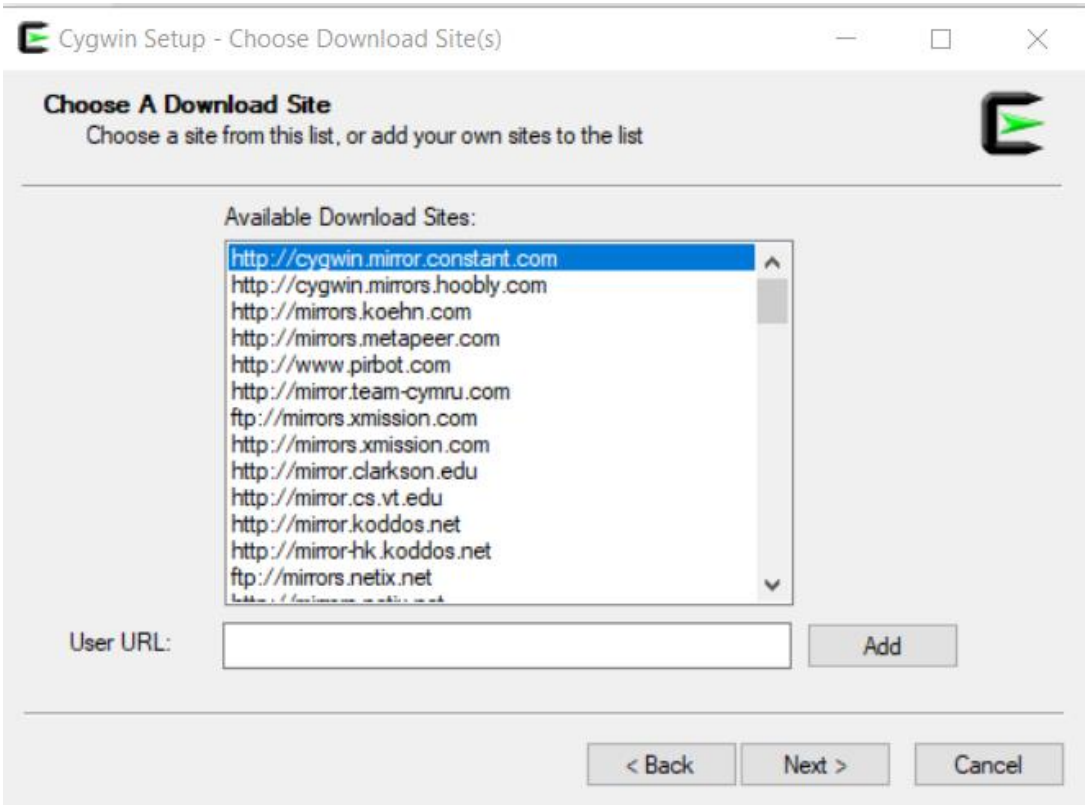
Εικόνα 14 Βήμα Τρίτο



Εικόνα 15 Βήμα Τέταρτο

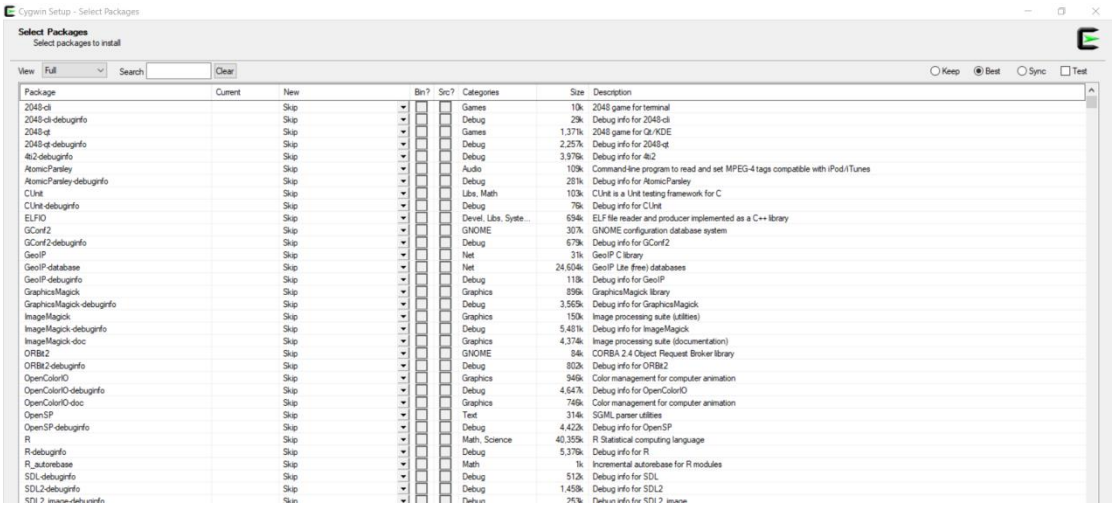


Εικόνα 16 Βήμα Πέμπτο



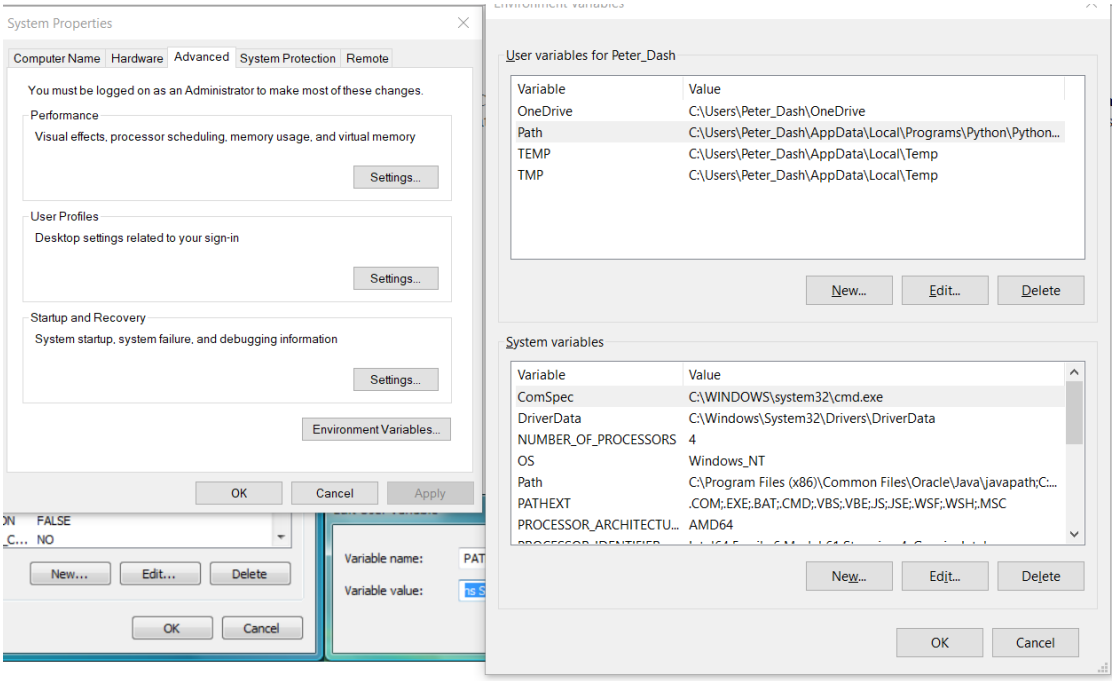
Εικόνα 17 Βήμα Έκτο

Σε αυτή τη σελίδα πατάμε full και μετά next.



Εικόνα 18 Βήμα Έβδομο

Έπειτα πατάμε next σε ότι εμφανιστεί. Τέλος πρέπει να ρυθμίσουμε το Cygwin στα Windows για να αναγνωρίζουν την διαδρομή του προγράμματος. Πάμε στον πίνακα ελέγχου του υπολογιστή, μετά πατάμε σύστημα και ασφάλεια και μετά ασφάλεια. Έπειτα πατάμε αλλαγή ρυθμίσεων, πατάμε για προχωρημένους και τέλος πατάμε μεταβλητές περιβάλλοντος.



Εικόνα 19 Βήμα Όγδοο

Τέλος πατάμε επεξεργασία και βάζουμε τη διαδρομή που εγκαταστήσαμε το πρόγραμμα π..χ C:\cygwin\bin και είμαστε έτοιμοι.

3.2 Bootcat

Το Bootcat αυτοματοποιεί τη διαδικασία εύρεσης κειμένων αναφοράς στο διαδίκτυο και τη σύγκρισή τους σε ένα ενιαίο σώμα. Επιτρέπει ποικίλα επίπεδα ελέγχου. Στο πρώτο στάδιο, οι χρήστες παρέχουν έναν κατάλογο όρων μιας ή πολλαπλών λέξεων που χρησιμοποιούνται ως σπόροι (seeds) για συλλογή κειμένου. Αυτά στη συνέχεια συνδυάζονται σε "πλειάδες" ποικίλου μήκους και αποστέλλονται ως ερωτήματα σε μια μηχανή αναζήτησης, η οποία επιστρέφει μια λίστα δυνητικά σχετικών διευθύνσεων URL. Σε αυτό το σημείο ο χρήστης έχει τη δυνατότητα να επιθεωρεί τις διευθύνσεις URL και να αφαιρεί όσες δεν του ταιριάζουν σύμφωνα με την έρευνά του. Οι πραγματικές ιστοσελίδες ανακτώνται στη συνέχεια, μετατρέπονται σε απλό κείμενο και αποθηκεύονται σε μορφή "txt". Το σώμα μπορεί έτσι να διερευνηθεί χρησιμοποιώντας τα συμφραζόμενα.

Χρησιμοποιώντας το BootCat μπορεί κανείς να κατασκευάσει ένα σχετικά μεγάλο και γρήγορο σώμα σε πολύ λίγο χρόνο αλλά χωρίς επιλογή των κειμένων. Αυτή η σχετικά ευέλικτη προσέγγιση καθιστά το BootCaT ένα χρήσιμο εργαλείο για μεταφραστές το οποίο έχει χρησιμοποιηθεί στον τομέα της μετάφρασης και της ορολογίας για την κατασκευή μικρών κορμών διαφορετικού μεγέθους και εξειδίκευσης.

3.2.1 Τρόπος Λειτουργίας Bootcat

Η διαδικασία που ακολουθεί το BootCaT μπορεί να χωριστεί σε δύο κύριες φάσεις: Χρησιμοποιείται πρώτα ένας επαναλαμβανόμενος αλγόριθμος για να δεθούν σώματα κειμένων και όροι unigram[9] από τον ιστό. Στη συνέχεια, προχωράμε για να εξάγουμε όρους πολλαπλών λέξεων με βάση τον τελικό κορμό και τη λίστα όρων unigram που εξάγουμε στην προηγούμενη φάση.

[9] [9] Τα N-grams αντιπροσωπεύουν ακολουθίες από tokens. Tokens : συνολικός αριθμός λέξεων Διαφορετικά είδη tokens : N-grams χαρακτήρων, N-grams λέξεων, N-grams Part of Speech N = πόσοι όροι εξετάζονται ως εξής (Unigrams: 1 όρος Bigrams: 2 όροι Trigrams: 3 όροι).

Η διαδικασία δεσίματος σωμάτων κειμένων και Unigrams (bootstrapping) ξεκινά με μια μικρή λίστα σπόρων που αναμένεται να είναι αντιπροσωπευτικοί του τομέα υπό διερεύνηση. Για καλά καθορισμένους εξειδικευμένους τομείς, ένας μικρός κατάλογος σπόρων (5 έως 15) είναι συνήθως επαρκής. Οι όροι των σπόρων συνδυάζονται τυχαία και κάθε συνδυασμός χρησιμοποιείται ως ερώτημα στη μηχανή αναζήτησης Google. Οι κορυφαίες σελίδες που επιστρέφονται για κάθε ερώτημα ανακτώνται και μορφοποιούνται ως κείμενο.

Νέοι σπόροι unigram εξάγονται από το σώμα των ανακτημένων σελίδων, συγκρίνοντας τη συχνότητα εμφάνισης της κάθε λέξη σε αυτό το σετ με τη συχνότητα εμφάνισής της στο σώμα κειμένων αναφοράς. Συγκρίνουμε τις συχνότητες χρησιμοποιώντας το αρχείο καταγραφής. Τυχαίοι συνδυασμοί των πρόσφατα εξαχθέντων σπόρων χρησιμοποιούνται στη συνέχεια για έναν άλλο γύρο ερωτημάτων στη Google και ένα νέο σώμα δημιουργείται με την ανάκτηση και τη μορφοποίηση των κορυφαίων σελίδων που βρέθηκαν σε αυτόν τον γύρο. Αυτή η διαδικασία εξαγωγής όρων και λήψης σώματος κειμένων μπορεί να επαναληφθεί όσες φορές χρειαστεί ανάλογα με τις προτιμήσεις του χρήστη. (Bernardini, 2004)

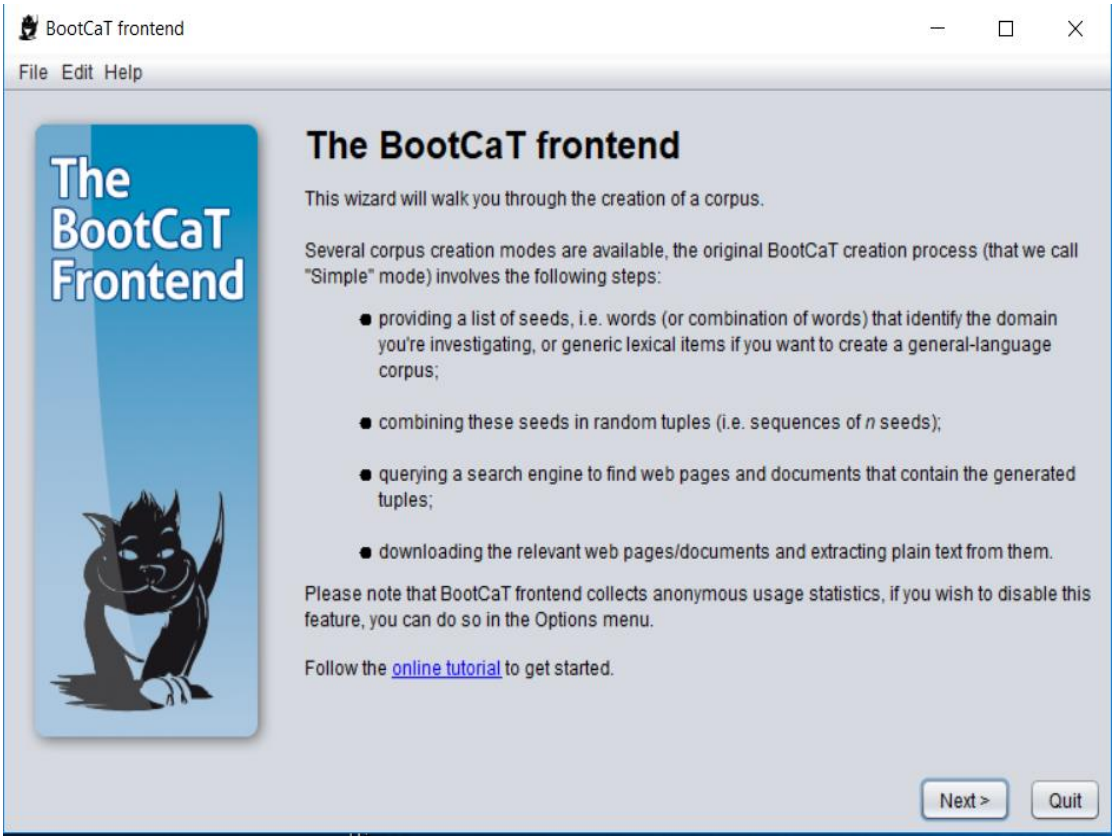
Η εξαγωγή των πολυλεκτικών όρων γίνεται ως εξής: Αρχικά το πρόγραμμα εξάγει έναν κατάλογο συνδέσεων ενός και δύο λέξεων από το σώμα, αναζητώντας λέξεις και bigrams που εμφανίζονται συχνά μεταξύ δύο μονολεκτικών όρων π.χ. (από).

Στη συνέχεια εξάγει μια λίστα με λέξεις όπως «*απ, από, γι, για, δι, δια, εις, εκ, ένα, έναν, ένας, ενός, εξ, επ, επί, καθ, και, κατ, κατά, με, μέσα, μια, μία, μιά, μιαν, μιας, περί, σε, στα, στη, στην, στις, στο, στον, στους, τα, τη, την, της, τις, το, τον, του, τους, των, υπό*» με πολύ υψηλή συχνότητα σε έγγραφα που δεν αναγνωρίστηκαν ως συνδετικές. Σε αυτό το σημείο, μπορούμε να αναζητήσουμε όρους πολλών λέξεων ή αλλιώς ακολουθίες λέξεων που πληρούν κάποιους περιορισμούς.

Οι πολυλεκτικοί όροι αναζητούνται επαναλαμβανόμενα. Ξεκινώντας με διγράμματα, κοιτάμε αριστερά και δεξιά για έναν όρο n + 1 που περιέχει το ρέον (ν)gram και ικανοποιεί τους περιορισμούς που έχουν δοθεί. Για τον κάθε σπόρο bigram επιστρέφεται ο μακρύτερος και πιο καλά διαμορφωμένος όρος που τον περιέχει.

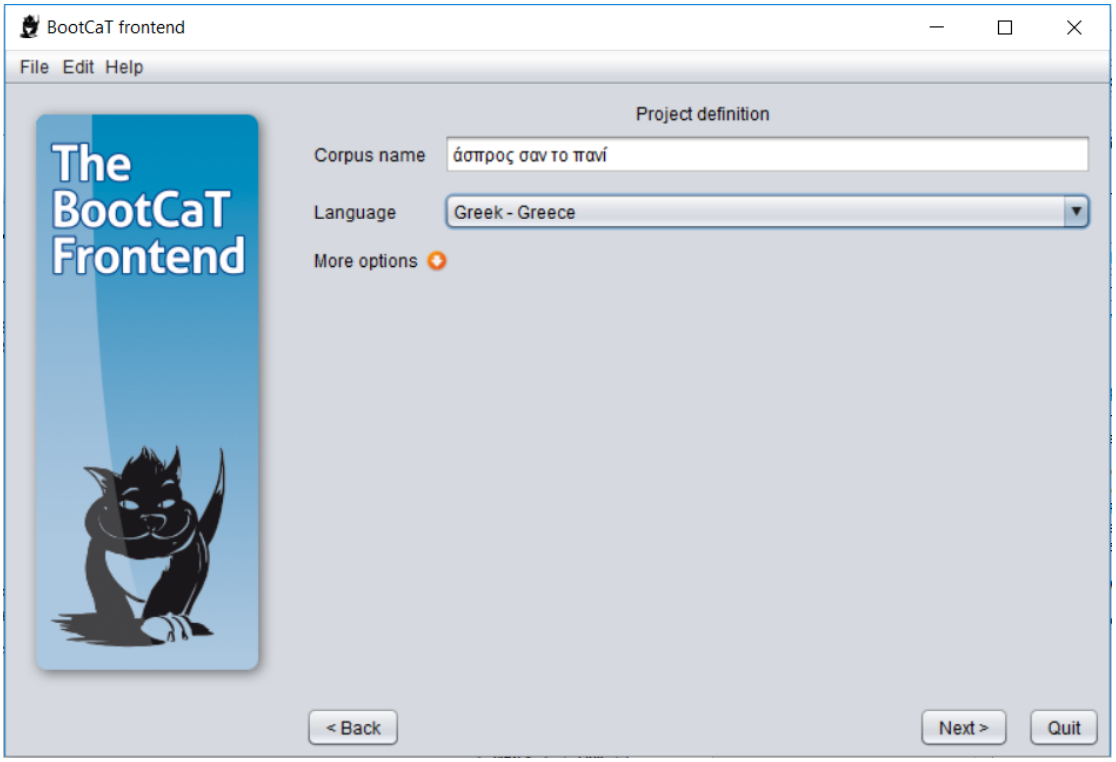
Και πάλι, ο χρήστης πρέπει να ορίσει διάφορες παραμέτρους, όπως την ελάχιστη συχνότητα για τους όρους bigram και την αξία της σταθεράς «k» (το ελάχιστο όριο

συχνότητας για τους μακρύτερους όρους θα προκύψει από αυτές τις δύο παραμέτρους).



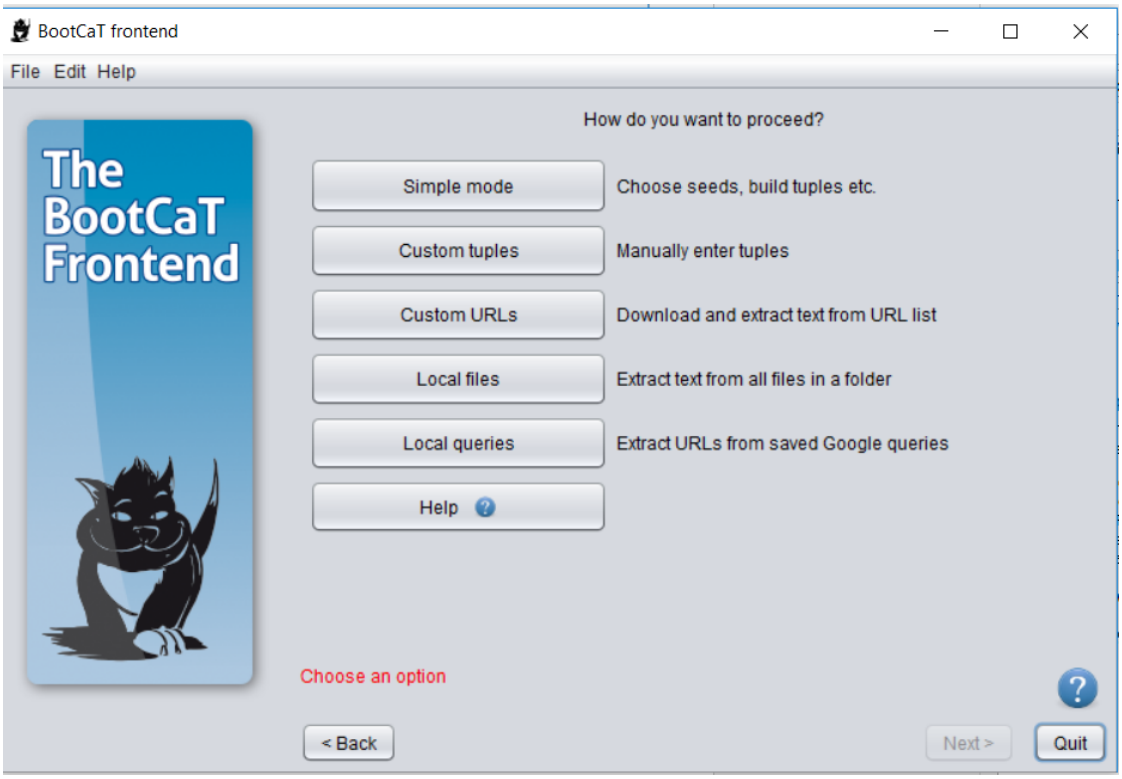
Εικόνα 20 Bootcat Βήμα 1ο

Μια πρώτη εικόνα του προγράμματος είναι αυτή. Στο σημείο αυτό αφού έχουμε διαβάσει τις οδηγίες πατάμε “next”.



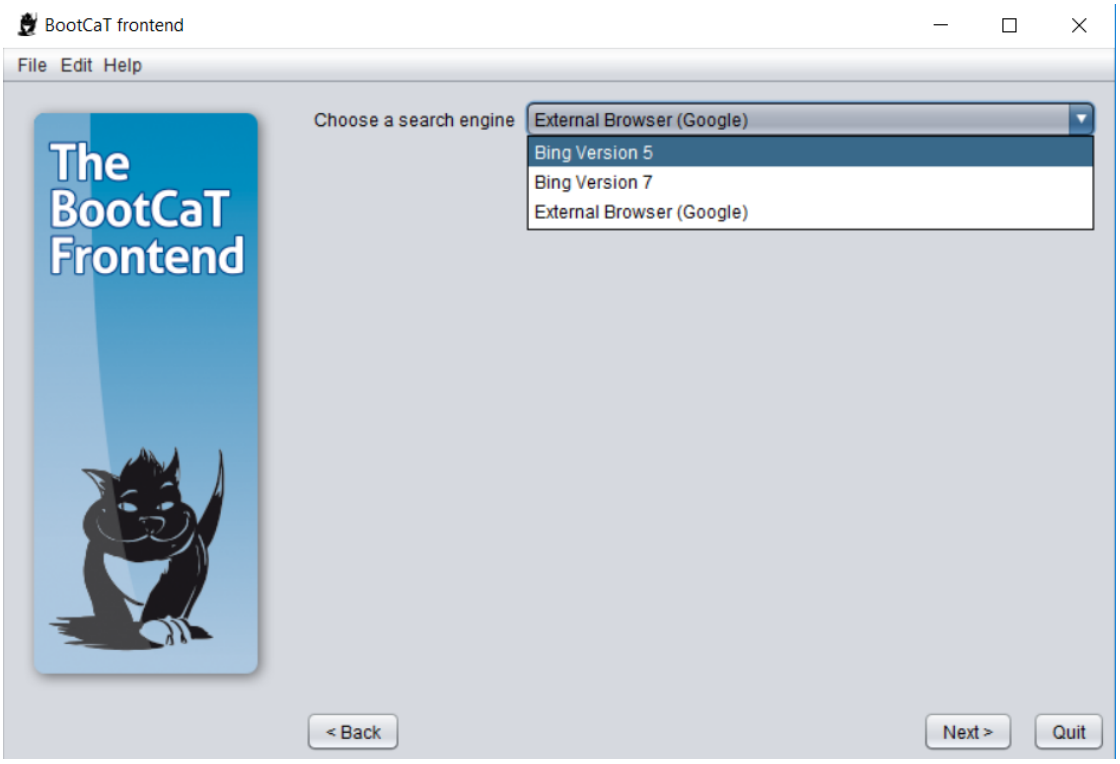
Εικόνα 21 Bootcat Βήμα 2ο

Σε αυτή τη σελίδα του προγράμματος επιλέγουμε στο πεδίο “corpus name” ένα όνομα για το σώμα κειμένων που θέλουμε να φτιάξουμε και στο πεδίο ”language” τη γλώσσα που θέλουμε. Έπειτα πατάμε “next”



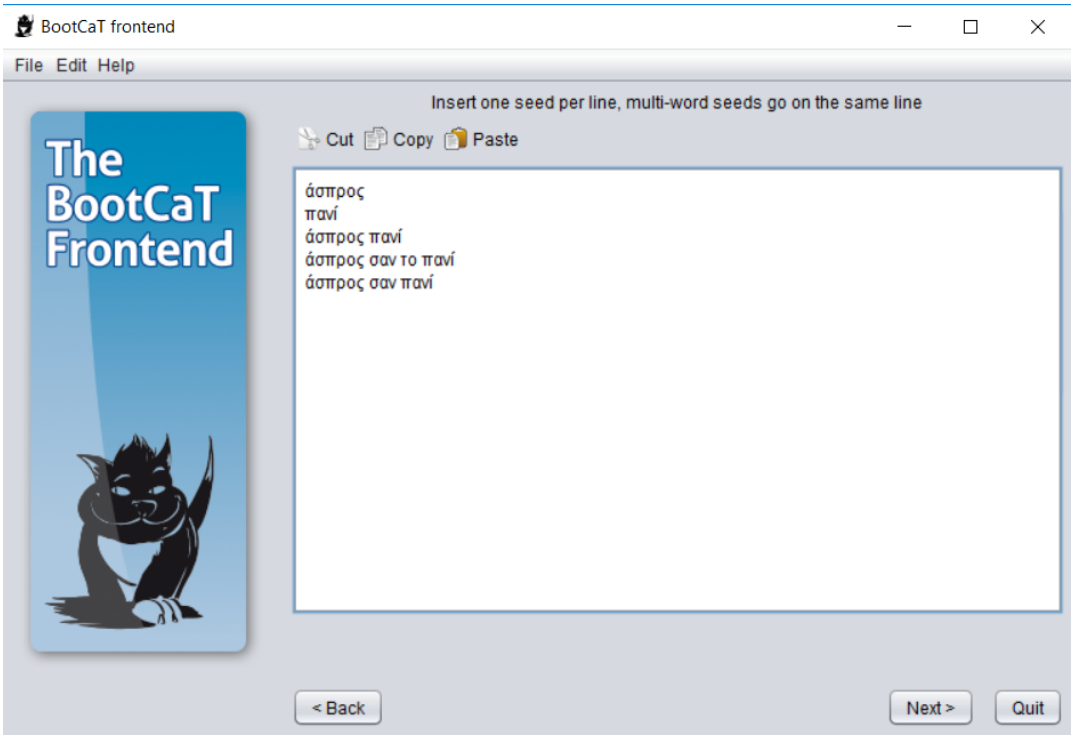
Εικόνα 22 Bootcat Βήμα 3ο

Έπειτα επιλέγουμε «simple mode».



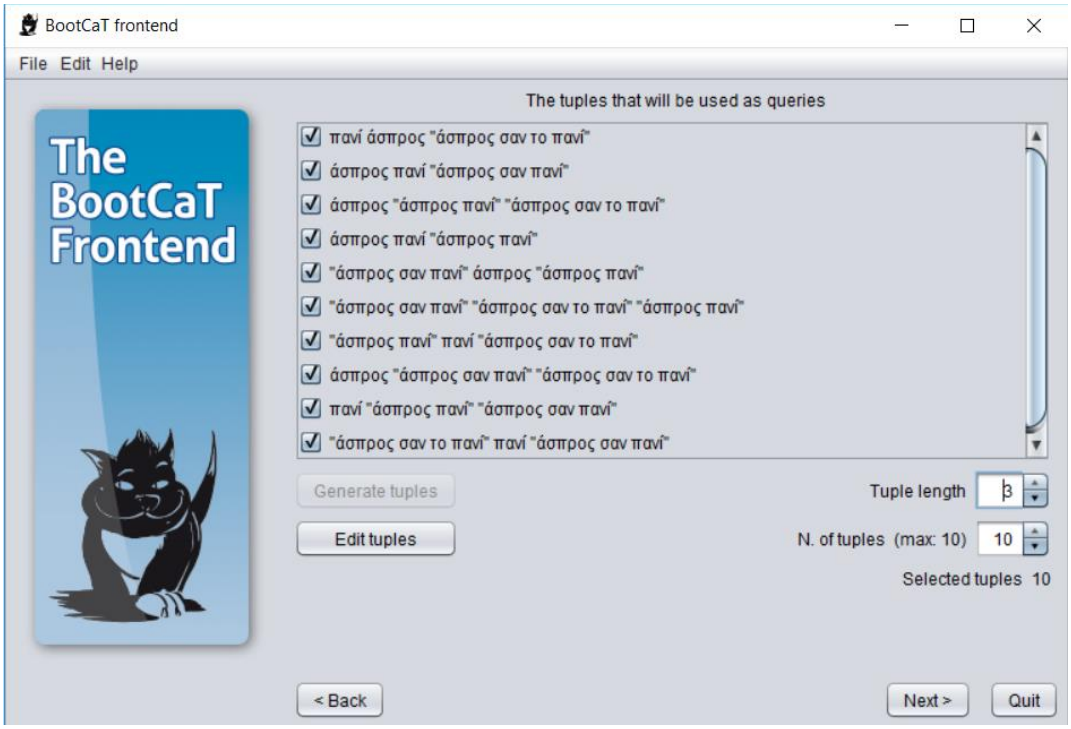
Εικόνα 23 Bootcat Βήμα 4ο

Επιλέγουμε «external browser google» και μετά «next».



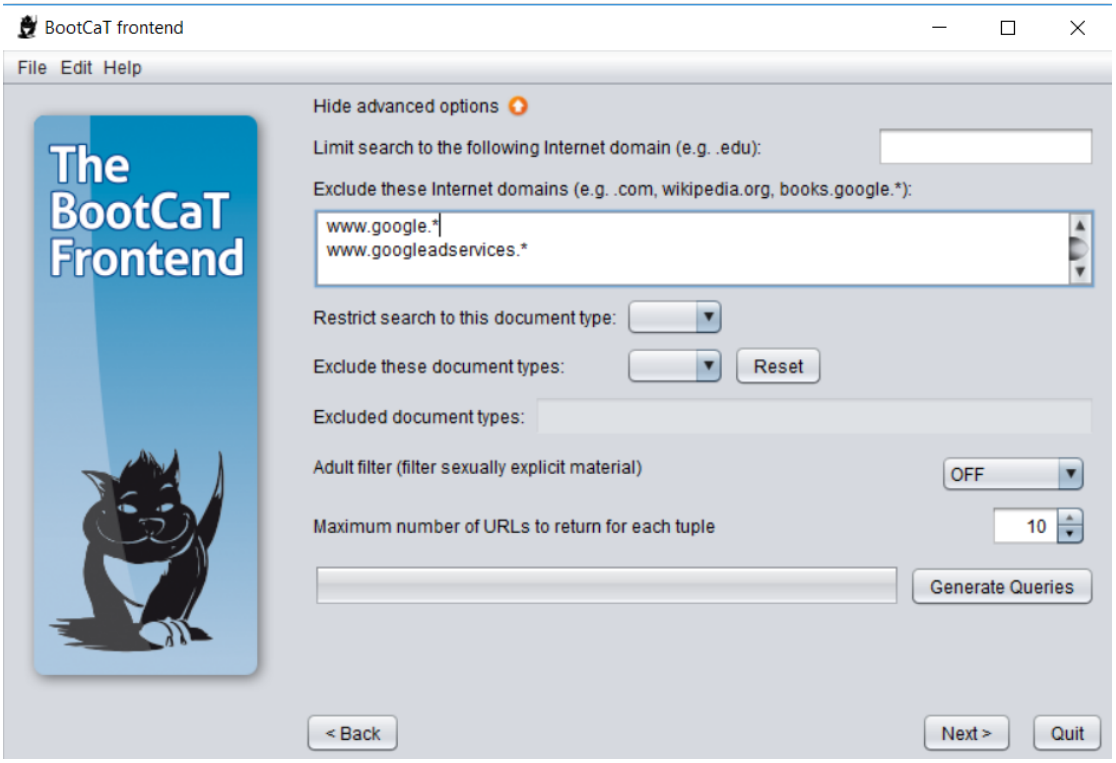
Εικόνα 24 Bootcat Βήμα 5ο

Έπειτα πληκτρολογούμε τις λέξεις σπόρους ή τις πολυλεκτικές εκφράσεις σπόρους και πατάμε «next».



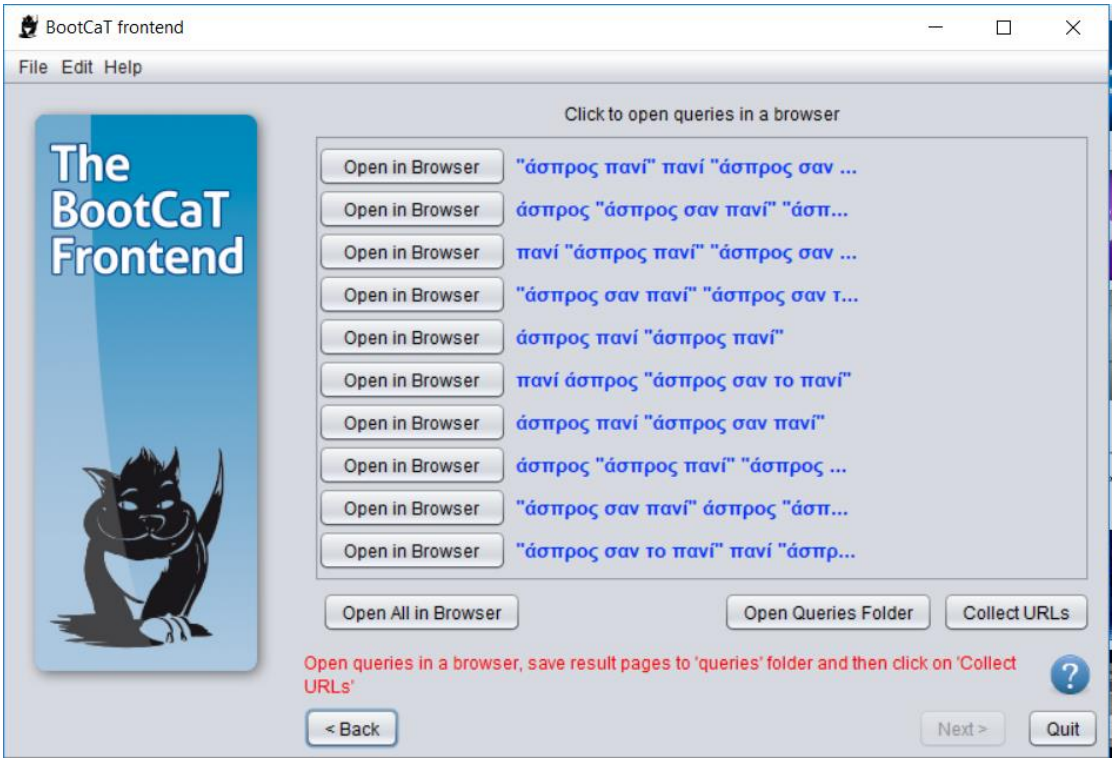
Εικόνα 25 Bootcat Βήμα 6ο

Έπειτα αφού πατήσουμε το «generate tuples» μας εμφανίζει τις εκφράσεις που θα αναζητήσει στο web και όπως και παραπάνω πατάμε «next».



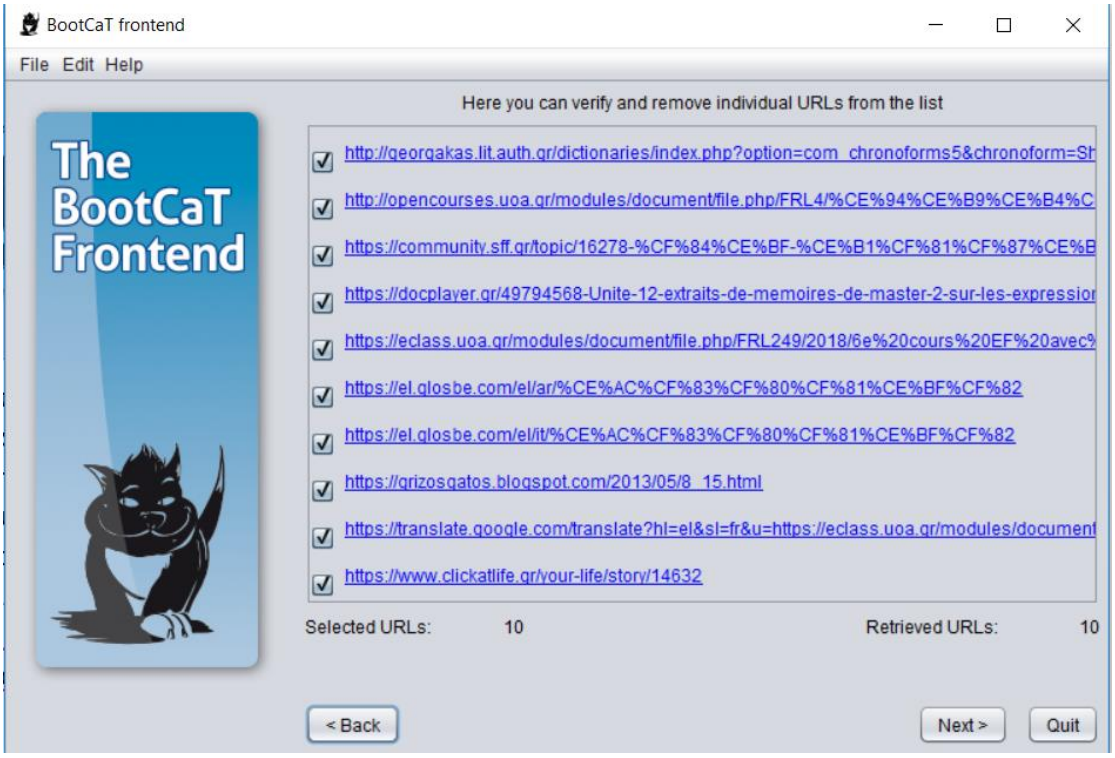
Εικόνα 26 Bootcat Βήμα 7ο

Αφήνουμε την σελίδα αυτή του προγράμματος ως έχει και πατάμε «generate queries»



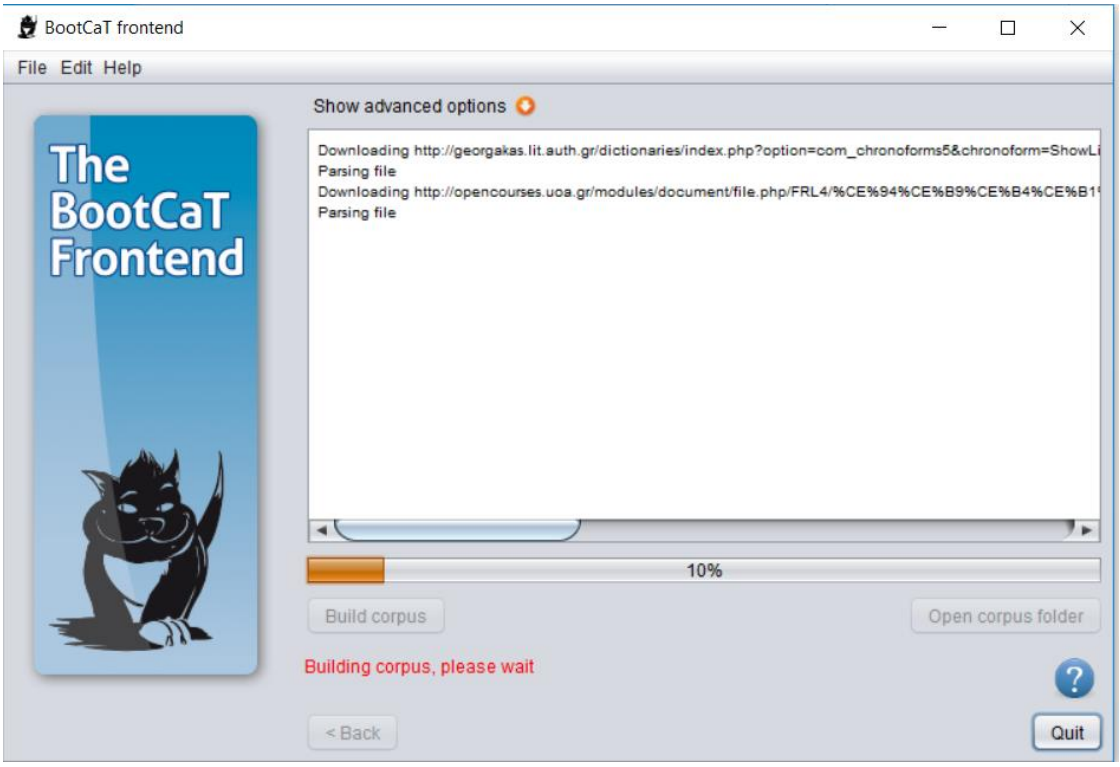
Εικόνα 27. Bootcat Βήμα 8ο

Εδώ πατάμε «open all in browser» για να μας ανοίξει τις σελίδες που έχει βρει σε έναν browser και στη συνέχεια τις αποθηκεύουμε χειροκίνητα στο «queries folder». Σε αυτό το σημείο για να δούμε ποια είναι η διαδρομή του queries folder πατάμε το «open queries folder». Πατώντας δεξί κλίκ στην αναζητήσεις που έχει κάνει στο google αποθηκεύουμε τις ιστοσελίδες ως πλήρεις στην εξής διαδρομή: «C:\Users\Peter_Dash\Documents\BootCaT Corpora\άσπρος σαν το πανί\queries» και πατάμε «collect urls».



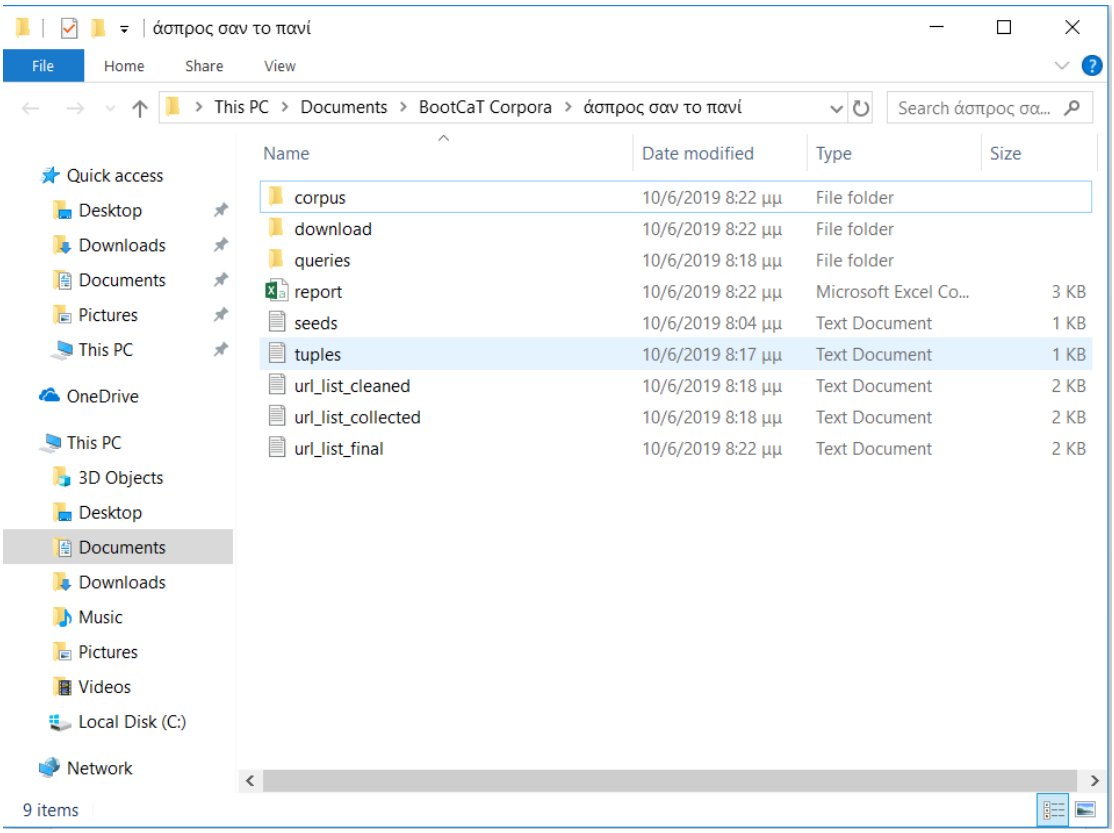
Εικόνα 28. Bootcat Βήμα 9ο

Εδώ αν θέλουμε μπορούμε να αφαιρέσουμε μεμονωμένα urls αν κρίνουμε ότι δεν είναι συναφή με την έρευνά μας και όταν έχουμε ολοκληρώσει την επιλογή μας πατάμε «next».



Εικόνα 29. Bootcat Βήμα 10ο

Πατάμε «build corpus» και περιμένουμε το πρόγραμμα να ολοκληρώσει τη συλλογή πληροφοριών. Τέλος πατάμε «open corpus folder».



Εικόνα 30. Bootcat Βήμα 11ο

Εδώ εμφανίζεται ο φάκελος με τα αποτελέσματα. Περιλαμβάνει τα αρχικά στοιχεία που δίνει ο χρήστης στο πρόγραμμα όπως και τα urls που κράτησε το πρόγραμμα και το σώμα κειμένων που έχει δημιουργήσει σε μορφή αρχείων «txt» και όχι ενιαίου αρχείου . Επίσης περιλαμβάνει τις σελίδες τις οποίες κατέβασε, αποθηκευμένες για πιθανό έλεγχο από το χρήστη.

3.3 Sketch Engine

Το Sketch Engine είναι ένα ακόμα εργαλείο για γλωσσική έρευνα. Λειτουργεί με μεγάλα δείγματα γλώσσας, που ονομάζονται σώματα κειμένων. Οι αλγόριθμοί του αναλύουν τα σώματα κειμένων για να εντοπίσουν γρήγορα τι είναι τυπικό στη γλώσσα και τι είναι σπάνιο, ασυνήθιστο, ξεπερασμένο, τι βγαίνει εκτός χρήσης ή ποιες νέες λέξεις αρχίζουν να χρησιμοποιούνται. Επίσης, έχει σχεδιαστεί για εφαρμογές ανάλυσης ή εξόρυξης κειμένου.

Το Sketch Engine χρησιμοποιείται κυρίως από γλωσσολόγους, λεξικογράφους, μεταφραστές, φοιτητές και δασκάλους. Πρόκειται για μια λύση πρώτης επιλογής για εκδότες, πανεπιστήμια, μεταφραστικούς οργανισμούς και ινστιτούτα εθνικών γλωσσών σε ολόκληρο τον κόσμο. Το Sketch Engine περιέχει 500 έτοιμα προς χρήση σώματα κειμένων σε πάνω από 90 γλώσσες, καθένα από τα οποία έχει μέγεθος μέχρι και 30 δις λέξεις για να παρέχει ένα πραγματικά αντιπροσωπευτικό δείγμα γλώσσας. (Sketch Engine, 2004)

3.3.1 Τρόπος Λειτουργίας Sketch Engine

Η λεξικογραμματική σύνοψη είναι αυτό που δίνει το όνομά του στο πρόγραμμα το οποίο αντιπροσωπεύει μία σύνοψη των γραμματικών και συμφραστικών συμπεριφορών μιας λέξης. Χρησιμοποιήθηκαν για πρώτη φορά στην παραγωγή του Macmillan English Dictionary και παρουσιάστηκαν στο Euralex 2002 (Kilgarriff and Rundell 2002). (Kilgarriff A. a., 2004, p. 3)

Η κύρια λοιπόν λειτουργία του προγράμματος είναι να δημιουργεί, παίρνοντας ως βάση ένα σώμα κειμένων οποιασδήποτε γλώσσας, σκίτσα λέξεων για τις λέξεις της προκαθορισμένης γλώσσας. Το σκίτσο της λέξης χρησιμοποιώντας γραμματικά μοτίβα ασχολείται με ένα αυθαίρετο κομμάτι κειμένου γύρω από την λέξη που το δίνουμε και επισημαίνει γραμματικές σχέσεις στις οποίες συμμετέχει η λέξη. Παρέχει δηλαδή έναν κατάλογο των συμφραζόμενων για κάθε γραμματική σχέση στην οποία συμμετέχει η λέξη. Για παράδειγμα για ένα ρήμα θα μας εμφανίσει το υποκείμενο, τα αντικείμενα, τα συνοδευτικά ρήματα, τα επιρρήματα, προθέσεις και θα τα παρουσιάσει σε διαφορετικές λίστες.

Ακόμα περιλαμβάνει έναν θησαυρό με βάση τα σώματα κειμένων και «διαφορές σκίτσων», οι οποίες προσδιορίζουν για δύο σημασιολογικά συναφείς λέξεις ποια συμπεριφορά τους είναι κοινή και σε τι διαφέρουν. (Kilgarriff A. a., 2004, σ. 8)

Το Sketch Engine είναι ένα σύστημα ερωτημάτων σε ένα σώμα κειμένων που επιτρέπει στο χρήστη να βλέπει σκίτσα λέξεων, όπως είναι οι λέξεις που έχουν την ίδια έννοια, και «sketch diff».

Τα σκίτσα λέξεων είναι πλήρως ενσωματωμένα με τα συμφραζόμενα (concordance) : κάνοντας κλικ σε μια θέση ενδιαφέροντος για τη λεξικογραμματική σύνοψη, ο χρήστης παίρνει μια αντιστοιχία των αποδεικτικών στοιχείων του σώματος που οδηγούν σε αυτή την αλληλεξάρτηση σε αυτή τη γραμματική σχέση. Αν ο χρήστης κάνει κλικ στη λέξη π.χ «toast» στη λίστα των αντικειμένων υψηλής ανάλυσης στο σκίτσο για το ρήμα «spread» θα ληφθούν υπόψη οι συνάψεις όπου το «toast» (n) εμφανίζεται ως αντικείμενο του «spread» (v).

***pray* (v)** BNC freq= 2455

~ for	680	3.4	~ to	142	1.1	and/or	179	1.7	modifier	338	0.5	object	183	-1.2	subject	1361	0.5
rain	12	19.8	god	32	24.0	hope	20	20.8	silently	15	13.3	god	13	10.5	we	306	12.3
soul	14	19.3	God	22	17.7	hop	13	15.5	together	35	9.3	God	11	9.6	petitioner	7	8.3
-	117	17.3	lord	16	11.4	fast	6	12.2	fervently	4	7.6	prayer	6	7.6	knee	5	6.9
God	11	16.5	saint	4	10.0	pray	16	11.2	aloud	6	7.5	day	9	3.8	congregation	4	6.8
peace	25	16.5	jesus	2	5.4	kneel	5	9.9	earnestly	5	7.3	heaven	2	3.3	i	263	6.2
miracle	8	13.9	emperor	2	5.2	read	9	9.5	inwardly	3	5.5	hook	2	3.3	she	130	5.8
him	26	13.7	Jesus	2	4.5	talk	6	7.4	hard	7	5.3	time	13	3.2	muslim	3	5.7
forgiveness	7	13.4	spirit	2	4.3	sing	4	6.4	daily	3	4.4	night	5	3.1	follower	3	5.0
you	23	13.2	image	2	4.0	watch	4	5.0	only	20	3.8	lord	2	2.7	Jesus	5	4.8
me	24	13.1	wind	2	3.9	live	3	3.9	continually	3	3.7	pardon	2	2.7	jew	3	4.5
deliverance	6	13.0	him	6	3.3	work	5	3.5	regularly	5	3.5	soul	2	2.4	church	7	4.5
them	23	12.2				wish	2	3.4	often	10	3.3	silence	3	2.4	fellowship	2	4.0
church	12	11.7				believe	2	2.9	ever	9	3.0				Singh	2	3.7
guidance	8	11.6				learn	2	2.8	secretly	2	2.7				Family	6	3.6
us	16	11.6				tell	2	2.3	quietly	3	2.4						
chance	5	10.3							still	11	2.3						

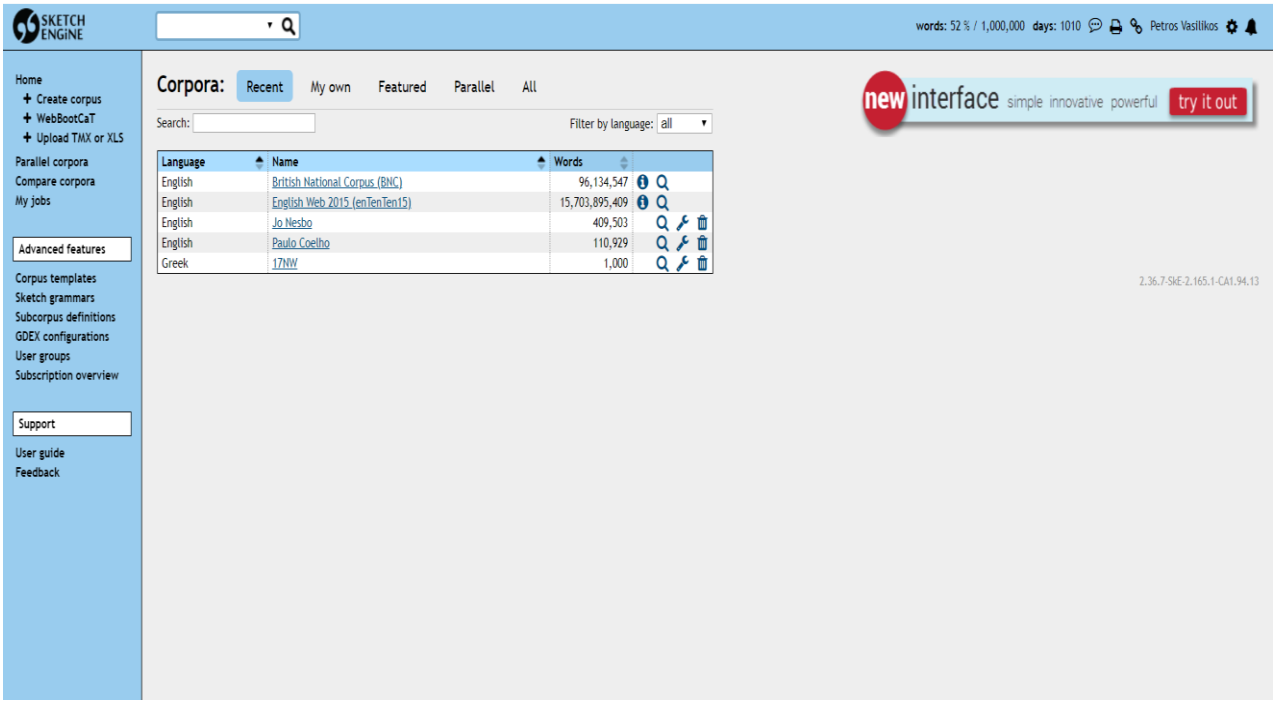
Word sketch for *pray (v)*

Εικόνα 31 Παράδειγμα σκίτσου του ρήματος Pray

Τέλος το Sketch Engine χρησιμεύει ως λογισμικό κατασκευής σωμάτων κειμένων. Χρησιμοποιεί την τεχνολογία WebBootCaT, για να δημιουργήσει αυτόματα σώμα κειμένων από σχετικές ιστοσελίδες. Τα δεδομένα που λαμβάνονται από το διαδίκτυο καθαρίζονται, και το μη κείμενό τους εξαλείφεται. Ο χρήστης μπορεί να καθορίσει ποιο περιεχόμενο πρέπει να μεταφορτωθεί μέσω μιας από τις παρακάτω επιλογές: 1) παρέχοντας μερικές τυπικές λέξεις που ορίζουν το θέμα (λέξεις σπόρων), 2)

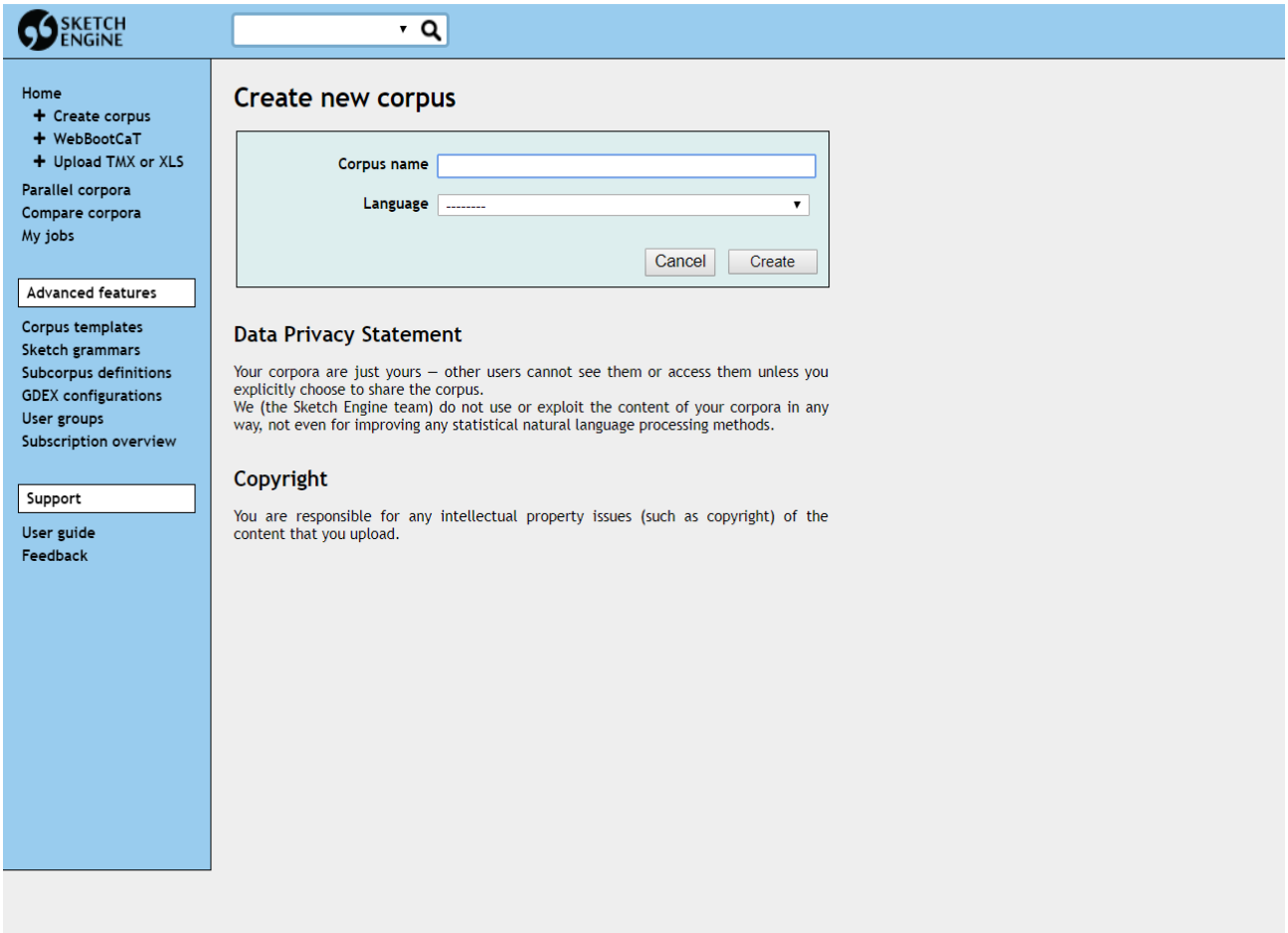
παρέχοντας μια λίστα με τις διευθύνσεις URL που πρέπει να μεταφορτωθούν, 3) με τη λήψη ενός πλήρους ιστότοπου. (Sketch Engine, 2004).

3.3.2 Δημιουργία Corpus με το Sketch Engine



Εικόνα 32 Δημιουργία Corpus με Sketch Engine Βήμα 1ο

Η παραπάνω φωτογραφία είναι η αρχική σελίδα μετά τη σύνδεση. Να διευκρινίσουμε ότι για λόγους ασυμβατότητας χρησιμοποιήθηκε το παλιό Interface του προγράμματος : https://old.sketchengine.co.uk/ . Έπειτα πατάμε «Create Corpus»



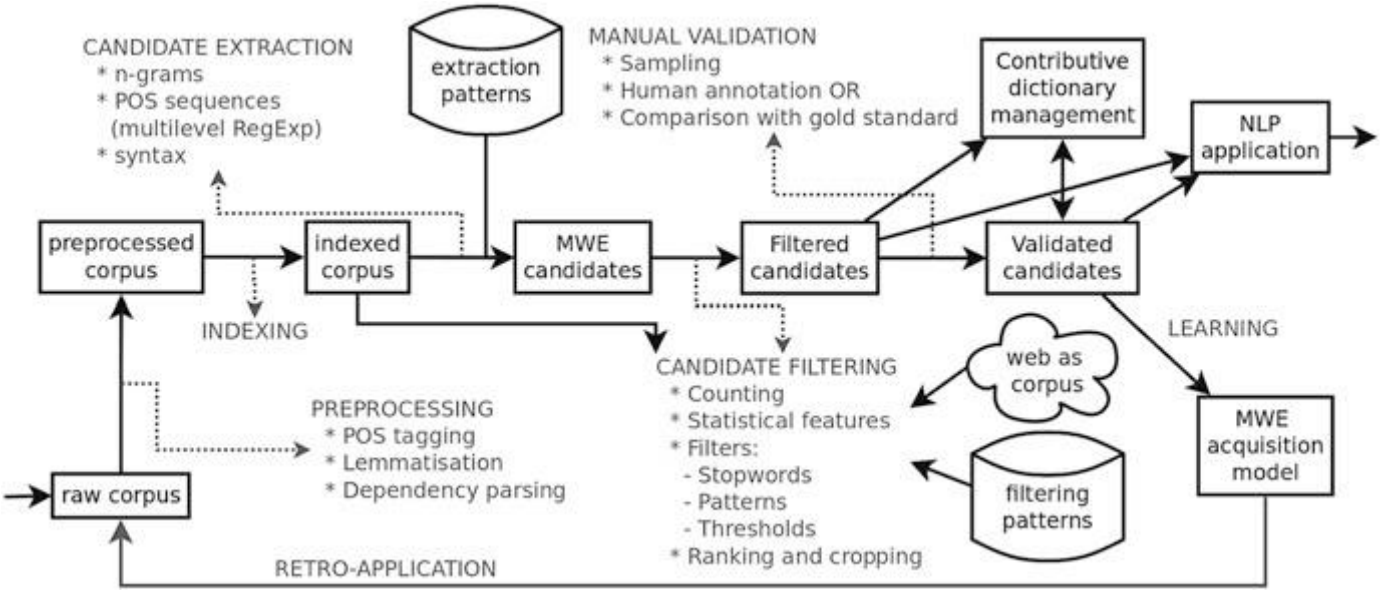
Εικόνα 33 Δημιουργία Corpus με Sketch Engine Βήμα 2ο

Εδώ εμφανίζεται η σελίδα που θέτουμε το όνομα του corpus μας και τη γλώσσα την οποία θέλουμε.

3.4 Mwe Toolkit

Το εργαλείο Mwe Toolkit μας βοηθάει στην αυτόματη αναγνώριση των πολυλεκτικών εκφράσεων λειτουργώντας πάνω σε ένα δοσμένο κείμενο σε μορφή txt. Η βασική του λειτουργία είναι η επισημείωση των πολυλεκτικών εκφράσεων πάνω στο κείμενο αφού προηγηθεί η αναγνώριση τους. Αυτό το επιτυγχάνει αντίστοιχα με τα προηγούμενα προγράμματα, χρησιμοποιώντας τα διάφορα υποπρογράμματα που ενσωματώνονται στον πυρήνα του. Το εργαλείο αυτό αρχικά προοριζόταν για χρήση από λεξικογράφους και ορογράφους αφού θεωρείται δεδομένο ότι κάθε φορά που δίνεται ένα κείμενο μιας συγκεκριμένης γλώσσας και τομέα μπορεί το πρόγραμμα να αναγνωρίσει ομάδες λέξεων που πιθανόν να είναι πολυλεκτικές εκφράσεις. Στη

συνέχεια χρησιμοποιήθηκε και για λειτουργίες όπως μηχανική μετάφραση κ.α. (Ramisch, 2015, pp. 129-139)



Εικόνα 34 Τρόπος Λειτουργίας Mwe Toolkit

ΚΕΦΑΛΑΙΟ 4

ΑΞΙΟΛΟΓΗΣΗ ΕΡΓΑΛΕΙΩΝ

Στην ενότητα αυτή θα αξιολογήσουμε τα τρία εργαλεία συγκέντρωσης κειμενικού υλικού από το διαδίκτυο (crawlers), δηλαδή τα Sketch Engine, Bootcat και Ilsp Focused Crawler. Στη συνέχεια, θα δώσουμε μερικά πειραματικά δεδομένα από τη χρήση του Ilsp Crawler και θα συγκρίνουμε την απόδοσή του εν σχέσει με την χειρωνακτική συλλογή κειμενικού υλικού από το διαδίκτυο.

4.1. Τα τρία εργαλεία συγκέντρωσης κειμενικού υλικού από το διαδίκτυο (crawlers)

Όπως έδειξε η έρευνά μας ο πιο αποτελεσματικός τρόπος συλλογής κειμενικού υλικού ήταν η χρήση του Crawler. Συνοψίζουμε τους λόγους για τους οποίους καταλήξαμε σε αυτήν την επιλογή.

Το Bootcat αν και είναι ένα ελεύθερο λογισμικό με σχετικά απλό τρόπο χρήσης, αφού προορίζεται και για χρήστες που δεν έχουν ιδιαίτερες γνώσεις υπολογιστών, δεν επιτρέπει την εύρεση μεγάλου όγκου αποτελεσμάτων. Επίσης χρησιμοποιεί μόνο τη μηχανή αναζήτησης της Google ως πηγή για να τροφοδοτήσει την έρευνά του. Τέλος ένας μέσος χρήστης δεν θα μπορούσε να διαθέτει τις απαραίτητες γνώσεις για να παραμετροποιήσει το πρόγραμμα σε τεχνικό επίπεδο κάτι που αναφέρεται στην ιστοσελίδα του κατασκευαστή του.

Το Sketch Engine αν και αρκετά αποτελεσματικό όσον αφορά τη χρήση του για την εξαγωγή αποτελεσμάτων, βασίζεται κυρίως σε ήδη κατασκευασμένα σώματα κειμένων που διαθέτει στη βάση του. Αυτό κατά συνέπεια περιορίζει τα αποτελέσματα αφού τα δεδομένα που υπάρχουν στη διάθεση του προγράμματος είναι πεπερασμένα. Βέβαια υπάρχει η δυνατότητα κατασκευής σώματος κειμένων από τον χρήστη αλλά αυτό κρύβει κάποιες δυσκολίες. Αρχικά εξ ορισμού τα links μιας σελίδας που επισκέπτεται το Sketch Engine μέσω της μηχανής αναζήτησης Bing δεν θα ακολουθηθούν άρα περιορίζεται σημαντικά το μέγεθος του Corpus μας. Επομένως αν θέλαμε να δώσουμε κάποια αρχικά links ως seed-urls από τη χειροκίνητη αναζήτησή που κάναμε νωρίτερα, αυτό δεν θα ήταν εφικτό να πετύχει το σκοπό που θα θέλαμε. Άλλη επιλογή είναι το

Sketch Engine να κάνει αναζήτηση κατευθείαν στο διαδίκτυο όπου δίνουμε 3 ή παραπάνω λέξεις σπόρους «seeds» που καθορίζουν το θέμα της αναζήτησης μέσω του Bing. (Sketch Engine, 2004)

Καταλήγοντας, ο Ilsp Focused Crawler, με τον οποίο ασχοληθήκαμε ιδιαίτερα σε αυτήν την εργασία, αποδείχτηκε όπως θα δούμε και στη συνέχεια ο πιο κατάλληλος και αποτελεσματικός για την έρευνά μας. Με τον Crawler μπορεί ο χρήστης να παραμετροποιήσει την αναζήτησή του με διάφορους τρόπους με βάση το αποτέλεσμα που θέλει να επιτύχει. Έπειτα χάρη στον τρόπο λειτουργίας του ο χρήστης μπορεί να οδηγηθεί στη δημιουργία μεγαλύτερων σωμάτων κειμένων και συνεπώς σε περισσότερα συναφή αποτελέσματα. Επειδή ο Crawler έχει την δυνατότητα να χρησιμοποιεί τα «θέματα» των κλιτών λέξεων, χειρίζεται σε σημαντικό βαθμό την μορφολογία της Ελληνικής. Τέλος, έχει τη δυνατότητα να αποφύγει την συλλογή διπλοτύπων, πράγμα που είναι σημαντικό για την γλωσσική έρευνα που παίρνει υπόψη τις συχνότητες εμφάνισης των φαινομένων.

Λογισμικό	Δυνατότητα ρυθμίσεων	Απαλοιφή διπλοτύπων	Αυτοματοποιημένη Ανάλυση	Χειρισμός της μορφολογίας	Φιλικό περιβάλλον χρήστη
Bootcat	-	-	-	-	+
Ilsp Crawler	+	+	-	+	-
Sketch Engine	-	+	+	-	+

Πίνακας 1 Σύγκριση Εργαλείων

4.2 Συλλογή υλικού με τον crawler σε σχέση με την χειρωνακτική μέθοδο

Για την έρευνα μας, δόθηκε χειρωνακτικά συγκεντρωμένο υλικό από προηγούμενη έρευνα. Με μια εξατομικευμένη εφαρμογή στο Facebook 260 ομιλητές της νεοελληνικής κλήθηκαν να προσδιορίσουν ποιες από τις 152 παρομοιώσεις που είχαν στη διάθεσή τους θα χρησιμοποιούσαν στον καθημερινό τους λόγο. Από αυτές βρέθηκαν 85 συγκρίσεις να χρησιμοποιούνται με μεγάλη συχνότητα και κατέληξαν σε 20 από αυτές προς έρευνα. Η συλλογή παραδειγμάτων χρήσης τους έγινε από το

διαδίκτυο με αναζήτηση όλων των πιθανών συντακτικών και λεκτικών παραλλαγών με ανάλογα ερωτήματα στη μηχανή αναζήτησης Google. Έπειτα μετρήθηκε στις παρομοιώσεις που βρέθηκαν μέσα σε αυτά τα παραδείγματα, η σχέση μεταξύ της σημασίας τους και της συντακτικής ευελιξίας τους. Τέλος καθαρίστηκε το σώμα κειμένων από διπλότυπα και από παράγωγα μηχανικής μετάφρασης. Για την δική μας έρευνα χρησιμοποιήθηκαν 4 από αυτές τις Π.Λ.Ε. (Stella Markantonatou, Panagiotis Kouris, Katerina Selimi, Dimitra Stasinou, Yianis Maistros, 2019)

ΠΛΕ	Αρχικά παραδείγματα	Ενεργά links	Επισκέψιμα links	Παραδείγματα Ilsp crawler	Τελικά links	MWE Toolkit patterns
Άσπρος σαν το πανί	408	374	113	571	571	460
Στολισμένος σαν φρεγάτα	16	22	13	6	6	-
Ντυμένος σαν αστακός	50	25	20	-	-	-
Κόκκινος σαν αστακός	313	-	-	-	-	-

Πίνακας 2 Παραδείγματα αποτελεσμάτων Ilsp Crawler από δοσμένο υλικό)

Αρχικά παραδείγματα: όσοι δεσμοί δόθηκαν από την χειρωνακτική έρευνα

Ενεργά links: όσοι δεσμοί από τα αρχικά παραδείγματα ήταν λειτουργικοί κατά τη φάση των πειραμάτων με τον ILSP Crawler. (Λόγω του χρόνου που μεσολάβησε από τη συλλογή των links της αρχικής έρευνας μέχρι της δικής μας, πολλά links δεν ήταν πλέον ενεργά)

Παραδείγματα ILSP Crawler: τα παραδείγματα που συγκέντρωσε ο ILSP Crawler

Τελικά links: οι δεσμοί για τα παραδείγματα που συγκέντρωσε ο ILSP Crawler

4.3 Εντοπισμός ΠΛΕ στο Τρίτο Στεφάνι

Χρησιμοποιήσαμε την δυνατότητα των φραστικών σχημάτων (patterns) του MWE Toolkit για να βρούμε τις συχνότερες ρηματικές ΠΛΕ στο Τρίτο Στεφάνι. Χρησιμοποιήσαμε τα εξής φραστικά σχήματα:

(Pn)+(Vb)+Vb+(Ad)+(At)+(Aj)+No+(Pn)

(At)+No+(Pt)+(Pn)+(Vb)+Vb

Vb+Cj+Vb

(Pn)+No+(Pt)+(Pn)+(Vb)+Vb

(Pn)+Pn+(Vb)+Vb

(Pn)+(Vb)+Vb+(At)+(No)+(Ad)+AsPp+(At)+No

Οι ρηματικές ΠΛΕ με δύο ή περισσότερες εμφανίσεις που ανασύρθηκαν με τον μηχανισμό αυτόν δίνονται στον Πίνακα 3. Παρατηρούμε ότι δεν υπάρχουν παγιωμένες παρομοιώσεις σε αυτές.

λέω με το νου	58
λέω λέξη	13
περνάω την ώρα μου	13
παίρνω αέρα	10
κάνω παρέα	8
ρίχνω μια ματιά	7
ανοίγω το στόμα μου	6
ένας θεός το ξέρει	6
βγάζω τα μάτια μου	6
πέφτω στα χέρια	6
κάνω σκηνή	6
να μη βλέπω στα μάτια μου	6
πέφτω στα πόδια κάποιου	6
ορκίζομαι στα κόκκαλα κάποιου	6
πατάω πόδι	5
δε δίνω πεντάρα	5
χαρά σε κάποιον/κάτι	5
χτυπάω την πόρτα στη μούρη στα μούτρα κάποιου	5
βάζω στην άκρη	5
γίνομαι θηρίο	5
καρφώνω κάποιον με τα μάτια	5
αλλάζω χρώμα	4

κάνω γούστο	4
αδειάζω τη γωνιά	4
γίνομαι άνθρωπος	4
ματώνει η καρδιά μου	4
κλείνω τα μάτια σε κάτι	4
κακό χρόνο να 'χεις	4
να πάρει ο διάβολος	4
κάνω τη ζωή μαύρη	3
δίνω καιρό	3
μου ανακατεύονται τα άντερα	3
δίνω στα νεύρα	3
έχω στο μάτι κάτι	3
σφίγγεται η καρδιά μου	3
γίνομαι βαπόρι	3
για να σκάσεις	3
η ψυχή μου το ξέρει	3
βρίσκω το μπελά μου	3
τρώω κάποιον	3
μου κάνει καρδιά	3
ας όψεται	3
θα κάνω κάποιον κομματάκια	3
ρουφάω το μεδούλι κάποιου	3
το ρίχνω σε κάτι	3
το λέω με την καρδία μου	3
δε βαριέσαι	3
κάνω θέλημα κάποιου	3
φέρνω κάποιον ως το λαιμό	3
παίρνω κάτι χαμπάρι	3
λέω τη μοίρα	3
είμαι άνθρωπος	2
ποιο είναι το ποιόν κάποιου	2
έχω στην καμπούρα μου	2
γίνομαι λέαινα	2
ανάβω και κορώνω	2
το παίρνω πάνω μου	2
ψήνω το ψάρι στα χείλη	2
παγώνει το (χαμόγελο)γέλιο στα χείλια μου	2
τρώω το κεφάλι μου	2
σπάω το κεφάλι μου	2
πατώντας στα νύχια	2
πέφτω με τα μούτρα	2
κόβει το μάτι μου	2
μένω στο ράφι	2
την παθαίνω	2
τι χρωστάει ο άνθρωπος κάποιος σε κάποιον	2
χάνω το χρώμα μου	2

λύνεται ο αφαλός μου	2
κάνω χωρατά	2
με παίρνει η μυρωδιά από τη μύτη	2
κάνω τα γούστα κάποιου	2
ξέρω τη γλύκα	2
μένω στον τόπο	2
αφήνω σε χλωρό κλαρί	2
καταλαβαίνω την καρδιά κάποιου	2
τρώω τα ψωμιά μου	2
εννοώ να κάνω κάτι	2
κρεμιέμαι από τη φούστα κάποιας	2
κάνω του κεφαλιού μου	2
είμαι στο χώμα	2
τρίβω κάποιον στην κασίδα μου	2
κάνω τεμενάδες	2
πιάνω/ομαι κορόϊδο	2
τρώω κάποιον ζωντανό	2
γίνομαι μπαρούτι	2
κάνω τα στραβά μάτια	2
λέω την αμαρτία μου	2
πάει κάποιος	2
ανοίγω το σπίτι μου	2
ζω και καζαντίζω	2
κόβεται η αναπνοή μου	2
χαϊρι και προκοπή να μη δω	2
ανεβάζω, κατεβάζω κάποιον κάτι	2
κάτι/κάποιος είναι σε κακά χάλια	2
φέρνω κάποιον σβούρα	2
χάνω κάποιον	2
με τσιμπάει μύγα τσε τσε	2
κάνω κάποιον σαν τα μούτρα μου	2
έτσι κι έτσι έχουν τα πράγματα	2
η διαίσθησή/ενστικτό μου με γελάει	2
κάνω τα χατίρια κάποιου	2
τρομάρα στα μπαντζάκια μου	2
μαλώνω για ψύλλου πήδημα	2
κουρδίζω κάποιον	2
έχω κάποιον άξιο για κάτι	2

Πίνακας 3 Οι ρηματικές ΠΛΕ που ανασύρθηκαν με δύο ή περισσότερες εμφανίσεις.

Παρατηρούμε ότι η ΠΛΕ λέω με τον νου μου εμφανίζεται πολύ συχνά (58 φορές). Μία αναζήτηση στο κείμενο την έφερε 70 φορές, πράγμα το οποίο σημαίνει ότι τα φραστικά σχήματα έχασαν κάποια παραδείγματα, προφανώς αυτά στα οποία η χρήση εμφάνιζε μεγαλύτερη ευελιξία. Στα παραδείγματα που βλέπουμε στην συνέχεια, υπάρχει γλωσσικό υλικό ανάμεσα στα συστατικά της ΠΛΕ το οποίο δεν προβλέπεται

από τα φραστικά σχήματα. Δεν βρήκαμε όμως άλλες μορφές της ΠΛΕ πλην αυτής και της κανονικής.

Ίσως -λέω καμιά φορά με το νου μου- να 'ταν γραφτό απ' το θεό να τον πάρω για να τραβήξω όσα τράβηξα.

Στο τέλος είπα κι εγώ με το νου μου: στο διάβολο!

Από αυτό το δεδομένο και επειδή η συχνότητα χρήσης των διαφόρων ΠΛΕ στο Τρίτο Στεφάνι είναι πολύ χαμηλή υποθέτουμε ότι ο Ταχτσής χρησιμοποιεί την κυρίως την κανονική μορφή των ΠΛΕ. Η κανονική μορφή των ΠΛΕ υπερισχύει συντριπτικά και στα δεδομένα μας από τον Παγκόσμιο Ιστό.

Επειδή στις ΠΛΕ με περισσότερες από δύο εμφανίσεις δεν βρέθηκαν παγιωμένες παρομοιώσεις, χρησιμοποιήσαμε ειδικά φραστικά σχήματα για την ΠΛΕ «άσπρος σαν το πανί». Χρησιμοποιήθηκαν δύο μοτίβα, το δεύτερο με ανεστραμμένη την σειρά των όρων. Τα μοτίβα περιέχουν τις δύο βασικές λέξεις της παγιωμένης παρομοίωσης με την σειρά που δίνονται και επιτρέπουν μηδέν έως άπειρες λέξεις στα ενδιάμεσα. Δεν βρέθηκε η συγκεκριμένη παρομοίωση στο Τρίτο Στεφάνι μολονότι αξιοποιήθηκαν και οι δυνατότητες συντακτικής ευελιξίας της συγκεκριμένης παγιωμένης παρομοίωσης.

Τέλος, ενδεικτικά χρησιμοποιήσαμε την δυνατότητα των μοτίβων (patterns) του MWE Toolkit για τα αρχεία «άσπρος σαν το πανί» και «ένας θεός το ξέρει» με τα μοτίβα που δίνονται στους Πίνακες 4 και 5. Για την ΠΛΕ «άσπρος σαν το πανί» χρησιμοποιήθηκαν δύο μοτίβα, το δεύτερο με ανεστραμμένη την σειρά των όρων. Τα μοτίβα περιέχουν τις δύο βασικές λέξεις της κάθε ΠΛΕ με την σειρά που δίνονται και επιτρέπουν μηδέν έως άπειρες λέξεις στα ενδιάμεσα.

```
<?xml version="1.0" encoding="UTF-8"?>
<patterns>
<pat>
<w pos="No*" lemma="θεός" />
<!--  ignore anything intervening -->
<pat repeat="*" ignore="true"><w/></pat>
<w pos="Vb*" lemma="ξέρω"/>
</pat>
</patterns>
```

Πίνακας 4 Μοτίβα για την ΠΛΕ «ένας θεός το ξέρει».

<?xml version="1.0" encoding="ISO-8859-1"?>
<patterns>
<pat> <w lemma="άσπρος" pos="Aj*"/> <pat ignore="true" repeat="*"><w/></pat> <w lemma="πανί" pos="No*"/> </pat>
<pat> <w lemma="πανί" pos="No*"/> <pat ignore="true" repeat="*"><w/></pat> <w lemma="άσπρος" pos="Aj*"/>
</pat>
</patterns>

Πίνακας 5 Μοτίβα για την ΠΛΕ «άσπρος σαν το πανί»

Στη συνέχεια χρησιμοποιήσαμε τον ILSP crawler για να συγκεντρώσουμε παραδείγματα για μερικές από τις ΠΛΕ που εντοπίστηκαν στο Τρίτο Στεφάνι. Τα αποτελέσματα φαίνονται στον Πίνακα 3.

ΠΛΕ	Αρχικά παραδείγματα	Ενεργά links	Επισκέψιμα links	Παραδείγματα Ilsp crawler	Τελικά links	MWE Toolkit patterns
Ένας θεός το ξέρει	12	12	12	38547	38547	1313
Γίνομαι άνθρωπος	10	-	-	-	-	-
Πέφτω στο χέρι κάποιου	12	12	12	-	-	-
Μόνος και έρημος σαν την καλαμιά στον κάμπο	16	16	16	29	29	-

Πίνακας 6 Παραδείγματα που συγκεντρώθηκαν από το διαδίκτυο με τον crawler. Οι ΠΛΕ εμφανίζονται στο Τρίτο Στεφάνι με συχνότητα μεγαλύτερη του 2 εκτός της παγιωμένης παρομοίωσης που εμφανίζεται μία φορά μόνο.

Η χρήση της δυνατότητας των μοτίβων με το MWE Toolkit μας δίνει μία άμεση εικόνα για το πόσες χρήσεις της ΠΛΕ έχει ανασύρει ο crawler πραγματικά. Έτσι, ο crawler για την ΠΛΕ «άσπρος σαν το πανί» έχει ανασύρει 460 εμφανίσεις ενώ για την «ένας θεός το ξέρει» 1313.

4.4 Συζήτηση

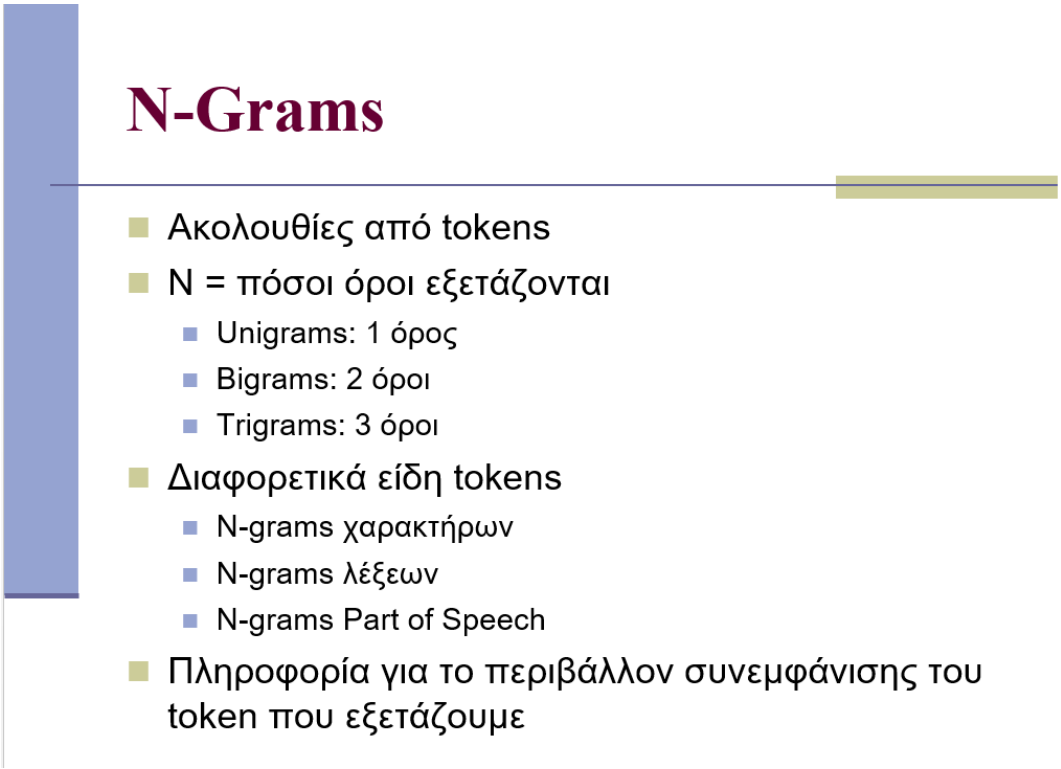
Παρατηρούμε ότι τα αποτελέσματα δεν μας επιτρέπουν να συνάγουμε βάσιμα συμπεράσματα αλλά μόνο να προχωρήσουμε σε κάποιες υποθέσεις που απαιτούν περισσότερη διερεύνηση:

1. Ο crawler απεδείχθη πιο αποτελεσματικός στην συγκέντρωση κειμένων από τον άνθρωπο για την υψηλής συχνότητας παγιωμένη παρομοίωση *άσπρος σαν το πανί* αλλά λιγότερο αποτελεσματικός για τις χαμηλότερης συχνότητας παγιωμένες παρομοιώσεις που χρησιμοποιήθηκαν στο πείραμα. Χρειάζεται περισσότερος πειραματισμός για να γίνει κατανοητό εάν η συμπεριφορά του crawler εξαρτάται από το εάν μια ΠΛΕ είναι σπάνια ή όχι ή εάν συνδέεται με κάποιο είδος λόγου (π.χ. έντονα λαϊκό λόγο, προσβλητικό λόγο, τυπικό λόγο κλπ).

2. Σε σχέση με το Τρίτο Στεφάνι βλέπουμε ότι η μάλλον συχνή στο μυθιστόρημα ΠΛΕ ένας θεός το ξέρει έφερε πάρα πολλά αποτελέσματα (1313 περιπτώσεις). Θα μπορούσε κανείς να πει ότι εδώ η γλώσσα του Ταχτσή φαίνεται να συμβαδίζει με την αποτύπωση της Ελληνικής στον Παγκόσμιο Ιστό. Όμως, η επίσης συχνή στον Ταχτσή ΠΛΕ γίνομαι άνθρωπος δεν έφερε αποτελέσματα. Εάν είχαμε εμπιστοσύνη στον crawler θα λέγαμε ότι εδώ έχουμε μια ένδειξη για το πώς διαφέρει η γλώσσα που χρησιμοποιείται στο Τρίτο Στεφάνι από την ομιλούμενη γλώσσα. Εδώ, βέβαια, γίνεται η συζητήσιμη παραδοχή ότι ο Παγκόσμιος Ιστός δίνει μια αντιπροσωπευτική εικόνα της ομιλούμενης γλώσσας.

ΠΑΡΑΡΤΗΜΑ

N-Grams



(McCoy, 2018)

ILSP Crawler: log file

log_test (1) - Notepad

File Edit Format View Help

τεραστ ασπρ χερ εκοβ καθ αντιπαλ ασχημ γιατ ητ πιο ευφυ αστει χαμογελ στραβ δοντ εφερν παντ τ νικητ τ χαμεν κοντ συμφων μπουλινγκ ητ καθημερινοτητ χαρακτηρισμ επεφτ βροχ λε φανοποι γον φωναζ απ μπαλκον ξαναπ ετσ φιλ σου ανεβ σπιτ ενα βραδ αστοχ τριποντ ταπ μασ εκ εξαλλ κουκουβαγ εχασ λει απ νικ περασ στ αντιπαλ κοσμητ βροχ απειλ σπρωξιμα αλληλοκατηγορι σκληρ μπουλινγκ ολ ολ μεχρ φτασ πρασιν λαστιχ νερ ξεπλεν παντ υπηρχ νικητ ηττημεν χοντρ ασχημ μυωπ κοντ ημαστ παρε φιλ παιδ καθ ολοκληρ καλοκαιρ πεζουλ μπαλ ελειπ καποι πηγαιν σπιτ ποδ δ εχ κατ καν εμ μπουλινγκ μον παρε επεβαλ τ κανον ορ στεγαν σκληρ παιδ χωματοδρομ μοιραζ δυνατ αδυναμ γκ στρειτ πλουσι φτωχ ασχημ ομορφ παρε ητ μια καν εμπαιν περιθωρι καν ητ καλ κανεν ολ ειχ ιδ αξ σιγουρ λογ πληγων στιγμ σιγουρ ητ λαθ σιγουρ καν μπουλινγκ ανοιξ πρασιν λαστιχ ηπ λιγ νερ ζητ συγγνωμ ταπ φλοκ χοντρ ασχημ ητ γιατρ καθηγητ δικηγορ φανοποι ετρωγ κατ πευκ συγγνωμ μασ ρε κουκουβαγ μου ειπ παιδ ξεπλυν μπουλινγκ σ ενα πρασιν λαστιχ πριν ποτισ τισ καρδι μασ παρε μασ διδαξ τι σημαιν μοναδικ ετικετ

---

http://www.clickatlife.gr/your-life/story/14632

κλισ φρασ αποκωδικοπο επιστημον πεμπτ μαι γνωριζ τι σημαιν νιωθ πεταλουδ στομαχ μηπως τι συμβαιν οτ κοιμ ποδ σασ οτ γιν ασπρ σαν παν ολ φρασ τυχαι εχ προκυψ επειδ ακριβως σωμ αντιδρ ορισμεν καταστασ ειδικ απαντ αποκαλυπτ ολα ιατρ κλισ κατεληξ αποτελ καθημερινρητ νιωθ πεταλουδ στομαχ αισθημ πεταλουδ στομαχ πριν μια μεγαλ παρουσιασ ειδικ εκδηλωσ πραγματικοτητ αντιδρασ εντερ νευρ συστημ οργανισμ κατ πιεσ οργανισμ ενεργοποι αντιδρασ φυγ παλ επομενως αισθησ πεταλουδ αφορ φοβ τη νευρικοτητ οποι μερ στρ αισθαν οργανισμ ιδι ακριβως προκαλ ακομ αυξησ καρδιακ ρυθμ αρτηριακ πιεσ καταστασ προετοιμασι μια μαχ κυριολεκτ μεταφορ αρρωστ ερωτ βαθ συναισθημα ενα αλλ προσωπ προκαλ αυπν ταχυκαρδ απωλει βαρ προβλημα συγκεντρωσ μπορ σημαιν κατ περισσ ενα απλ φλερτ μπορ συμπτωμα εν ερωτοχτυπημεν συμπτωμα μπορ εμφανιζ ειτ οτ καποι ευτυχισμεν ειτ δυστυχισμεν πρωτ περιπτωσ μαλιστ μπορ περιλαμβαν ιδεοφυχαναγκαστικ τασ ελεγχ επανελεγχ τηλεφων μηνυμα συναισθημα αναδυ λογ αυξησ ορισμεν νευροδιαβιβαστ ντοπαμιν νορεπινεφριν οξυτοκιν περιπτωσ αναπτυξ πον στηθ δυσπνοι γενικ αδυναμ καλ ητ επισκεφτ γιατρ σασ αδρεναλιν στα υψη αισθαν συνεπαρμεν μια προπονησ εξοικειωμεν εξαψ δρομε προπονησ αντοχ γνωστ προκαλ πολλ θετικ ψυχικ σωματικ αλλαγ συμπεριλαμβανομεν μειωσ στρ αγχ αντιληψ πον ταυτοχρον αυξημεν διαθεσ μαλιστ εξαψ πολλ φορ συγκριν δρασ οργανισμ ναρκωτικ ουσι ευτυχωσ χημικ ουσι εγκεφαλ παραγ εξαψ εχ επικινδυν παρενεργει υπν ομορφι εχ απαλλαξ ποτ εαυτ σασ μια βαρετ εκδηλωσ γιατ χρειαζ υπν ομορφι σασ χρειαζ δικαιολογειστ καθωσ επιστημ σασ υποστηριζ υπαρχ αδιασειστ στοιχει συνδε υγει δερμα υπν σημαντικ αναφερ υπν κυτταρ αναζωογον προσφερ λαμψ ομορφ επιπλεον στερησ υπν αυπν συνδε αυξημεν ορμον στρ μειων λειτουργ ανοσοποιητ συστημα αυξησ κινδυν φλεγμον δερμα κοιμ ποδ σασ οτ παραμεν καθισμεν ιδ θεσ μεγαλ χρον διαστημ μπορ αισθανθ σα σασ τρυπ βελον οτ προσπαθ σηκωθ συνθηκ οπ ποδ σασ κοιμ πραγματικοτητ συμβαιν μια παρατεταμεν πιεσ στα ποδ οποι οφειλ περιορισμ ρο αιμα στα νευρ προκαλ λειτουργ σωστ αρτηρι τροφοδοτ αιμ θρεπτ συστατ νευρ ποδ συμπιεζ οτ συμβαιν νευρ κυτταρ μπορ λειτουργ σωστ εγκεφαλ στελν αναμεικτ μηνυμα αποτελεσμ εχ αισθησ καψιμα μουδιασμα νομιζ σασ τσιμπ καρφιτσ ραγιζ καρδ εκτ φυσιολογ αισθημ θρην προκαλ μια σοβαρ συναισθηματικ απωλει θανατ ραγισμεν καρδ συνηθ φαινομεν βαρι θλιψ επισ γνωστ καρδιομυοπαθει takotsubo takotsubo ενα ιαπων ονομ μια παγιδ τη μορφ χταποδ στρ μυοκαρδιοπαθει συνδρομ εμφανιζ πολ πιο συχν στισ γυναικ τι τ ανδρ αγνωστ λογ κυριως αγχωτικ καταστασ μια σοβαρ ιατρικ διαγνωσ καταστροφικ οικονομικ απωλει φυσικ καταστροφ συχν συνδε σημαντικ καταθλιψ μπορ μιμ μια καρδιακ προσβολ στεφανιαι αρτηρι παραμεν ανεπηρεαστ επιστημον πιστευ σημαντ ρολ τισ καταστασ παιζ κατεχολαμιν χημικ ουσι σωμ νορεπινεφριν γνωστ ορμον στρ πεθαν φοβ μου μπορ καποι φιλ σασ κρυφτ μεσ σκοταδ σασ τρομαξ τοσ ωστ πεθαν φοβ σασ σιγουρ απιθαν καθως ειδικ αναφερ καρδ μπορ σταματ φοβ δυνητ μπορ συγκεκριμεν καταστασ ωστοσ ανθρωπ μπορ φοβ τοσ πολ συνεχει υποφερ οξει καρδιακ προσβολ οποι οφειλ τερασ απελευθερωσ αδρεναλιν εξαιτι φοβ απελευθερωσ αδρεναλιν μπορ προκαλεσ καρδιακ αρρυθμ ανωμαλ καρδιακ ρυθμ τη σειρ στεφανιαι νοσ ετσ ενα ατομ μπορ πεθαν καρδιακ προσβολ αποτελεσμ μι αιφνιδι τρομακτικ καταστασ πιθαν πεθαν ιδι φοβ εγιν ασπρ σαν παν εχ υπαρξ στιγμ εχ γιν ασπρ σαν παν εχ χλομιασ τοσ πολ ωστ αλλαξ χρωμ δερμα σασ οτ συμβαιν ουσ αιμ απομακρυν συγκεκριμεν περιοχ σωμα διαφορ λογ απλ λογ συμβαιν ιδι καταστασ οπ καποι αρρωστ οργανισμ μπορ προσπαθ ελεγξ τη θερμοκρασ σωμα γεγον προκαλ μια χλωμ εμφανισ ιδ λογικ φοβισμεν απελευθερωσ ορμον στρ προκαλ συστολ αιμοφορ αγγει προσωπ συνεχει στελν αιμ στ μυσ μερ σωμα μπορ αναγκαι επιβιωσ ποδ μπορεσ τρεξ χρειαστ ad links

---

ΒΙΒΛΙΟΓΡΑΦΙΑ

Apache Hadoop. (2019, Μάϊος 7). Ανάκτηση από https://hadoop.apache.org/

Apache Hadoop. (2019, Μάϊος 7). Ανάκτηση από http://hadoop.apache.org/

Baldwin, T. a. (2010). Multiword Expressions. (N. I. Damerau, Επιμ.) *Handbook of Natural Language Processing*, 267-292.

Beaton, R. (1996). *Εισαγωγή στη νεότερη λογοτεχνία. Ποίηση και Πεζογραφία 1821-1992*. (Ε. Ζ.-Μ. Σπανάκη, Μεταφρ.) Αθήνα: Νεφέλη.

Bernardini, M. B. (2004). *BootCaT: Bootstrapping Corpora and Terms from the Web*. Ανάκτηση από http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf

Bixo. (2019, Μάϊος 22). Ανάκτηση από https://openbixo.org/

Bootcat. (2018, Ιούνιος 6). Ανάκτηση από https://bootcat.dipintra.it/

Cacciari, C. (1993). *The place of idioms in a literal and metaphorical world.*

Cygwin. (2019). Ανάκτηση από https://www.cygwin.com/

Cygwin. (2019). Ανάκτηση από https://www.cygwin.com/

Foundation, T. A. (2019). *Apache Tika - a content analysis toolkit*. Ανάκτηση από https://tika.apache.org/

https://openbixo.org/. (2019). *Bixo A Web Mining Toolkit*. Ανάκτηση 2019, από https://openbixo.org/

Kazazis, K. (1979). Learnedisms in Costas Taktsis's Third Wedding. *Byzantine and Modern Greek Studies*.

Kilgarriff, A. a. (2004). Itri-04-08 the sketch engine. *Information Technology*.

Kilgarriff, A. a. (2014). The Sketch Engine: ten years on. *Lexicography*, 7-36.

Markantonatou, S., Panagiotis, M., George, Z., Vassiliki, M., & Maria, C. (2019). *IDION: A database for Modern Greek multiword expressions* (Τόμ. Proceedings of the Joint Workshop on Multiword Expressions and WordNet

(MWE-WN 2019). Florence Italy: Association for Computational Linguistics. Ανάκτηση από https://www.aclweb.org/anthology/W19-5115

McCoy. (2018, Σεπτέμβριος). *Διαχείριση Web Περιεχομένου & Γλωσσικά Εργαλεία*. Ανάκτηση από http://www.cis.udel.edu/~mccoy/courses/cisc882.03f/lectures/lect5-ngrams.ppt

PANACEA. (2019). Ανάκτηση από http://www.panacea-lr.eu/

Prokopis Prokopidis, V. P. (2019, April 8). *ILSP Focused Crawler*. Ανάκτηση 2019, από http://nlp.ilsp.gr/redmine/projects/ilsp-fc

Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework.* (J. H. Graeme Hirst, Επιμ.) Springer.

Schönpos, M. (2008). *Russisch-deutsche Äquivalenzbeziehungen in der Phraseologie. Dargestellt an Werken N. V.*

Sketch Engine. (2004). Ανάκτηση από http://www.sketchengine.eu

Sketch Engine. (2004). Ανάκτηση από https://www.sketchengine.eu/guide/create-a-corpus-from-the-web/

Stella Markantonatou, Panagiotis Kouris, Katerina Selimi, Dimitra Stasinou, Yianis Maistros. (2019). Ψυχή άσπρη σαν το χιόνι but never ψυχή άσπρη σαν το γάλα: semasio-syntactic comments on the fixed similes of Modern Greek. Μελέτες για την ελληνική γλώσσα 39. Στο *Μελέτες για την ελληνική γλώσσα 39* (σσ. 639-650).

T.E. Jisha, M. T. (2015, February 10). *Identification of Multiword Expressions: A Literature Study*. Ανάκτηση από http://marymathacollege.org/data/downloads/2019-01-21-2-25-58_identification-of-multi-word-expressions-jisha.pdf

The Web Robots Pages. (2007). Ανάκτηση από http://www.robotstxt.org/

Vassilis Papavassiliou, P. P. (2012, October 24). *A modular open-source focused crawler for mining monolingual and bilingual corpora from the web*. Ανάκτηση από https://www.aclweb.org/anthology/W13-2506.pdf

Vitti, M. (1926). *Ιστορία της Νεοελληνικής Λογοτεχνίας.* (Μ. Ζορμπά, Μεταφρ.) Αθήνα: Οδυσσέας.

Wikipedia. (2019, 7 14). *Wikipedia*. Ανάκτηση 5 4, 2019, από Wikipedia: https://el.wikipedia.org

Wikipedia, t. f. (2019, 7 14). *Named-entity recognition*. Ανάκτηση από https://en.wikipedia.org/wiki/Named-entity_recognition

Άννα Αναστασιάδη - Συμεωνίδη, Α. Ε. (2006). *Οι στερεότυπες εκφράσεις και η διδακτική της νέας ελληνικής ως δεύτερης γλώσσας.* Αθήνα: Πατάκης.

Βικιλεξικό. (2013, 6 16). Ανάκτηση από https://el.wiktionary.org

Βικιλεξικό. (2013, 6 16). Ανάκτηση από https://el.wiktionary.org

ΒΙΚΙΠΑΙΔΕΙΑ. (2018, 6 9). Ανάκτηση 5 4, 2019, από https://el.wikipedia.org/wiki/%CE%A4%CE%BF_%CF%84%CF%81%CE%AF%CF%84%CE%BF_%CF%83%CF%84%CE%B5%CF%86%CE%AC%CE%BD%CE%B9

Η Πύλη για την ελληνική γλώσσα. (2019). Ανάκτηση από http://www.greek-language.gr/

Kohlschuetter et al. (2010). Ανάκτηση από https://www.l3s.de/~kohlschuetter/publications/wsdm187-kohlschuetter.pdf

Μαρκαντωνάτου, Σ. (2017). *Ιδιώματα & Ατομική άσκηση.* Σημειώσεις Μαθήματος, Καλαμάτα. Ανάκτηση από https://eclass.uop.gr/modules/document/file.php/LITD318/5_6_Lectures.pdf

Μπαμπινιώτης, Γ. (1998). *Λεξικό της νέας Ελληνικής γλώσσας.* Αθήνα: Κέντρο Λεξικολογίας.

Παπαβασιλείου Βασίλης, Σ. Σ. (2018). The ILSP/ARC submission to the WMT 2018 Parallel Corpus Filtering Shared Task. (σσ. 928--933). Belgium, Brussels: Association for Computational Linguistics. Ανάκτηση από https://www.aclweb.org/anthology/W18-6484

Σαραντάκος, Ν. (1997). *Το αλφαβητάρι των ιδιωματικών εκφράσεων.* Αθήνα: Δίαυλος.

Συμεωνίδης, Χ. (2000). *Εισαγωγή στην ελληνική φρασεολογία.* Θεσσαλονίκη: Κώδικας.

Ταχτσής, Κ. (1987). *Το τρίτο στεφάνι.* Εξάντας.

Χιώτη, Α. Ν. (2010). Οι παγιωμένες εκφράσεις της νέας ελληνικής: ιστορική διάσταση, ταξινόμηση και στερεοτυπικότητα. doi:10.12681/eadd/22381

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ & ΠΙΝΑΚΩΝ

Εικόνες

Πίνακες