Θέματα Επιστημών και Τεχνολογίας στην Εκπαίδευση

Τόμ. 3, Αρ. 1 (2010)

Greeklish Converter v1.00, ένα νέο περιβάλλον μετατροπής χαρακτήρων Greeklish

Αλέξανδρος Καράκος

Βιβλιογραφική αναφορά:

Καράκος Α. (2010). Greeklish Converter v1.00, ένα νέο περιβάλλον μετατροπής χαρακτήρων Greeklish. *Θέματα Επιστημών και Τεχνολογίας στην Εκπαίδευση*, *3*(1), 49–67. ανακτήθηκε από https://ejournals.epublishing.ekt.gr/index.php/thete/article/view/44633

Greeklish Converter v1.00, ένα νέο περιβάλλον μετατροπής χαρακτήρων Greeklish

Ιωάννης Παπαϊωάννου[1], Αλέξανδρος Καράκος[2], Αναστασία Γεωργιάδου[3]
karakos@ee.duth.gr

[1]Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών, Δημοκρίτειο Πανεπιστήμιο Θράκης
[2]Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστώ, Δημοκρίτειο Πανεπιστήμιο Θράκης
[3]Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών, Δημοκρίτειο Πανεπιστήμιο Θράκης

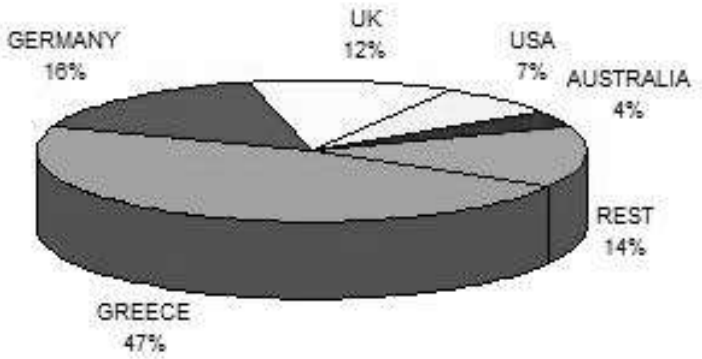
Περίληψη. Κατά την επικοινωνία στο διαδίκτυο (π.χ. αποστολή E-mails ή συμμετοχή σε forum) πολύ συχνά γίνεται χρήση του λατινικού αλφαβήτου για τη γραφή λέξεων άλλων γλωσσών. Ο τρόπος αυτός της αποτύπωσης λέξεων για την ελληνική γλώσσα είναι γνωστός ως Greeklish. Η τάση αυτή η οποία αρχικά ξεκίνησε λόγω της αδυναμίας άμεσης μεταφοράς των ελληνικών χαρακτήρων, έχει υιοθετηθεί, κυρίως χάριν ευκολίας και συντομίας, από ένα μεγάλο τμήμα των ελλήνων χρηστών του διαδικτύου, συμπεριλαμβανομένου και μεγάλου τμήματος της μαθητικής κοινότητας. Για τη διευκόλυνση των ελλήνων χρηστών του διαδικτύου προχωρήσαμε στην κατασκευή ενός εργαλείου για την εύκολη και αποτελεσματική μετατροπή κειμένων από τα Greeklish στα ελληνικά και αντίστροφα. Στόχος της εργασίας αυτής είναι η παρουσίαση του προτεινόμενου εργαλείου και του αντίστοιχου γραφικού περιβάλλοντος (Greeklish Converter v1.00) το οποίο μπορεί εύκολα να χρησιμοποιηθεί από οποιονδήποτε χρήστη σαν μια ανεξάρτητη εφαρμογή μετατροπής Greeklish στα ελληνικά και αντίστροφα με βέλτιστη απόδοση λόγω της ενσωμάτωσης και χρήσης νέων τεχνικών και ορθογραφικών λεξικών.

Εισαγωγή

Καθημερινά ο αριθμός του «διαδικτυακού πληθυσμού» αυξάνεται με αποτέλεσμα να χρησιμοποιούνται ολοένα και περισσότερο οι διαθέσιμοι στο διαδίκτυο τρόποι για σύγχρονη και ασύγχρονη επικοινωνία (Ανδρουτσόπουλος, 1998; Ανδρουτσόπουλος 2000). Σύμφωνα με τις μετρήσεις της Διεθνούς Ένωση Τηλεπικοινωνιών (ITU) των Ηνωμένων Εθνών (World Telecommunication, 2010), ο διαδικτυακός πληθυσμός παγκοσμίως έχει ξεπεράσει το έτος 2010 τα δύο δισεκατομμύρια και στην Ελλάδα ανέρχεται περίπου σε 5 εκατομμύρια (44,54% του πληθυσμού). Ένα σημαντικό ποσοστό του διαδικτυακού πληθυσμού καταλαμβάνουν και οι Έλληνες χρήστες (Lazarinis κ.α., 2007; Koutsogiannis κ.α., 2003).

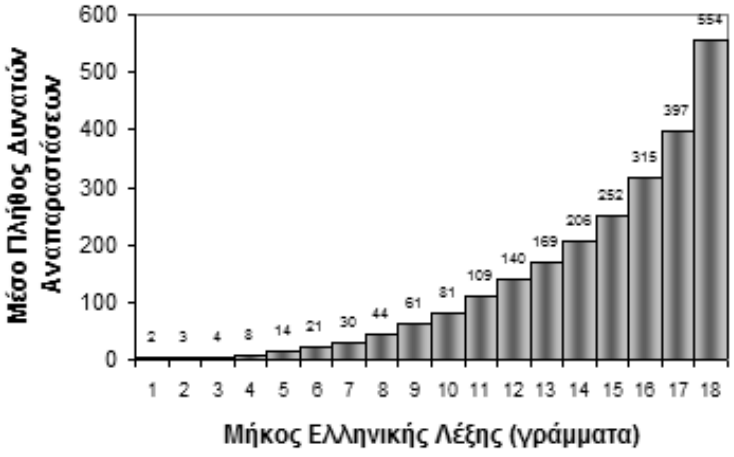
Μεταξύ των ελλήνων χρηστών του διαδικτύου καθιερώθηκε ένας ιδιαίτερος τρόπος γραφής για τη διαδικτυακή επικοινωνία, γνωστός ως λατινοελληνικά ή Greeklish, σύμφωνα με τον οποίο χρησιμοποιούνται οι χαρακτήρες του λατινικού αλφαβήτου για την απεικόνιση του ελληνικού αλφαβήτου (Androutsopoulos, 2006; Ανδρουτσόπουλος, 1999). Παρόλο που τα λατινοελληνικά αρχικά χρησιμοποιήθηκαν ως ένας βολικός τρόπος διακίνησης διαδικτυακής πληροφορίας σε μια εποχή όπου το πρότυπο κωδικοποίησης χαρακτήρων Unicode δεν υπήρχε (Karakos, 2002; Karakos, 2003), εξακολουθούν να χρησιμοποιούνται ευρέως από τους έλληνες χρήστες είτε από συνήθεια, είτε για λόγους συντομίας, είτε σε σπάνιες περιπτώσεις λόγω της χρήσης πεπαλαιωμένων υπολογιστών ή λειτουργικών συστημάτων όπου, δεν υπάρχουν διαθέσιμες ελληνικές γραμματοσειρές.

Η εταιρεία Innoetics (Τεχνοβλαστός του ΙΕΛ) εκτιμά ότι η χρήση των Greeklish ανά χώρα είναι αυτή του Σχήματος 1.



Σχήμα 1. Χρήση των greeklish ανά χώρα (http://services.innoetics.com/greeklish/Statistics.aspx)

Ένα βασικό πρόβλημα των Greeklish είναι η ασυνέπεια απεικόνισης των Ελληνικών χαρακτήρων με λατινικούς χαρακτήρες το οποίο οφείλεται στην έλλειψη τυποποίησης με αποτέλεσμα την πολλή μεγάλη ποικιλότητα αναπαράστασης. Το πλήθος των διαφορετικών τρόπων αναπαράστασης μίας λέξης αυξάνει γεωμετρικά όπως εμφανίζεται στο Σχήμα 2.



Σχήμα 2. Η ποικιλότητα αναπαράστασης (http://services.innoetics.com/greeklish/Statistics.aspx)

Η μετατροπή κειμένων από και προς την ελληνική γλώσσα γίνεται με τη διαδικασία της μεταγραφής (transliteration). Μεταγραφή είναι η πρακτική της αποτύπωσης μιας λέξης ή ενός γραπτού κειμένου από ένα σύστημα γραφής σε ένα άλλο τέτοιο σύστημα, καθώς και το σύνολο των κανόνων που διέπουν τη διαδικασία αυτή. Παρά την θέσπιση των προτύπων ISO 843 και ΕΛΟΤ 743 για τη μεταγραφή ελληνικών κειμένων από και προς λατινοελληνικά, η μεταγραφή των ελληνικών κειμένων γίνεται συνήθως με τρόπο ιδιοσυγκρασιακό με αποτέλεσμα τα λατινοελληνικά να χαρακτηρίζονται από ορθογραφική ποικιλότητα. (ISO 843, 1997; Marinis κ.α., 2007).

Άλλοτε ακολουθείται φωνητική μεταγραφή, με γνώμονα την αναπαράσταση φθόγγων και άρα απλοποίηση της ορθογραφίας και άλλοτε η ορθογραφική μεταγραφή, με γνώμονα την αναπαράσταση των ελληνικών φθόγγων στο μέτρο που αυτό είναι δυνατό με οπτικά

πανομοιότυπους λατινικούς χαρακτήρες ή ακόμη και μεταγραφή πληκτρολογίου, όπου η μετατροπή σε λατινοελληνικά γίνεται με χρήση ειδικών προγραμμάτων που μεταγράφουν το κείμενο με βάση την αντιστοίχιση του ελληνικού και του λατινικού γράμματος που βρίσκονται στο ίδιο πλήκτρο ενός πληκτρολογίου. (Greeklish, Wikipedia, the free encyclopedia; Tsourakis et al., 2007)

Για παράδειγμα, η φωνητική μεταγραφή χρησιμοποιεί το λατινικό i για τα ελληνικά ι, η, υ, οι, ει, το λατινικό ο για τους ελληνικούς φθόγγους ο και ω και το u για το δίψηφο φωνήεν ου. Η ορθογραφική μεταγραφή κάνει χρήση του λατινικού χαρακτήρα w για το ω ή των λατινικών χαρακτήρων ei για το δίψηφο φωνήεν ει καθώς και χρήση αριθμητικών χαρακτήρων σε αντικατάσταση γραμμάτων όπως 8 για το ελληνικό θ και 3 για το ελληνικό ξ. (Ανδρουτσόπουλος, 2001; Ανδρουτσόπουλος, 1999)

Πίνακας 1. Παραδείγματα μεταγραφής ελληνικών γραμμάτων

	η	**υ**	**ω**	**ου**	**θ**	**ξ**	**χ**	**ψ**
Φωνητική Μεταγραφή	i	i/u	O	u/ou	th	x/ks	ch/h/x	ps
Ορθογραφική Μεταγραφή	h/n	y/u	w/v	oy/ou	8/0	ks/3	x	ps
Μεταγραφή Πληκτρολογίου	h	Y	w/v	oy	u/q	j	x	c

Εργαλεία μεταγραφής

Η ευρύτατη χρήση των λατινοελληνικών από την ελληνική διαδικτυακή κοινότητα οδήγησε στην ανάπτυξη εργαλείων για εύκολη και γρήγορη μεταγραφή κειμένων από και προς τα λατινοελληνικά. Οι αρχές πάνω στις οποίες βασίζεται η λειτουργία των αλγορίθμων που χρησιμοποιούν τα εργαλεία αυτά πρέπει να είναι σωστά τεκμηριωμένες. Υπάρχουν διάφορα εργαλεία που έχουν αναπτυχθεί και πραγματοποιούν μεταγραφή από και προς τα λατινοελληνικά.

Το e-Chaos (Παράσχης, 2010), ο Μετατροπέας Χαρακτήρων (www.code.gr), το Greek Textbox Firefox extension (Κοσσυβάς, 2010) και διάφορες διαδικτυακές υπηρεσίες μετατροπής, όπως αυτές που παρέχονται από την ιστοσελίδα translatum.gr (www.translatum.gr Greeklish Converter), τον Αναπτυξιακό Σύνδεσμο Δυτικής Αθήνας (Αναπτυξιακός Σύνδεσμος Δυτικής Αθήνας, Μετατροπή ελληνικού κειμένου σε greeklish, http://home.asda.gr/active/GrLish2.asp), την εφαρμογή "Greeklish", που αναπτύχθηκε στο Δημοκρίτειο Πανεπιστήμιο Θράκης (Karakos, 2003) και το «Greeklish OUT» (Μάρκου, 2010), είναι υλοποιήσεις που προσεγγίζουν το πρόβλημα της μεταγραφής. Μια πιο εξελιγμένη και αρτιότερη από πλευράς ποιότητας των αποτελεσμάτων εφαρμογή, το «All Greek To Me» (Ινστιτούτο Επεξεργασίας του Λόγου (ΙΕΛ). "All Greek To Me!") έχει αναπτυχθεί από το Ινστιτούτο Επεξεργασίας του Λόγου και διατίθεται ως εμπορικό προϊόν από την εταιρεία Innoetics (http://services.innoetics.com/greeklish/Service.aspx). Πρόσφατα, η εταιρεία Google ανακοίνωσε μια νέα διαδικτυακή υπηρεσία, το πρόγραμμα Google Transliteration (http://www.google.com/transliterate/), το οποίο παρέχει τη δυνατότητα μετατροπής λατινικών χαρακτήρων στους φωνητικά ισοδύναμους χαρακτήρες της γλώσσας την οποία επιλέγει ο χρήστης μέσα από ένα πλήθος 22 διαθέσιμων γλωσσών. Η υπηρεσία αυτή χρησιμοποιεί τον κώδικα Unicode και είναι ήδη ενσωματωμένη σε αρκετές εφαρμογές παρέχοντας ένα API και διατίθεται ως bookmarklet για επέκταση και σε άλλες ιστοσελίδες.

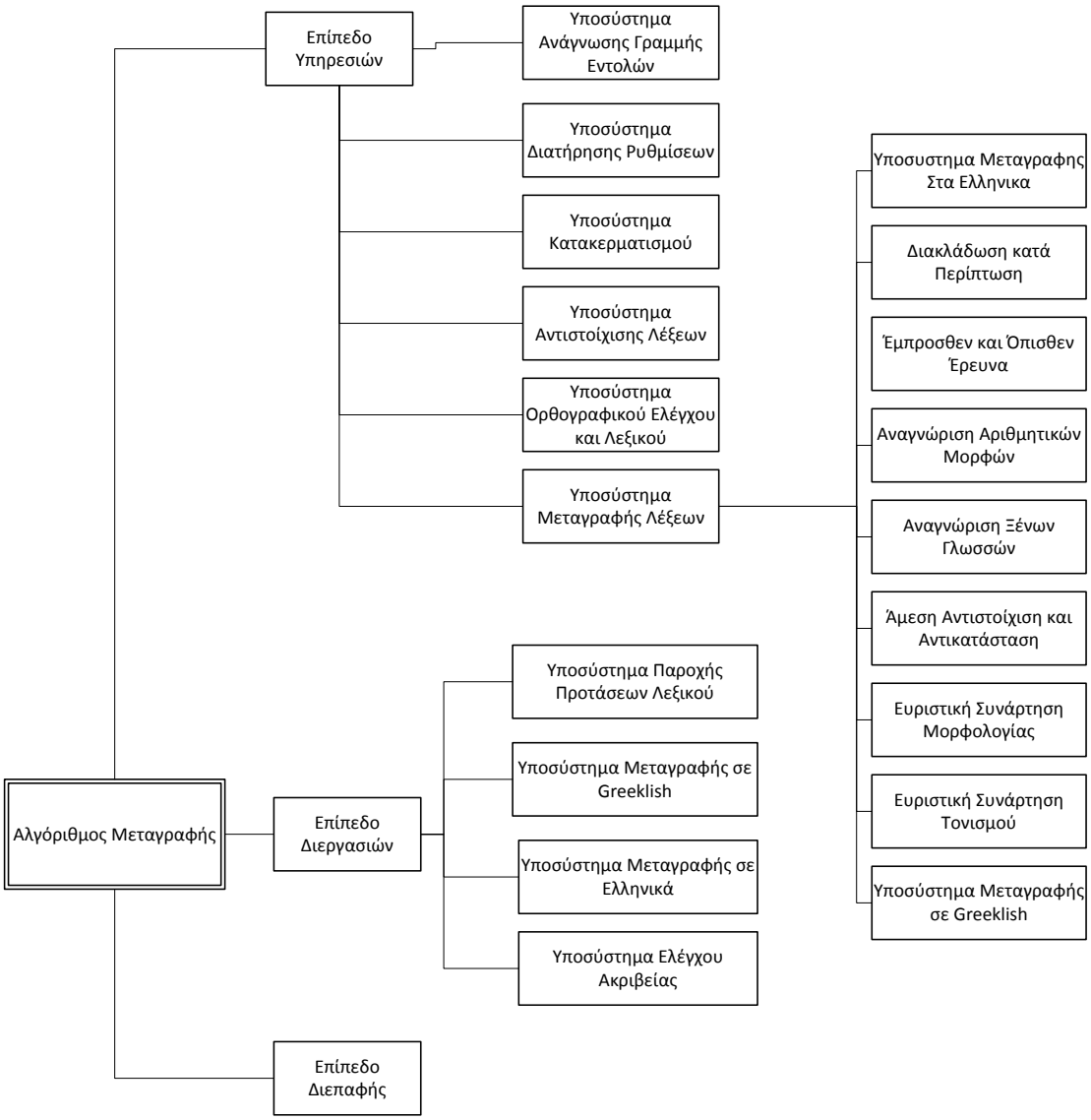
Η πρότασή μας

Οι βασικοί στόχοι του προτεινόμενου συστήματος ήταν η χρήση λεξικών της ελληνικής γλώσσας στη διαδικασία της μεταγραφής, η ευκολία στη χρήση του προγράμματος τόσο σε αλληλεπιδραστικά όσο και με αυτοματοποιημένα σενάρια χρήσης, η δυνατότητα λειτουργίας του με μηδενική πρότερη ρύθμιση από την πλευρά του τελικού χρήστη και η εύκολη κατανόηση και τροποποίησή του από οποιονδήποτε χρήστη στο μέλλον. Επίσης επιδιώχθηκε το τελικό αποτέλεσμα να είναι εύκολα παραμετροποιήσιμο και επεκτάσιμο και η συντήρησή του να μην παρουσιάζει ιδιαίτερα προβλήματα. (Παπαϊωάννου, 2007)

Ο σχεδιασμός του συστήματος βασίστηκε στο διαχωρισμό σε λογικά και λειτουργικά ανεξάρτητα συνθετικά δομικά μέρη, με σκοπό τη μείωση της πολυπλοκότητας. Το κάθε ένα από αυτά παρέχει σε όλα τα υπόλοιπα μέρη ένα σύνολο από εντολές που μπορεί να δεχτεί αλλά αποκρύπτει από αυτά τον τρόπο με τον οποίο πρόκειται να τις εκτελέσει. Σαν αποτέλεσμα υπάρχει πάντοτε η δυνατότητα της αλλαγής των εσωτερικών αλγορίθμων κάθε τμήματος χωρίς αυτή να επηρεάσει οποιοδήποτε άλλο τμήμα. Επιπρόσθετα επιδιώκεται η σαφής τεκμηρίωση των προϋποθέσεων που απαιτεί κάθε τμήμα να ισχύουν προτού εκτελέσει κάποια διεργασία, όπως επίσης και των δεσμεύσεων που αυτό αναλαμβάνει να σεβαστεί. Το κάθε τμήμα μπορεί να ελεγχθεί ως προς την ορθότητά του ανεξάρτητα από τα υπόλοιπα, με αποτέλεσμα να γίνεται ευκολότερος ο εντοπισμός σφαλμάτων και να μειώνεται συνολικά το κόστος και η δυσκολία συντήρησης της εφαρμογής.

Τελικά πραγματοποιείται ομαδοποίηση των δομικών μερών σε επίπεδα ανάλογα με το ρόλο του καθενός. Τα μέρη που βρίσκονται στο χαμηλότερο επίπεδο φέρνουν σε πέρας πολύ συγκεκριμένες λειτουργίες ενώ αυτά που βρίσκονται σε υψηλότερα επίπεδα χρησιμοποιούν τις υπηρεσίες των προηγουμένων για να επιτελέσουν πιο σύνθετο έργο. Με τον τρόπο αυτό επίσης μειώνεται η πολυπλοκότητα του συστήματος εφόσον η κατανόησή του μπορεί να γίνει σταδιακά ανεβαίνοντας από τα χαμηλότερα στα υψηλότερα επίπεδα. (Σχήμα 3).

Στον αλγόριθμο που αναπτύχθηκε τα δομικά μέρη ομαδοποιούνται σε τρία επίπεδα. Το επίπεδο υπηρεσιών στο οποίο πραγματοποιούνται διαδικασίες που αποτελούν μικρό τμήμα των εργασιών που πρέπει να φέρει σε πέρας το αμέσως επόμενο επίπεδο, το επίπεδο διεργασιών στο οποίο ολοκληρώνονται οι διεργασίες που πραγματοποιεί η εφαρμογή κατά τη βούληση του χρήστη και το επίπεδο διεπαφής στο οποίο πραγματοποιείται η ανταλλαγή πληροφοριών με το χρήστη.



Σχήμα 3. Τα δομικά μέρη του συστήματος

Ο αλγόριθμος μεταγραφής λέξης

Σύμφωνα με τον προτεινόμενο αλγόριθμο πραγματοποιείται αρχικά αναγνώριση και κατακερματισμός της συμβολοσειράς εισόδου. Κάθε απομονωμένο τμήμα της συμβολοσειράς εισόδου, το οποίο χαρακτηρίζεται ως «λέξη», μεταγράφεται ή όχι ύστερα από την κατάλληλη επεξεργασία.

Στη συνέχεια, επεξεργάζεται κάθε λέξη με μια διαδικασία η οποία αποτελείται από τα παρακάτω στάδια, στο τέλος των οποίων έχει παραχθεί η μετεγγραμμένη λέξη:

1. Αναγνώριση μορφής λέξης
2. Σχηματισμός μετεγγραμμένης μορφής
3. Πραγματοποίηση διορθώσεων

Στο πρώτο στάδιο γίνεται αναγνώριση της μορφής της λέξης με σκοπό την ταξινόμησή της σε ένα από τα σύνολα λέξεων τα οποία αναγνωρίζει ο αλγόριθμος. Τα σύνολα των λέξεων είναι:

1. Αριθμητικές μορφές: στο σύνολο αυτό ανήκουν λέξεις όπως «2os», «3h» οι οποίες έχουν συγκεκριμένη έννοια στην ελληνική γλώσσα.
2. Λέξεις προς άμεση αντικατάσταση: στο σύνολο αυτό ανήκουν όλες οι λέξεις οι οποίες είναι επιθυμητό είτε να παραμείνουν αναλλοίωτες είτε να μετεγγραφούν σε συγκεκριμένη μορφή (για παράδειγμα, ιδιωματικές εκφράσεις όπως «tpt»).
3. Μη-πραγματικές μορφές: στο σύνολο αυτό ανήκουν όλες οι λέξεις οι οποίες συχνά εμφανίζονται σε ηλεκτρονικά μέσα επικοινωνίας χωρίς όμως να ορίζονται σε κανένα λεξικό της ελληνικής (επιφωνήματα και άλλες παρόμοιες μορφές).
4. Ξενόγλωσσες λέξεις: στο σύνολο αυτό ανήκουν όλες οι λέξεις οι οποίες περιλαμβάνονται σε κάποιο από τα ξενόγλωσσα λεξικά της εφαρμογής.
5. Ελληνικές λέξεις: στο σύνολο αυτό ανήκουν όλες οι λέξεις οι οποίες δεν ανήκουν σε κανένα από τα παραπάνω σύνολα.

Για κάθε ένα από αυτά τα σύνολα υπάρχουν συγκεκριμένοι κανόνες βάσει των οποίων γίνεται αναγνώριση της λέξης ως μέρος του συνόλου και στη συνέχεια ακολουθείται διαφορετική σε κάθε περίπτωση διαδικασία για το σχηματισμό της μετεγγραμμένης μορφής από τη λέξη εισόδου. Η διαδικασία για όλα τα σύνολα πέραν αυτού των ελληνικών λέξεων μπορεί εύκολα να προκύψει λαμβάνοντας υπόψη τη μορφή που έχουν οι λέξεις του κάθε συνόλου. Στη συνέχεια θα περιγράψουμε εκτενέστερα τη διαδικασία μεταγραφής για τις ελληνικές λέξεις, η οποία έχει εξαιρετικά μεγάλη επιρροή τόσο στην ταχύτητα όσο και στην αποτελεσματικότητα του αλγορίθμου σαν σύνολο.

Η μεταγραφή ελληνικών λέξεων ξεκινά με το σχηματισμό όλων των πιθανών μορφών της βάσει των διαφορετικών τρόπων με τους οποίους μπορεί να μεταγραφεί κάθε γράμμα ή συνδυασμός γραμμάτων αυτής λαμβάνοντας υπόψη τόσο θεωρητικά (κανόνες γραμματικής) όσο και πειραματικά (το σύνολο των λέξεων που αναγνωρίζονται από το ελληνικό λεξικό του αλγορίθμου, μέσα στο οποίο συμπεριλαμβάνεται και μεγάλος αριθμός από κύρια ονόματα) δεδομένα, μια διαδικασία που περιγράφουμε ως «μεταγραφή με διακλαδώσεις». (Σχήμα 4).

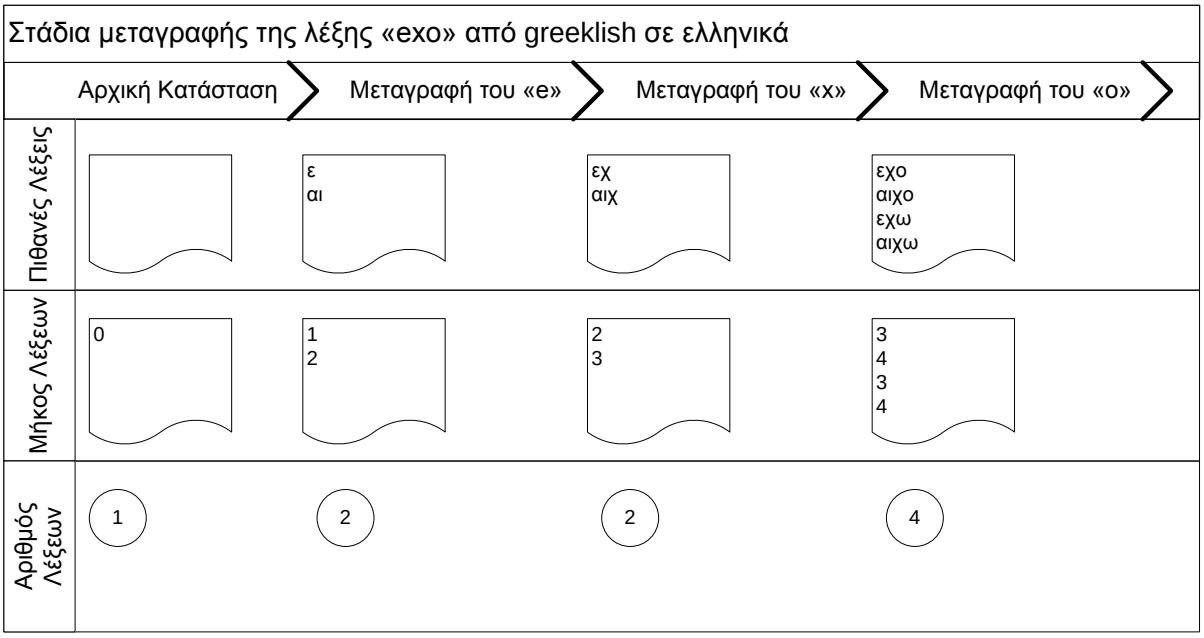
Στη συνέχεια, γίνεται επιλογή της επικρατέστερης μετεγγραμμένης μορφής μέσω έρευνας στο ελληνικό λεξικό άτονων λέξεων που περιλαμβάνει η εφαρμογή, από την οποία θα προκύψει η τελική μετεγγραμμένη μορφή. Σε περίπτωση που καμία από τις μορφές που έχουν προκύψει δεν αναγνωρίζεται από το λεξικό, επιλέγεται ως επικρατέστερη αυτή η οποία είναι πιθανότερο να αντιστοιχεί στη λέξη εισόδου βάσει γνωστών ιδιωμάτων των λατινοελληνικών.

Αν η επικρατέστερη μορφή έχει αναγνωριστεί ως γνωστή λέξη, στο επόμενο στάδιο γίνεται προσδιορισμός του τονούμενου φωνήεντος κάνοντας χρήση του ελληνικού λεξικού τονούμενων λέξεων με ειδικά επιλεγμένο τρόπο προκειμένου η ταχύτητα εκτέλεσης του αλγορίθμου να παραμένει υψηλή. Στην αντίθετη περίπτωση όπου η μορφή δεν έχει αναγνωριστεί ως λέξη (κάτι που εύκολα συμβαίνει σε περιπτώσεις λάθος πληκτρολογημένων λέξεων ή λέξεων με σημαντικά ορθογραφικά λάθη), ο αλγόριθμος δημιουργεί ένα σύνολο υποψήφιων μορφών που περιλαμβάνει παρόμοιες λέξεις επιλεγμένες μέσω των συνηθισμένων αλγορίθμων εύρεσης παρόμοιων λέξεων που υλοποιεί το ελληνικό λεξικό της εφαρμογής. Από το σύνολο αυτό αφαιρούνται λέξεις τις οποίες ο αλγόριθμος μεταγραφής θεωρεί λιγότερο πιθανές (για παράδειγμα, λέξεις που δεν τελειώνουν σε «ς» ενώ η λέξη

εισόδου τελειώνει σε «s») και κατόπιν επιλέγεται η περισσότερο όμοια από τις εναπομείναντες.

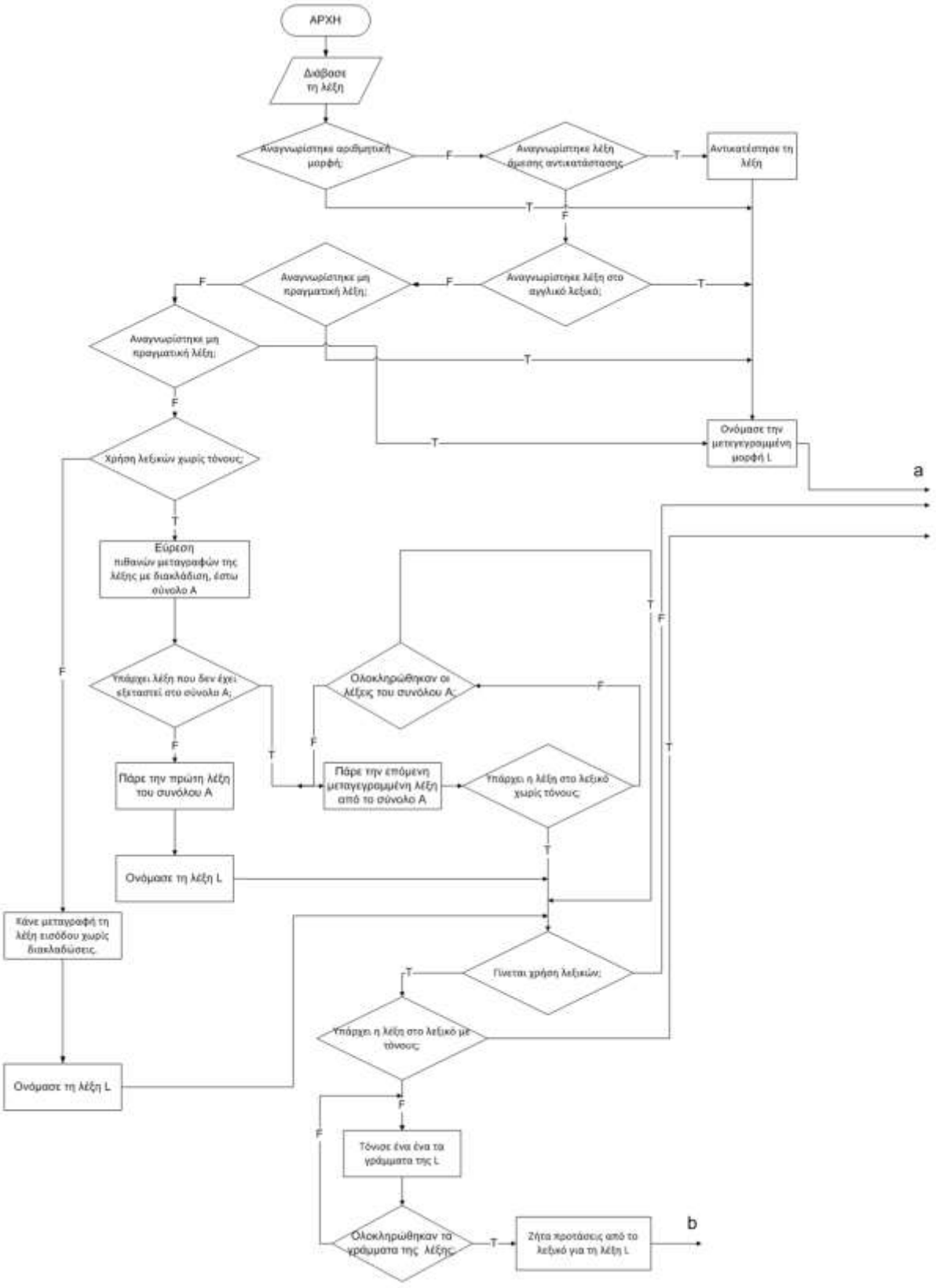
Στο τελικό στάδιο του αλγορίθμου πραγματοποιούνται διορθώσεις πάνω στη μετεγγραμμένη μορφή της λέξης προκειμένου η λέξη που τελικά θα προκύψει ως το αποτέλεσμα του αλγορίθμου να είναι πιο κοντά στα κοινώς αποδεκτά πρότυπα γραφής (για παράδειγμα, αν η λέξη αποτελείται εξ' ολοκλήρου από κεφαλαία γράμματα, αφαιρείται ο τυχόν τόνος από αυτή).

Ολοκληρώνοντας την περιγραφή του αλγορίθμου μεταγραφής είναι σημαντικό να αναφέρουμε ότι ο αλγόριθμος είναι δομημένος με τρόπο τέτοιο ώστε το σύνολο σχεδόν των δυνατοτήτων του μπορεί να ενεργοποιηθεί ή να απενεργοποιηθεί δυναμικά και κατ' επιλογή του χρήστη. Η χρήση αυτής της δυνατότητας έχει σαν αποτέλεσμα τη μείωση της αποτελεσματικότητας του αλγορίθμου και τη μεταβολή του χρόνου εκτέλεσης (τόσο αύξηση όσο και μείωση μπορεί να παρατηρηθούν, ανάλογα με τις συγκεκριμένες επιλογές που έγιναν), αλλά δίνει τη δυνατότητα στο χρήστη να αντλήσει πολύτιμα πειραματικά δεδομένα σχετικά με το πώς κάθε υποσύστημα του αλγορίθμου ή συνδυασμός αυτών επηρεάζει τελικά τη συνολική του αποτελεσματικότητα.

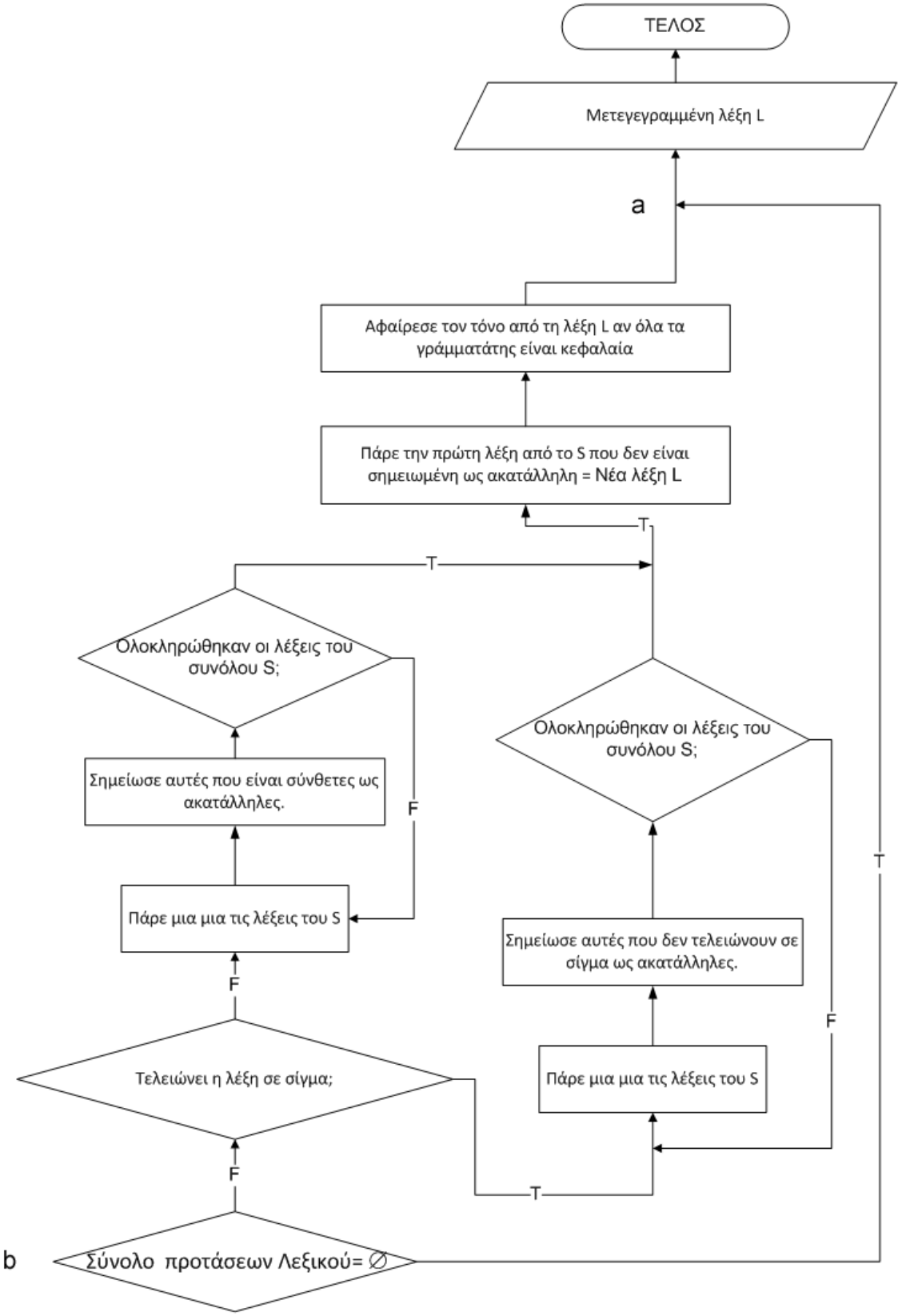


Σχήμα 4. Στάδια μεταγραφής από greeklish σε ελληνικά

Το διάγραμμα ροής του αλγόριθμου για τη μεταγραφή μιας λέξης λόγω της μεγάλης έκτασής του εμφανίζεται στα Σχήματα 5 και 6 με σημεία σύνδεσης τα γράμματα a και b.



Σχήμα 5. Το πρώτο μέρος του διαγράμματος ροής του αλγορίθμου μεταγραφής λέξης



Σχήμα 6. Το δεύτερο μέρος του διαγράμματος ροής του αλγορίθμου μεταγραφής λέξης

Χρήση λεξικών

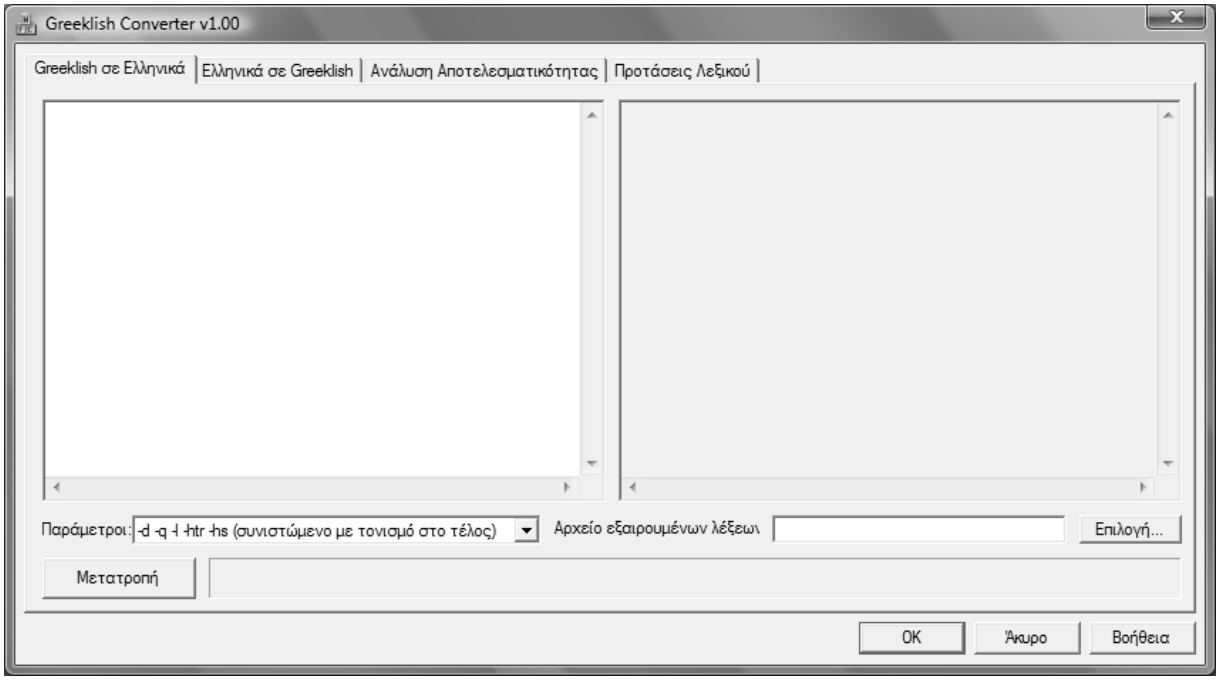
Για τη μεταγραφή από λατινοελληνικά σε ελληνικά ο αλγόριθμος βασίζεται στη χρήση λεξικών της ελληνικής γλώσσας. Η χρήση ορθογραφικών λεξικών προσφέρει εγγύηση αλάνθαστης μεταγραφής για λέξεις που είναι σωστά γραμμένες σε μορφή λατινοελληνικών, ή που μπορούν να μετασχηματιστούν σε σωστά γραμμένες μέσω προηγούμενων σταδίων του αλγορίθμου μεταγραφής. Επίσης μπορούν να επιτευχθούν υψηλά ποσοστά σωστής απόδοσης στη μεταγραφή καθώς οι βιβλιοθήκες λεξικών είναι σχεδιασμένες για να προσφέρουν ανταγωνιστική απόδοση ακόμα και σε πολύ μεγάλα σύνολα δεδομένων. Επιπλέον, η χρήση λεξικών επιτρέπει σε πολλές περιπτώσεις ακόμα και την αυτόματη διόρθωση τυπικών λαθών πληκτρολόγησης που συναντώνται σε πάρα πολλά κείμενα γραμμένα σε λατινοελληνικά. Το λεξικό που χρησιμοποιείται στον αλγόριθμο είναι το GNU Aspell (http://aspell.net/) ανοικτού κώδικα.

Το πρόγραμμα Greeklish Converter v1.00

Ο προτεινόμενος αλγόριθμος υλοποιήθηκε σε πρόγραμμα με τη χρήση της γλώσσας προγραμματισμού C++. Η μηχανή λεξικού GNU Aspell που χρησιμοποιήθηκε παρέχει δύο βασικές λειτουργίες που αξιοποιήθηκαν στην υλοποίηση του υποσυστήματος μεταγραφής, τη λειτουργία ελέγχου για την ύπαρξη μιας λέξης στο σύνολο του λεξικού και τη λειτουργία πρότασης λέξεων. Η ενσωμάτωσή του λεξικού στην εφαρμογή επιτεύχθηκε με την ανάπτυξη ενός μικρού επιπέδου αφαιρετικής πρόσβασης για να επιτευχθεί σύνδεση με το Aspell σύμφωνα με τα πρότυπα της γλώσσας C++.

Το πρόγραμμα Greeklish Converter v1.00 δημιουργήθηκε για χρήση στο γραφικό περιβάλλον των Microsoft Windows και είναι πλήρως εξελληνισμένο. Χρησιμοποιεί το σύνολο χαρακτήρων Unicode εξασφαλίζοντας τη σωστή απεικόνιση όλων των χαρακτηριστικών στην ελληνική γλώσσα και σε οποιοδήποτε έκδοση των Microsoft Windows, ανεξαρτήτως γλωσσικής έκδοσης, εφόσον η υποστήριξη της ελληνικής γλώσσας προσφέρεται ως βασικό χαρακτηριστικό του συστήματος.

Το πρόγραμμα Greeklish Converter v1.00 διατίθεται δωρεάν προς χρήση από κάθε ενδιαφερόμενο στη διεύθυνση: http://morphou.ee.duth.gr/~karakos/greeklish.html. Το διαθέσιμο αρχείο είναι σε συμπιεσμένη μορφή greeklish.rar, και μετά την αποσυμπίεσή του αρκεί να ενεργοποιηθεί το αρχείο με όνομα greeklish_gui.



Σχήμα 7. Το αρχικό περιβάλλον της εφαρμογής

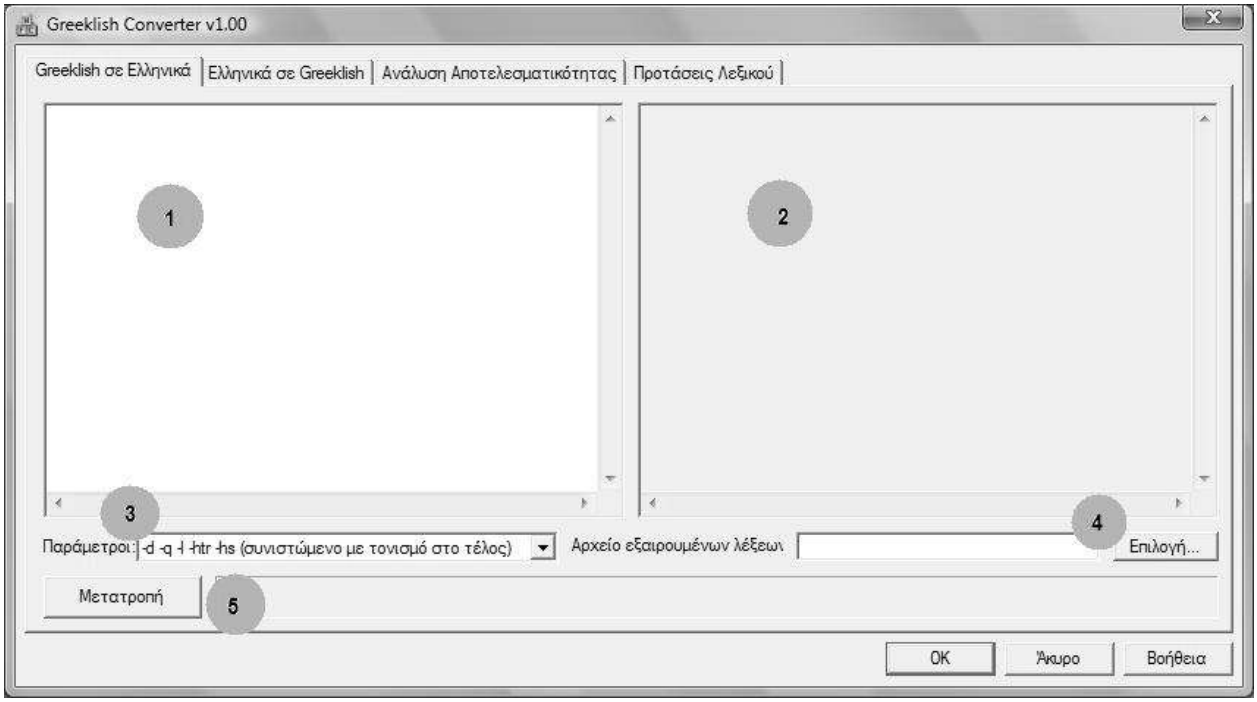
Το περιβάλλον της εφαρμογής

Κατά την εκκίνηση της εφαρμογής στην οθόνη εμφανίζεται το αρχικό περιβάλλον εργασίας. (Σχήμα 7). Ο χρήστης έχει τη δυνατότητα να επιλέξει μία από τις τέσσερεις καρτέλες που βρίσκονται στο πάνω μέρος του παραθύρου της εφαρμογής προκειμένου να μεταβεί στην οθόνη από όπου γίνονται οι ρυθμίσεις της λειτουργίας που επέλεξε και στην οποία εμφανίζονται τα αποτελέσματά της.

Οι διαθέσιμες λειτουργίες της εφαρμογής είναι μεταγραφή σε ελληνικά, μεταγραφή σε λατινοελληνικά, παροχή προτάσεων λεξικού, και έλεγχος ακριβείας.

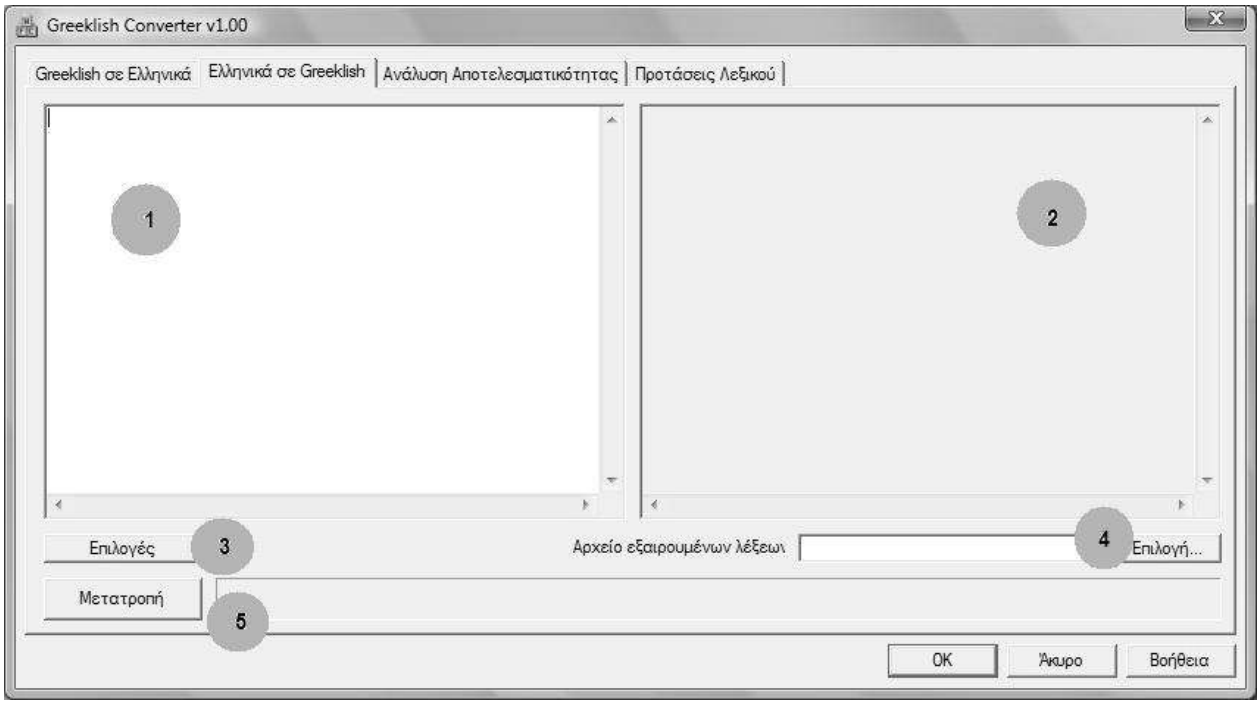
Μεταγραφή στα Ελληνικά

Η μεταγραφή στα ελληνικά είναι η λειτουργία που εμφανίζεται εξ' ορισμού στο χρήστη με την εκκίνηση της εφαρμογής. Η χρήση της είναι εξαιρετικά απλή, καθώς στο παράθυρο εμφανίζονται πέντε μόνο περιοχές. (Σχήμα 8).



Σχήμα 8. Οθόνη μεταγραφής σε ελληνικά

1. Πεδίο εισαγωγής λατινοελληνικού κειμένου (greeklish). Στο πεδίο αυτό πληκτρολογείται το κείμενο για μετατροπή σε ελληνικά.
2. Περιοχή παρουσίασης ελληνικού κειμένου. Στην περιοχή αυτή εμφανίζεται το ελληνικό κείμενο που προκύπτει από τη μεταγραφή.
3. Μενού επιλογής παραμέτρων. Στο μενού αυτό μπορούμε να επιλέξουμε εάν κατά τη διαδικασία της μεταγραφής είναι επιθυμητό οι λέξεις που μπορούν να τονιστούν σε πολλές συλλαβές να τονίζονται κατά προτίμηση στην προπαραλήγουσα ή στη λήγουσα.
4. Πεδίο επιλογής αρχείου εξαιρουμένων λέξεων. Στο πεδίο αυτό μπορεί να γίνει εισαγωγή του πλήρους ονόματος ενός αρχείου κειμένου και έτσι να εξαιρεθούν από τη διαδικασία της μεταγραφής οι λέξεις που υπάρχουν μέσα σε αυτό.
5. Κουμπί μετατροπής. Πιέζοντας το κουμπί αυτό ξεκινά η διαδικασία μεταγραφής, κατά τη διάρκεια της οποίας παρουσιάζεται η εξέλιξη της διαδικασίας στη μπάρα προόδου που βρίσκεται δεξιά του κουμπιού.



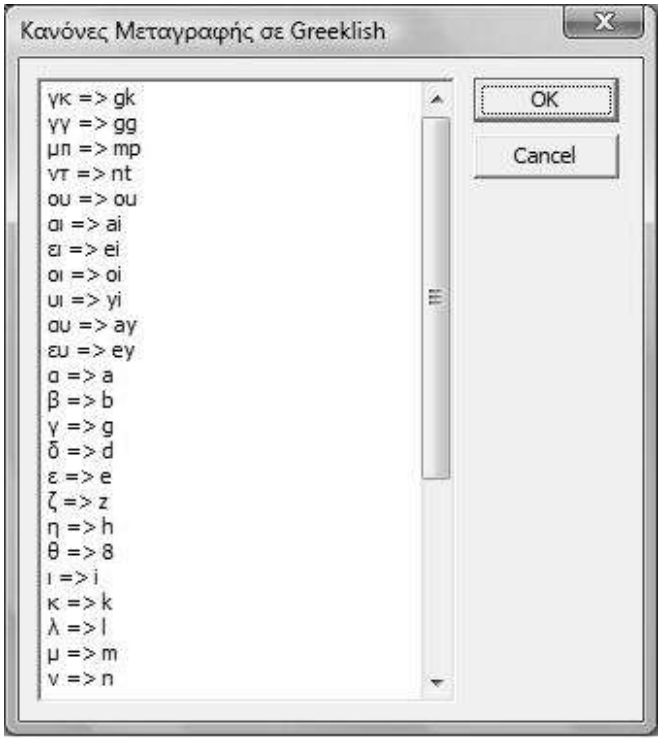
Σχήμα 9. Οθόνη μεταγραφής σε greeklish

Μεταγραφή στα Greeklish

Η οθόνη της λειτουργίας μεταγραφής σε greeklish παρουσιάζει ελάχιστες διαφορές με την οθόνη της λειτουργίας μεταγραφής στα ελληνικά. (Σχήμα 9)

Οι περιοχές 1, 2, 4 και 5 εκτελούν τις ίδιες ακριβώς λειτουργίες όπως και στην οθόνη μεταγραφής στα ελληνικά, με τη διαφορά ότι στη λειτουργία αυτή το κείμενο εισόδου είναι σε μορφή ελληνικών και το κείμενο εξόδου σε μορφή λατινοελληνικών (greeklish).

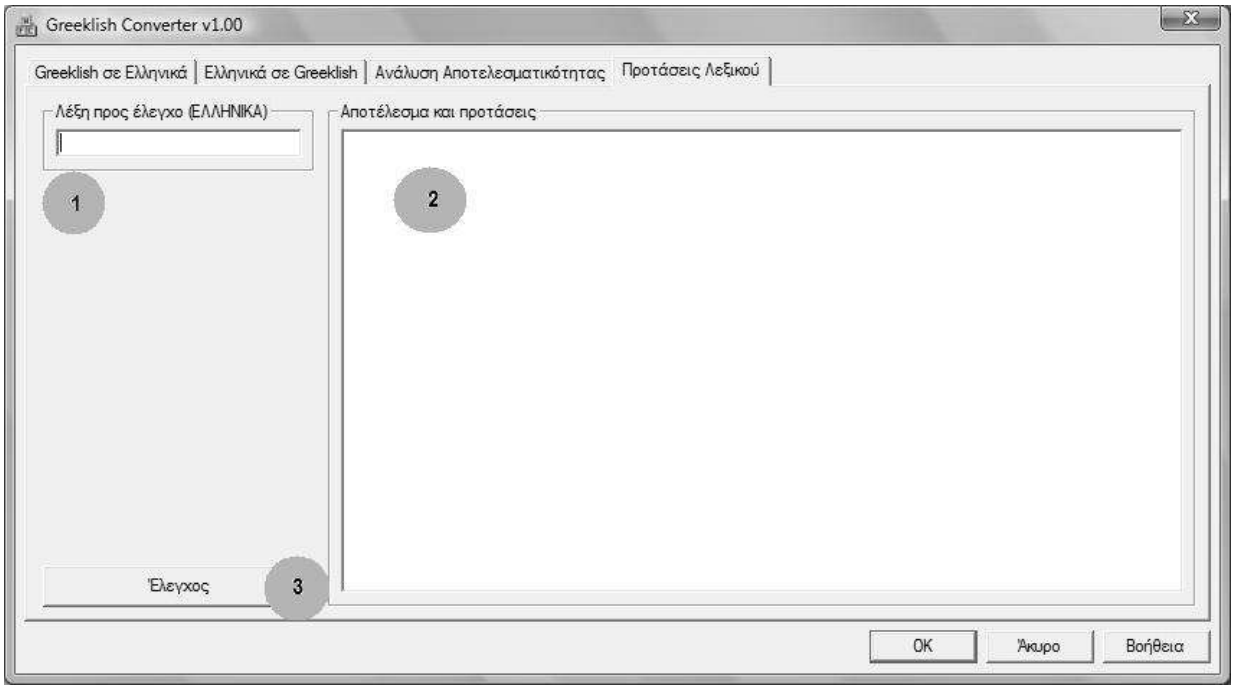
Το κουμπί Επιλογές που είναι σημειωμένο με τον αριθμό 3 επιτρέπει τον καθορισμό των κανόνων μεταγραφής που θα χρησιμοποιηθούν για τη μετατροπή σε λατινοελληνικά με τη βοήθεια ενός επιπλέον παραθύρου διαλόγου (Σχήμα 10). Ο χρήστης μπορεί είτε να διορθώσει δηλαδή, να αλλάξει ένα κανόνα (π.χ. το γράμμα ξ είναι δηλωμένο να μεταγράφεται ως 3 και μπορεί να γίνει ks) είτε να καθορίσει πέρα από τους δηλωμένους κανόνες του Σχήματος 10 και νέους κανόνες δικής του επιλογής, πληκτρολογώντας τους απλά στο τέλος της λίστας των κανόνων. Η μόνη προϋπόθεση είναι να συμπληρώνει ένα κανόνα σε κάθε γραμμή και στο τέλος να πατήσει το πλήκτρο OK. Η δυνατότητα αυτή δεν προσφέρεται από κανένα άλλο υπάρχων πρόγραμμα μεταγραφής.



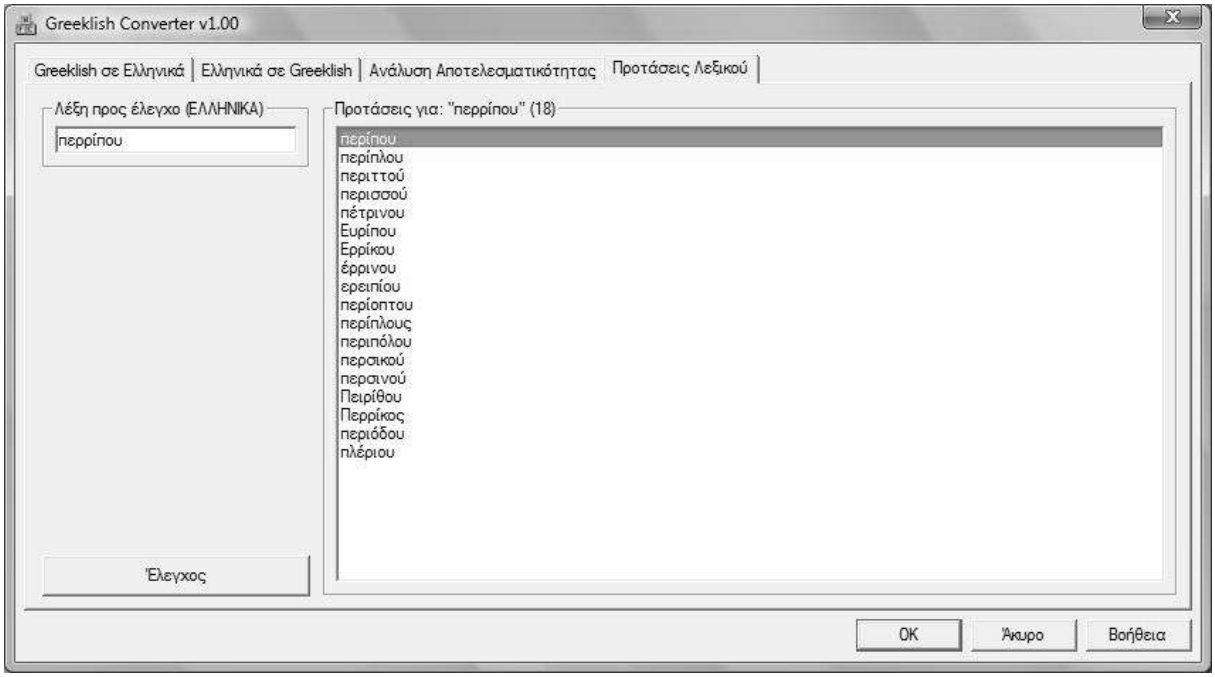
Σχήμα 10. Παράθυρο επιλογής κανόνων μεταγραφής σε greeklish

Παροχή Προτάσεων Λεξικού

Η οθόνη της λειτουργίας παροχής προτάσεων λεξικού είναι πολύ απλή στη χρήση της, καθώς περιέχει 3 μόνο σημεία αλληλεπίδρασης με το χρήστη. (Σχήμα 11)



Σχήμα 11. Οθόνη παροχής προτάσεων λεξικού



Σχήμα 12. Αποτελέσματα παροχής προτάσεων λεξικού

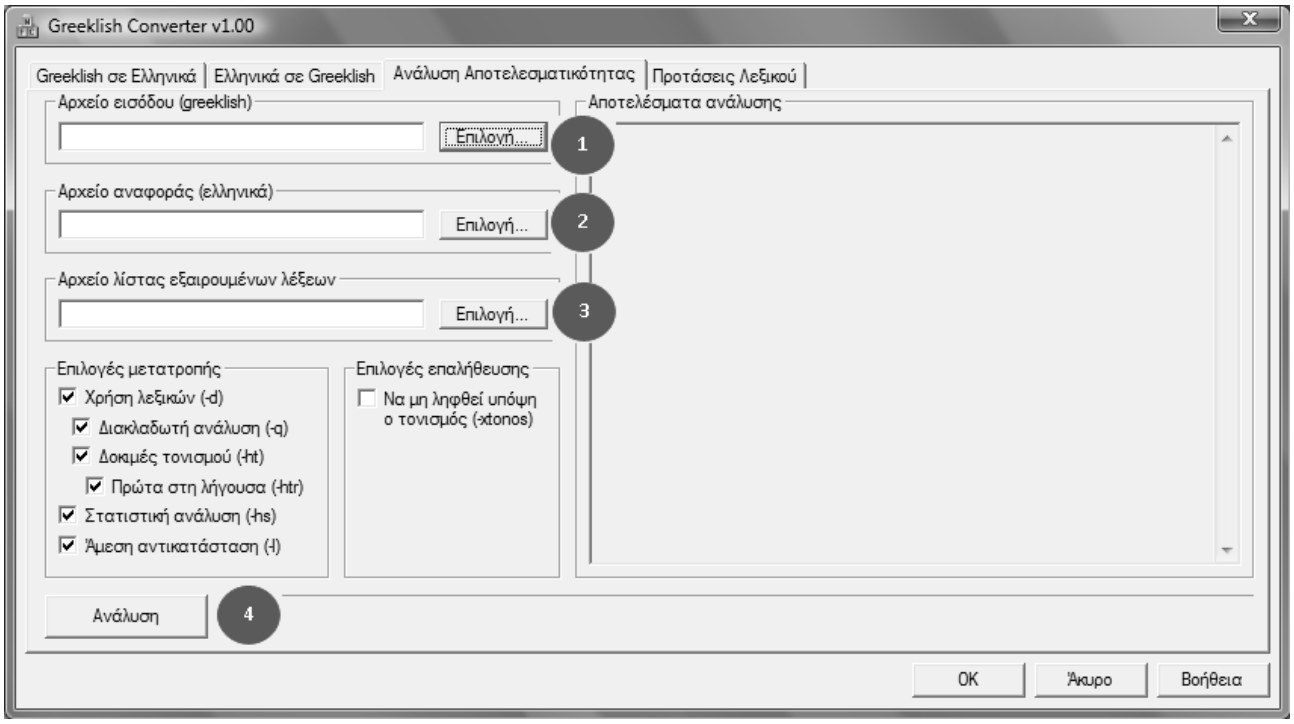
Οι περιοχές αλληλεπίδρασης με το χρήστη είναι:

1. Πεδίο εισαγωγής λέξης προς εξέταση όπου πληκτρολογείται η λέξη για την οποία επιθυμούμε να γίνει ορθογραφικός έλεγχος και να λάβουμε προτάσεις διόρθωσης από το λεξικό.
2. Περιοχή εμφάνισης αποτελεσμάτων. Στην περιοχή αυτή εμφανίζονται οι προτάσεις διόρθωσης που παρέχει το λεξικό εάν η λέξη προς εξέταση είναι άγνωστη.
3. Κουμπί εκτέλεσης λειτουργίας. Πιέζοντάς το εμφανίζονται οι προτάσεις του λεξικού για τη λέξη που έχουμε εισάγει.

Το αποτέλεσμα της λειτουργίας παροχής προτάσεων για μία άγνωστη λέξη εμφανίζεται στο Σχήμα 12.

Έλεγχος Ακριβείας

Η οθόνη της λειτουργίας ελέγχου ακριβείας είναι συγκριτικά η πιο σύνθετη από τις οθόνες των λειτουργιών που παρέχει η εφαρμογή. Παρά το γεγονός αυτό, η χρήση της δεν παρουσιάζει δυσκολίες καθώς εμφανίζονται σε αυτή μόνο 4 σημεία αλληλεπίδρασης με το χρήστη. (Σχήμα 13)



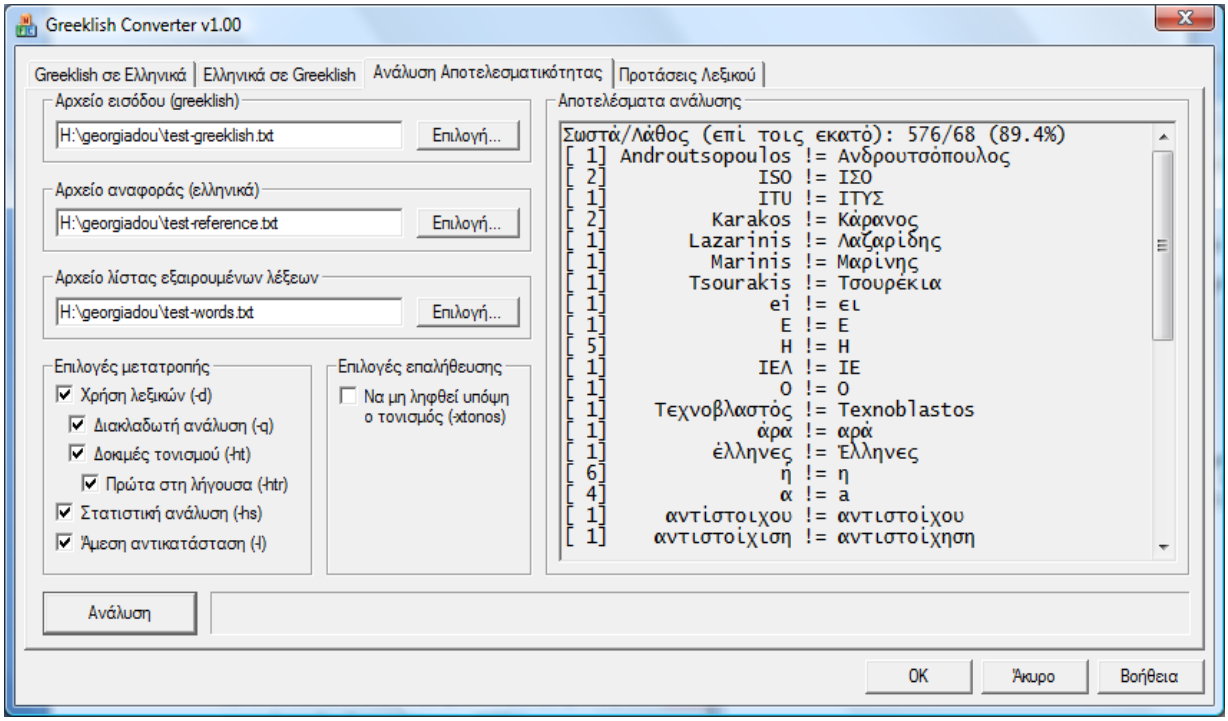
Σχήμα 13. Οθόνη ελέγχου ακριβείας της μεταγραφής

Οι περιοχές αλληλεπίδρασης με το χρήστη είναι :

1. Πεδίο εισαγωγής greeklish αρχείου εισόδου που θα μεταγραφεί. Μπορούμε είτε να πληκτρολογήσουμε το πλήρες όνομα του αρχείου απευθείας είτε να το επιλέξουμε μέσα από ένα παράθυρο διαλόγου χρησιμοποιώντας το κουμπί «Επιλογή…».
2. Πεδίο εισαγωγής του αρχείου αναφοράς δηλαδή, του αρχείου το οποίο περιέχει την ορθή μεταγραφή του κειμένου εισόδου. Μπορούμε είτε να πληκτρολογήσουμε το πλήρες όνομα του αρχείου απευθείας είτε να το επιλέξουμε μέσα από ένα παράθυρο διαλόγου χρησιμοποιώντας το κουμπί «Επιλογή…».
3. Πεδίο εισαγωγής αρχείου που περιέχει λίστα λέξεων οι οποίες επιθυμούμε να μη μεταγραφούν όπου εμφανίζονται στο κείμενο εισόδου αλλά να παραμείνουν ως έχουν. Μπορούμε είτε να πληκτρολογήσουμε το πλήρες όνομα του αρχείου απευθείας είτε να το επιλέξουμε μέσα από ένα παράθυρο διαλόγου χρησιμοποιώντας το κουμπί «Επιλογή…».
4. Κουμπί έναρξης. Πιέζοντας το κουμπί Ανάλυση ξεκινά η διαδικασία μεταγραφής και ελέγχου, κατά τη διάρκεια της οποίας παρουσιάζεται η εξέλιξη της διαδικασίας στη μπάρα προόδου που βρίσκεται δεξιά του κουμπιού.

Με τον έλεγχο ακρίβειας, μπορούμε να διαπιστώσουμε την ακρίβεια της μεταγραφής και στο παράθυρο «Αποτελέσματα ανάλυσης» εμφανίζονται το πλήθος των λέξεων του κειμένου οι οποίες δεν έχουν αποδοθεί ολόσωστα κατά τη μεταγραφή. Τα σχήματα 14 και 15 περιέχουν το αποτέλεσμα της ακρίβειας της μεταγραφής των δύο πρώτων κεφαλαίων αυτής της εργασίας. Το Σχήμα 14 περιέχει τον έλεγχο ακριβείας του κειμένου το οποίο δημιουργήθηκε από το πρόγραμμα Greeklish Converter v1.00, ενώ το Σχήμα 15 περιέχει τον έλεγχο ακριβείας του ιδίου κειμένου το οποίο όμως δημιουργήθηκε από το πρόγραμμα Greeklish to Greek!, το οποίο έχει αναπτυχθεί από το Ινστιτούτο Επεξεργασίας του Λόγου. Παρατηρούμε ότι το πρόγραμμα Greeklish Converter v1.00, είναι ελαφρώς καλλίτερο με ποσοστό 89,4% έναντι 88,9% του προγράμματος Greeklish to Greek!. Ο λόγος της μικρής

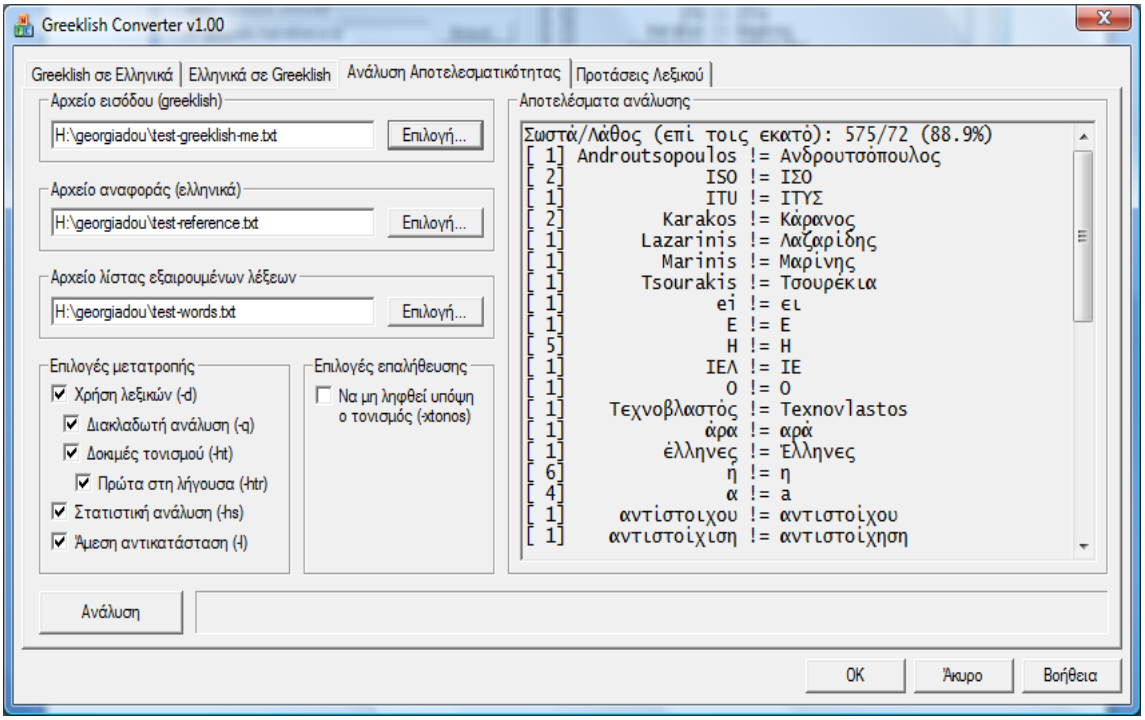
έκτασης του κειμένου ελέγχου (4213 χαρακτήρες ή 671 λέξεις) οφείλεται στο γεγονός ότι το πρόγραμμα Greeklish to Greek!, το οποίο διατίθεται δωρεάν στο διαδίκτυο επιτρέπει την μετατροπή κειμένου μέχρι 5000 χαρακτήρες ενώ το πλήρες πακέτο είναι εμπορικό προϊόν. Ο χρόνος εκτέλεσης και των δύο προγραμμάτων είναι σχεδόν παρόμοιος και διαρκεί περίπου ένα δευτερόλεπτο για κείμενο της τάξης των 5000 χαρακτήρων σε ένα σύγχρονο υπολογιστή.



Σχήμα 14. Οθόνη αποτελεσμάτων ελέγχου του προγράμματος Greeklish Converter v1.00

Συμπεράσματα

Το βασικότερο πρόβλημα των Greeklish είναι η ασυνέπεια απεικόνισης των Ελληνικών χαρακτήρων με λατινικούς χαρακτήρες με αποτέλεσμα τη μεγάλη ποικιλότητα αναπαράστασης. Για τη μεταγραφή λατινοελληνικών κειμένων στα ελληνικά, η χρήση ορθογραφικών λεξικών σαν μέρος της διαδικασίας μεταγραφής κρίθηκε απαραίτητη προκειμένου να επιτευχθεί το βέλτιστο αποτέλεσμα.



Σχήμα 15. Οθόνη αποτελεσμάτων ελέγχου του προγράμματος Greeklish to Greek!

Επειδή σε πολλές περιπτώσεις δεν είναι δυνατόν να γίνει σωστή μεταγραφή με απόλυτη βεβαιότητα, γίνεται κατανοητό ότι επιλέγοντας την πραγματοποίηση της συχνότερα σωστής μεταγραφής σε βάρος της πιο ακριβούς, ο αλγόριθμος μπορεί στατιστικά να βελτιώσει τα αποτελέσματά του. Ο ευκολότερος τρόπος να γίνει αυτό είναι να αφαιρεθούν από το λεξικό οι λέξεις οι οποίες είναι λιγότερο πιθανές να εμφανιστούν.

Επίσης, δεν αρκεί απλώς να χρησιμοποιείται κάποιο λεξικό της ελληνικής γλώσσας σαν διορθωτής των σφαλμάτων ενός αλγορίθμου μεταγραφής. Αυτό γίνεται άμεσα αντιληπτό εάν χρησιμοποιηθεί η λειτουργία ελέγχου ακριβείας της εφαρμογής με όλες τις επιλογές πλην αυτής της χρήσης λεξικών απενεργοποιημένες. Ο προτεινόμενος αλγόριθμος μεταγραφής μπορεί να κάνει χρήση εξελιγμένων τεχνικών για να πετύχει πολύ καλύτερα αποτελέσματα από το συνδυασμό ενός απλού αλγορίθμου και ενός ορθογραφικού διορθωτή, καταναλώνοντας μάλιστα πολύ λιγότερο χρόνο για την πραγματοποίηση της μεταγραφής.

Εκτός από τη χρήση τεχνικών βελτιστοποίησης της μεταγραφής, ένας άλλος τομέας που αποδεικνύεται πολύ σημαντικός είναι αυτός του κατακερματισμού της εισόδου. Ο αλγόριθμος κατακερματισμού μπορεί να αναγνωρίσει στο κείμενο εισόδου τμήματα που δεν αποτελούν λέξεις και έτσι να αποφευχθεί η μεταγραφή τους.

Επίσης, δεδομένου ότι οι αλγόριθμοι μεταγραφής δυσκολεύονται ή και αδυνατούν να αντιμετωπίσουν σωστά κάποιες κατηγορίες λέξεων, όπως είναι για παράδειγμα τα κύρια ονόματα σε ξένες γλώσσες, το πρόγραμμα Greeklish Converter v1.00 παρέχει τη δυνατότητα στο χρήστη να παρεμβαίνει στην εφαρμογή και να παρέχει επιπλέον πληροφορίες, προκειμένου να βελτιωθεί το αποτέλεσμα της μεταγραφής. Επιπλέον, παρέχει τη δυνατότητα στο χρήστη να ορίσει τις λέξεις οι οποίες δεν θα μεταγράφονται όπου και αν εντοπιστούν στο κείμενο εισόδου.

Τέλος, το πρόγραμμα Greeklish Converter v1.00 έχει βέλτιστη απόδοση λόγω της ενσωμάτωσης και χρήσης ορθογραφικών λεξικών και μπορεί να χρησιμοποιηθεί εύκολα σαν μια ανεξάρτητη εφαρμογή μετατροπής Greeklish στα ελληνικά και αντίστροφα από

οποιονδήποτε χρήστη δωρεάν, αρκεί να το «κατεβάσει» από το διαδίκτυο και να το εγκαταστήσει στον υπολογιστή του.

Αναφορές

Androutsopoulos, J. (2006). 'Greeklish': Transliteration practice and discourse in a setting of computer-mediated Digraphia. Forthcoming. In A Georgakopoulou & M Silk (eds.) Standard Languages and Language Standards: Greek, Past and Present.

ISO 843. (1997). Information and documentation – Conversion of Greek characters into Latin characters. International Organization for Standardization, Retrieved November, 2010, from http://www.iso.org

Karakos, A. (2003). Greeklish: An experimental interface for automatic transliteration. *Journal of the American Society for Information Science and Technology*, 54(11):1069-1074.

Koutsogiannis, D., & Mitsikopoulou, B. (2003). Greeklish and Greekness: Trends and Discourses of "Glocalness". *JCMC*, 9(1).

Lazarinis, F., & Vilares Ferro, J. & Tait, J. (2007). Improving Non-English Web Searching. Sigir 2007 Workshop Report**.**

Marinis, T., Papangeli, A., & Tseliga, T. (2007). "Potizo" or "Potizw"? The influence of morphology in the processing of Roman-alphabeted Greek. In E. Agathopoulou, M. Dimitrakopoulou, & D. Papadopoulou (eds.), *Selected Papers in Theoretical and Applied Linguistics* (pp. 443-452). Thessaloniki: Monochromia,.

Tsourakis, N., & Digalakis, V. (2007). "A generic methodology of converting transliterated text to phonetic strings case study: greeklish", In *INTERSPEECH-2007*, 1785-1788.

World Telecommunication/ICT Indicators Database, (2010). Retrieved November 30, 2010, from http://www.itu.int/ITU-D/ict/statistics/

Ανδρουτσόπουλος, Γ. Κ. (1998), 'Ορθογραφική ποικιλότητα στο ελληνικό ηλεκτρονικό ταχυδρομείο: μια πρώτη προσέγγιση', *Γλώσσα*, 46, 49-67.

Ανδρουτσόπουλος, Γ. Κ. (1999). Από τα Φραγκοχιώτικα στα greeklish, Ανακτήθηκε στις 30 Νοεμβρίου 2010 από http://www.tovima.gr/default.asp?pid=2&artid=114039&ct=114&dt=05/09/1999

Ανδρουτσόπουλος, Γ. Κ. (2000*)*. Λατινο-ελληνική ορθογραφία στο ηλεκτρονικό ταχυδρομείο: χρήση και στάσεις. *Studies in Greek Linguistics*, 20, 75-86.

Ανδρουτσόπουλος, Γ. Κ. (2001). Από dieuthinsi σε diey8ynsh. Ορθογραφική ποικιλότητα στη λατινική μεταγραφή των ελληνικών. Στο Α. Αγουράκη (επιμ.), *Πρακτικά 4ου Διεθνούς συνεδρίου Ελληνικής Γλωσσολογίας* (pp. 383-390). Θεσσαλονίκη: University studio press.

Καράκος, Α. (2002). Η ελληνική γραφή στα χρόνια του διαδικτύου. *Πληροφορική, – Τριμηνιαία έκδοση της Κυπριακής Εταιρείας Πληροφορικής*, 2, 22-24.

Κοσσυβάς, Φ. (2010). *«Greek Textbox» Firefox extension*. Ανακτήθηκε στις 30 Νοεμβρίου 2010 από https://addons.mozilla.org/en-US/firefox/addon/1193

Μάρκου, Ο (2010). *Greeklish OUT!. Υπηρεσία υπό ανάπτυξη*. Ανακτήθηκε στις 30 Νοεμβρίου 2010 από http://greeklishout.gr/

Παπαϊωάννου, Ι. (2007). Μελέτη και κατασκευή περιβάλλοντος μετατροπής χαρακτήρων για χρήση στο Internet. Δημοκρίτειο Πανεπιστήμιο Θράκης. Διπλωματική εργασία.

Παράσχης, Σ. (2010). *e-Chaos*. Ανακτήθηκε στις 30 Νοεμβρίου 2010 από http://www.paraschis.gr/files.php

Αναφορά στο άρθρο ως
Παπαϊωάννου, Ι., Καράκος, Α., Γεωργιάδου, Α. (2010). Greeklish Converter v1.00, ένα νέο περιβάλλον μετατροπής χαρακτήρων Greeklish. *Θέματα Επιστημών και Τεχνολογίας στην Εκπαίδευση*, 3(1), 49-67.
http://earthlab.uoi.gr/thete/index.php/thete