Δυσκολία κατανόησης του ξενόγλωσσου κειμένου και υφομετρία. Μια νέα προσέγγιση στην αναγνωσιμότητα κειμένων από έλληνες που μαθαίνουν την Ιταλική ως ξένη γλώσσα.

Γεώργιος Κ. Μικρός,
Επίκουρος Καθηγητής,
Τμήμα Ιταλικής και Ισπανικής Γλώσσας και Φιλολογίας

Περίληψη

Η παρούσα έρευνα χρησιμοποιεί ένα ευρύ φάσμα υφομετρικών μεταβλητών για να κατηγοριοποιήσει αυτόματα κείμενα ως προς την καταλληλότητα χρήσης τους στη διδασκαλία της Ιταλικής ως ξένης γλώσσας σε έλληνες μαθητές. Επιλέχθηκαν 4 φοιτητές ως κριτές και κατηγοριοποίησαν 450 ιταλικά κείμενα που συνέλεξαν από το Διαδίκτυο ως προς την καταλληλότητά τους για διδακτική χρήση με βάση την αντιλαμβανόμενη δυσκολία κατανόησής τους. Οι κρίσεις τους συσχετίστηκαν με τη χρήση πολυπαραγοντικού στατιστικού μοντέλου (λογιστική παλινδρόμηση) με μια σειρά από υφομετρικές μεταβλητές προερχόμενες από τις έρευνες της αναγνώρισης του συγγραφέα κειμένων αμφισβητούμενης πατρότητας. Τα αποτελέσματα έδειξαν ικανοποιητική ακρίβεια αυτόματης κατηγοριοποίησης με μέσο όρο 78,1%. Τα ποσοστά αυτά κρίνονται ικανοποιητικά αφού ένας τέτοιος αλγόριθμος θα μπορούσε να εξετάσει μεγάλο αριθμό κειμένων από το Διαδίκτυο και από τα 10 κείμενα που θα

σηματοδοτεί ως κατάλληλα για διδακτική αξιοποίηση τα 8 να είναι πράγματι αυτά που θα διάλεγε ένας καθηγητής ξένης γλώσσας.

1 Εισαγωγή[1]

Η καθιέρωση της επικοινωνιακής προσέγγισης στη διδασκαλία των ξένων γλωσσών άλλαξε ριζικά το διδακτικό υλικό και τη μεθοδολογία προσέγγισης του στην τάξη. Μέσα από την έμφαση στην μαθητοκεντρική δόμηση της διδακτικής ενότητας η επικοινωνιακή προσέγγιση ανέδειξε τη χρήση αυθεντικών κειμένων σε βασικό συντελεστή της αναδόμησης του γλωσσικού περιβάλλοντος της γλώσσας – στόχου [1, 47]. Παράλληλα τα τελευταία χρόνια η εκρηκτική διείσδυση των τεχνολογιών επικοινωνίας και ιδιαίτερα του Παγκόσμιου Ιστού (Web) επέτρεψε την ανεμπόδιστη παροχή αυθεντικών κειμένων σε πληθώρα γλωσσών σε μια ιδιαίτερα μεγάλη ποικιλία θεμάτων και κειμενικών γενών (text genres)[2].

Ο Παγκόσμιος Ιστός έτσι μετατράπηκε σε ένα τεράστιο αποθετήριο αυθεντικών ηλεκτρονικών κειμένων από όπου ο καθένας μπορεί να αντλήσει κειμενικό υλικό και να το εντάξει στην εκπαιδευτική διαδικασία. Αυτή η δυνατότητα αν και αρχικά αντιμετωπίστηκε με ενθουσιασμό από τους εκπαιδευτές στις ξένες γλώσσες, γρήγορα η τεράστια διαθέσιμη κειμενική ποικιλία προκάλεσε προβλήματα επιλογής. Αν και οι σύγχρονες

[1] Η παρούσα έρευνα χρηματοδοτήθηκε από τον Ειδικό Λογαριασμό Κονδυλίων Έρευνας του Πανεπιστημίου Αθηνών μέσω του ερευνητικού προγράμματος ΚΑΠΟΔΙΣΤΡΙΑΣ (2005) με κωδικό αριθμό 70/4/8871

[2] Η μηχανή αναζήτησης Google στις 03/08/2006 ανακοίνωσε ότι διαθέτει σε ηλεκτρονική μορφή κείμενα μεγέθους 1.024.908.267.229 λέξεων τα οποία έπειτα από επεξεργασία τα παραχωρεί στην ερευνητική κοινότητα [2].

μηχανές αναζήτησης μπορούν σχετικά εύκολα να προσδιορίσουν τη θεματική περιοχή ενός κειμένου και να το ανακτήσουν, ωστόσο δεν μπορούν να προσεγγίσουν την εκπαιδευτική καταλληλότητά του, απαραίτητη προϋπόθεση για την ενσωμάτωσή του σε ένα εκπαιδευτικό πρόγραμμα [3]. Ειδικότερα στον τομέα της διδασκαλίας των ξένων γλωσσών, η χρήση των μηχανών αναζήτησης μπορεί να προσφέρει χιλιάδες κείμενα τα οποία όμως παρατίθενται με κριτήριο τη θεματική συνάφεια ως προς τις λέξεις – κλειδιά που χρησιμοποιήθηκαν στην αναζήτηση και όχι τη δυσκολία του κειμένου ή την καταλληλότητά του ως προς συγκεκριμένο γλωσσικό επίπεδο. Η κρίση αυτή γίνεται από το διδάσκοντα ο οποίος θα πρέπει να διαβάσει αρκετές δεκάδες κείμενα στην οθόνη του υπολογιστή του[3] μέχρι να καταλήξει σε κάποιο που να ανταποκρίνεται στο γλωσσικό επίπεδο των μαθητών του, διαδικασία η οποία είναι κοπιώδης και χρονοβόρα [6]. Η αυτοματοποίηση αυτής της διαδικασίας θα μπορούσε να προσφέρει σημαντικές υπηρεσίες στην διδακτική της ξένης γλώσσας αφού θα περιόριζε σημαντικά το φόρτο προετοιμασίας του καθηγητή της ξένης γλώσσας. Επιπλέον θα συνέβαλε αποφασιστικά στην διεύρυνση της ποικιλίας ως προς τα κειμενικά γένη και θέματα τα οποία χρησιμοποιούνται στη γλωσσική διδασκαλία αφού θα ήταν διαθέσιμα πολλά κείμενα κατάλληλου γλωσσικού επιπέδου που θα κάλυπταν σε σημαντικό βαθμό τα

[3] Είναι ευρέως τεκμηριωμένο το γεγονός ότι η ταχύτητα ανάγνωσης στην οθόνη είναι 25-30% μικρότερη σε σχέση με το χαρτί [4, 457]. Ακόμα και όταν χρησιμοποιούνται οθόνες τεχνολογίας TFT, μια πρόσφατη έρευνα [5] έδειξε ότι το 50% των συμμετεχόντων παραπονέθηκαν ότι μεγαλύτερο πρόβλημα που αντιμετώπισαν διαβάζοντας στην οθόνη είναι η κούραση που τους προκαλεί στα μάτια. Ειδικότερα, κατά την διάρκεια γρήγορης επισκόπησης το 25% παραδέχτηκε ότι έχανε τις σειρές στην οθόνη και τους έπαιρνε πολύ χρόνο για να εντοπίσουν ξανά την λέξη ή τη σειρά στην οποία βρισκόντουσαν.

ενδιαφέροντα του μαθητικού κοινού. Η πραγμάτωση ενός τέτοιου στόχου σχετίζεται άμεσα με την υπολογιστική προσέγγιση στο θέμα της αναγνωσιμότητας και τον ποσοτικό υπολογισμό της με τον εντοπισμό και την καταμέτρηση κατάλληλων κειμενικών χαρακτηριστικών που μπορεί να χειριστεί ο Η/Υ.

2 Αναγνωσιμότητα κειμένων

Η αξιολόγηση της κειμενικής δυσκολίας υπήρξε ανέκαθεν πεδίο εντατικής έρευνας στην παιδαγωγική επιστήμη αφού σχετίζεται άμεσα με την αναγνωστική δεξιότητα και την επιτυχή ανάπτυξη του γραμματισμού. Αν και υπάρχουν πολλοί ορισμοί της αναγνωσιμότητας, στην παρούσα έρευνα θα υιοθετήσουμε αυτόν του George Klare ο οποίος την προσδιορίζει ως «η ευκολία της κατανόησης λόγω του ύφους της γραφής» [7]. Ο ορισμός αυτός, αν και δεν συμπεριλαμβάνει μεταβλητές που γνωρίζουμε ότι σχετίζονται άμεσα με την ευκολία κατανόησης του περιεχομένου, όπως τα κίνητρα και τα ενδιαφέροντα του αναγνώστη, η τυπογραφική μορφή του κειμένου κ.ά. ωστόσο επικεντρώνεται στην ερευνητική υπόθεση της παρούσας έρευνας, δηλαδή στη συμβολή του γλωσσικού ύφους στη διαμόρφωση της δυσκολίας κατανόησής του. Ο όρος «γλωσσικό ύφος» εδώ ακολουθεί την υφομετρική προσέγγιση και αναφέρεται στα γλωσσικά εκείνα χαρακτηριστικά που δεν αποτελούν συνειδητές επιλογές του συγγραφέα κατά τη διαδικασία της γραφής, όπως το μέγεθος των προτάσεων, το μέγεθος των λέξεων που χρησιμοποιεί, το ποσοστό των λέξεων που εμφανίζονται μία φορά μόνο στο κείμενο (άπαξ λεγόμενα) κ.ά. (βλ. και παρακάτω στο 3.2).

Η συσχέτιση γλωσσικού ύφους και αναγνωσιμότητας είχε ήδη παρατηρηθεί από τον 19ο αιώνα με την πρωτοποριακή για την εποχή του μελέτη του λογοτεχνικού ύφους [8] από τον καθηγητή Αγγλικής Λογοτεχνίας Lucius Adelno Sherman. Στον 20ο αιώνα μια σειρά από μελέτες καθιέρωσαν τη χρήση των υφομετρικών χαρακτηριστικών στη μελέτη της αναγνωσιμότητας με σημαντικότερη αυτή των Gray & Leary [9]. Μελέτησαν τη συσχέτιση 64 υφομετρικών μεταβλητών που μετρήθηκαν σε 48 κείμενα 100 λέξεων το καθένα και τα οποία αξιολογήθηκαν ως προς τη δυσκολία κατανόησης από 756 άτομα. Ακολουθώντας τη μέθοδο της πολλαπλής παλινδρόμηση (multiple regression) υπολόγισαν μια γραμμική εξίσωση η οποία συνέδεε τη δυσκολία του κειμένου με πέντε από τις σημαντικότερες μεταβλητές[4] πετυχαίνοντας δείκτη συνάφειας Pearson r 0,645. Αυτή ήταν και η αφετηρία μιας σειράς προσπαθειών η οποία συνεχίζεται μέχρι σήμερα για την εύρεση της καλύτερης εξίσωσης για την αξιολόγηση της δυσκολίας του κειμένου. Αν και οι εξισώσεις που έχουν προταθεί είναι πολυάριθμες, όλες μοιράζονται κάποια κοινά χαρακτηριστικά όσον αφορά τις μεταβλητές που χρησιμοποιούν αφού προσπαθούν να αναπαραστήσουν ποσοτικά τη συντακτική και την σημασιολογική πολυπλοκότητα του κειμένου. Παρακάτω περιγράφουμε τρεις από τις πιο γνωστές και χρησιμοποιημένες εξισώσεις που έχουν δημοσιευτεί για την Αγγλική γλώσσα:

[4] Οι μεταβλητές που χρησιμοποιήθηκαν ήταν: Μέσος όρος του μήκους των προτάσεων (μετρημένο σε λέξεις), Αριθμός διαφορετικών «δύσκολων» λέξεων, λέξεων δηλαδή που δεν ανήκουν στην λίστα των 769 πιο κοινόχρηστων λέξεων της Αγγλικής που δημοσίευσε ο Dale, Αριθμός αντωνυμιών α, β και γ προσώπου, Ποσοστό διαφορετικών λέξεων, Αριθμός προθετικών φράσεων.

1. Η εξίσωση Lorge [10] η οποία επαναδιατυπώθηκε το 1948 και κατηγοριοποιεί τα κείμενα σε τάξεις του εκπαιδευτικού συστήματος των Η.Π.Α. (από $3^{η}$ έως $12^{η}$).

   *Τάξη = 0,07*sl + 0,1073*wd +0,1301*pp + 1,6126* όπου

   - sl = Μέσο μήκος πρότασης μετρημένο σε λέξεις
   - wd = Αριθμός «δύσκολων» λέξεων ανά 100 λέξεις κειμένου
   - pp = Αριθμός προθετικών φράσεων ανά 100 λέξεις κειμένου

2. Η εξίσωση Dale-Challs [11] η οποία επαναδιατυπώθηκε το 1995, και κατηγοριοποιεί τα κείμενα σε τάξεις του εκπαιδευτικού συστήματος των Η.Π.Α. (από $3^{η}$ έως $12^{η}$).

   *Τάξη = 0,0596*sl + 0,1579*wd + 3,6365* όπου

   - sl = Μέσο μήκος πρότασης μετρημένο σε λέξεις
   - wd = Το ποσοστό των λέξεων που δεν βρίσκεται στη λίστα των 3000 πιο κοινόχρηστων λέξεων που δημοσίευσε ο Dale.

3. Η εξίσωση της αναγνωστικής ευκολίας του Flesch (Flesch Reading Ease) [12] η οποία έχει επαναδιατυπωθεί αρκετές φορές και υπολογίζει έναν δείκτη από το 0 έως το 100. Όσο περισσότερο πλησιάζει το 0 τόσο δυσκολότερο είναι το κείμενο.

   *Αναγνωστική Ευκολία = 206,835 − 1,015*sl − 0,846*wl* όπου

   - sl = Μέσο μήκος πρότασης μετρημένο σε λέξεις
   - wl = Αριθμός συλλαβών ανά 100 λέξεις κειμένου

   Αποτελεί μία από τις πιο δημοφιλείς εξισώσεις αναγνωσιμότητας και έχει χρησιμοποιηθεί ευρύτατα στις Η.Π.Α. Μια τροποποιημένη εκδοχή αυτής της εξίσωσης, η Flesch-Kincaid [13], χρησιμοποιείται για να κατηγοριοποιήσει τα κείμενα σε τάξεις του εκπαιδευτικού συστήματος των Η.Π.Α.

Οι εξισώσεις που προαναφέρθηκαν είναι υπολογισμένες για την Αγγλική γλώσσα και δεν μπορούν να εφαρμοστούν σε άλλες γλώσσες. Στην Ιταλική γλώσσα η οποία θα μας απασχολήσει στην παρούσα έρευνα έχουν υπολογιστεί δύο εξισώσεις οι οποίες αναλύονται παρακάτω:

1. Η μετατροπή της εξίσωσης της αναγνωστικής ευκολίας Flesch Reading Ease στην Ιταλική γλώσσα από τον Roberto Vacca γνωστή και ως Flesch – Vacca.

   $$Αναγνωστική\ Ευκολία = 206 – 0,65*wl – sl$$ όπου

   - sl = Μέσο μήκος πρότασης μετρημένο σε λέξεις
   - wl = Αριθμός συλλαβών ανά 100 λέξεις κειμένου

2. Η εξίσωση GULPEASE αναπτύχθηκε ειδικά για την Ιταλική γλώσσα από την ομάδα Gruppo Universitario Linguistico Pedagogico (GULP), στο Ινστιτούτο Φιλοσοφίας του Πανεπιστημίου «La Sapienza» της Ρώμης [14, 15].

   $$Αναγνωστική\ ευκολία = 89 – LP/10 + FR*3$$ όπου

   - LP = Ο συνολικός αριθμός των χαρακτήρων του κειμένου επί 100 διά το συνολικό αριθμό των λέξεων του κειμένου
   - FR = Ο αριθμός προτάσεων του κειμένου επί 100 διά το συνολικό αριθμό των λέξεων του κειμένου

   Ο δείκτης GULPEASE παίρνει τιμές από 0 έως 100 και όσο πλησιάζει το 0 το κείμενο γίνεται δυσκολότερο. Επειδή ο υπολογισμός του βασίζεται στον αριθμό γραμμάτων και όχι των συλλαβών στη λέξη, μπορεί να υπολογιστεί πιο εύκολα[5] και αυτή τη

[5] Ο υπολογισμός του δείκτη μπορεί να γίνει και στη σελίδα: http://xoomer.alice.it/roberto-ricci/variabilialeatorie/esperimenti/leggibilita.htm

στιγμή θεωρείται ο πιο αξιόπιστος δείκτης αναγνωσιμότητας της Ιταλικής γλώσσας [16].

Για τις ανάγκες της παρούσας έρευνας θα υπολογίσουμε το συγκεκριμένο δείκτη και θα συγκρίνουμε την απόδοσή του με την προτεινόμενη μεθοδολογία.

3 Κειμενική αναπαράσταση και γλωσσικά χαρακτηριστικά

3.1 Αναπαράσταση «Κείμενο – Γλωσσικά Χαρακτηριστικά»

Η εξελίξεις στην ποσοτική επεξεργασία Ηλεκτρονικών Σωμάτων Κειμένων (Corpora) με τη συνδρομή κλάδων όπως η Μηχανική Μάθηση (Machine Learning), η Ανάκτηση Πληροφορίας (Information Retrieval) και η Υφομετρία (Stylometry) μας δίνουν σήμερα σημαντικές δυνατότητες στο φιλτράρισμα μεγάλων συλλογών κειμένων και την κατηγοριοποίησή τους βάσει διαφόρων κριτηρίων όπως το θέμα [17], το κειμενικό γένος [18] και ο συγγραφέας [19] .

Η βασική μεθοδολογία που χρησιμοποιείται είναι αυτή της αναπαράστασης «Κείμενο – Γλωσσικά Χαρακτηριστικά» (document-feature representation) δηλαδή της αποτύπωσης κάθε κειμένου ως ακολουθίας γλωσσικών χαρακτηριστικών τα οποία εκφράζουν με ποσοτικό τρόπο την παρουσία τους σε αυτό. Συνήθως, τέτοιες συλλογές κειμένων περιέχουν κωδικοποιημένη τη σωστή κατηγορία (θέμα, γένος, συγγραφέας) που ανήκει το κάθε κείμενο. Με αυτή την πληροφορία εκπαιδεύονται στατιστικοί αλγόριθμοι μηχανικής μάθησης όπως τα Νευρωνικά Δίκτυα (Neural Networks), οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines), τα Δένδρα Αποφάσεων (Decision Trees) κ.ά. οι οποίοι

αναπτύσσουν έναν ταξινομητή (classifier) που μπορεί να κατηγοριοποιήσει κάθε νέο κείμενο που θα του παρουσιαστεί στη σωστή κατηγορία.

Η μέθοδος αυτή είναι εξαιρετικά ευέλικτη γιατί επιτρέπει τη χρήση διαφόρων χαρακτηριστικών αρκεί αυτά να μπορούν να μετρηθούν σε ένα κείμενο. Επίσης, στα ίδια δεδομένα μπορεί να γίνει παράλληλα εκπαίδευση πολλών ταξινομητών[6] και να επιλεγεί αυτός που αποδίδει καλύτερα την κατηγορία του κειμένου.

Στην παρούσα εργασία θα επεκτείνουμε τη χρήση της αναπαράστασης «Κείμενο – Γλωσσικά Χαρακτηριστικά» κατά τέτοιο τρόπο ώστε η κατηγορία στην οποία θα εντάσσεται ένα κείμενο, θα είναι ο βαθμός της δυσκολίας κατανόησης που παρουσιάζει από κάποιον που δεν είναι μητρικός ομιλητής της Ιταλικής αλλά την σπουδάζει ως ξένη γλώσσα.

3.2 Υφομετρικά γλωσσικά χαρακτηριστικά

Η υφομετρία ως κλάδος ασχολείται με την ποσοτική επεξεργασία των ύφους των γραπτών κειμένων με βασικότερη εφαρμογή τον εντοπισμό της πατρότητας ενός κειμένου όταν αυτός διεκδικείται από δύο ή περισσότερους πιθανούς συγγραφείς. Ήδη από το 1439 ο Lorenzo Valla, Ιταλός ουμανιστής, ρήτορας και εκπαιδευτής συνέγραψε το *De falso credita et ementita Constantini Donatione declamatio* στο οποίο εξετάζοντας τα υφολογικά χαρακτηριστικά της γλώσσας του κειμένου *Constitutum Constantini,* έδειξε ότι αυτό δεν μπορούσε να έχει γραφεί την περίοδο του Μεγάλου Κωνσταντίνου (4^{ος} αιώνας μ.Χ.), αλλά πολύ αργότερα, τον 8^{ο}

[6] Για την σύγκριση των περισσότερων αλγόριθμων μηχανικής μάθησης βλ. [20] και το σχετικό λογισμικό WEKA που έχουν αναπτύξει: http://www.cs.waikato.ac.nz/ml/weka/.

αιώνα μ.Χ. Η υφομετρία στα μέσα του 20ου αιώνα με τη συνδρομή των Η/Υ έγινε ευρύτερα γνωστή με τις προσπάθειες που καταβλήθηκαν για να διερευνηθεί η πατρότητα των επιστολών του Αποστόλου Παύλου [21] και ορισμένων έργων του Σαίξπηρ [22]. Σήμερα, η υφομετρία χρησιμοποιεί μεγάλος εύρος στατιστικών τεχνικών επιτυγχάνοντας υψηλά επίπεδα ταυτοποίησης του συγγραφέα ενός κειμένου. Η αξιοπιστία της μεθόδου είναι πλέον υψηλή και ήδη έχει αρχίσει να χρησιμοποιείται επικουρικά μαζί με άλλες τεχνικές σε κρίσιμους κλάδους όπως είναι η εγκληματολογική γλωσσολογία (forensic linguistics) [23].

Τα γλωσσικά χαρακτηριστικά τα οποία κατά καιρούς έχουν μετρηθεί είναι πολυάριθμα[7] και ανήκουν σε όλο το φάσμα των γλωσσικών επιπέδων. Αν και αυτά τα χαρακτηριστικά έχουν χρησιμοποιηθεί με έμφαση στον εντοπισμό του συγγραφικού ύφους, ωστόσο μέσα από σειρά ερευνών [25, 26, 27, 28] έχει φανεί ότι σχετίζονται και με άλλα μετακειμενικά χαρακτηριστικά, όπως είναι το θέμα. Αυτή η πολυεπίπεδη συσχέτιση των υφομετρικών χαρακτηριστικών με διαφορετικές κειμενικές λειτουργίες μας επιτρέπει να τα δοκιμάσουμε και στη μέτρηση της αναγνωσιμότητας. Δύο από τα πιο γνωστά υφομετρικά χαρακτηριστικά, το μέσο μήκος λέξης και το μέσο μήκος πρότασης αποτελούν τη βάση των γνωστότερων τύπων αναγνωσιμότητας όπως είδαμε και στην ενότητα 2, γεγονός που δείχνει ότι πιθανόν και άλλοι δείκτες μπορούν αξιοποιηθούν στην εκτίμηση της αναγνωστικής δυσκολίας. Στην παρούσα έρευνα μετρήσαμε τα ακόλουθα υφομετρικά χαρακτηριστικά:

[7] Ο Rudman [24, 360] υπολογίζει ότι μέχρι σήμερα έχουν χρησιμοποιηθεί πάνω από 1000 γλωσσικά χαρακτηριστικά ως υφομετρικές μεταβλητές.

- Words (Σύνολο λέξεων): Το συνολικό μέγεθος του κειμένου σε λέξεις.
- Type/token ratio (TTR): Ο λόγος του αριθμού των μοναδιαίων λεξιλογικών μονάδων (types) προς τον αριθμό των λέξεων (tokens) του κειμένου. Όσο μεγαλύτερος είναι ο δείκτης τόσο «πλουσιότερο» είναι λεξιλογικά το κείμενο.
- AWL (Μέσο Μήκος Λέξης): Υπολογίζεται ο μέσος όρος του μήκους των λέξεων του κάθε κειμένου με βασική μονάδα μέτρησης τον χαρακτήρα.
- WLsd (Τυπική Απόκλιση του Μέσου Μήκους Λέξης): Η τυπική απόκλιση του μέσου όρου του μήκους των λέξεων του κειμένου.
- Sentence length (Μέσο Μήκος της Πρότασης): Ο μέσος όρος του μήκους των προτάσεων του κειμένου μετρημένο σε λέξεις.
- SLsd (Τυπική Απόκλιση του Μέσου Μήκους της Πρότασης): Η τυπική απόκλιση του μέσου μήκους των προτάσεων του κάθε κειμένου.
- Perc_HapL (Άπαξ Λεγόμενα): Υπολογίζεται το ποσοστό των λέξεων που εμφανίζονται στο κείμενο με συχνότητα 1.
- Perc_DisL (Δις Λεγόμενα): Υπολογίζεται το ποσοστό των λέξεων που εμφανίζονται στο κείμενο με συχνότητα 2.
- Dis_HapL (Λόγος Δις προς Άπαξ Λεγόμενα): Ο λόγος αυτός έχει προταθεί στη βιβλιογραφία ως ενδεικτικός του συγγραφικού ύφους [29].
- LD (Λεξιλογική Πυκνότητα): Ο λόγος του ποσοστού των λέξεων «περιεχομένου» (content words) προς το ποσοστό των λειτουργικών λέξεων (function words). Ο όρος «Λεξιλογική Πυκνότητα» (Lexical

Density) χρησιμοποιείται στη θεωρία της λειτουργικής γραμματικής του Halliday με ελαφρά διαφορετικό τρόπο υπολογισμού. Εδώ ακολουθούμε την έκδοση που χρησιμοποιείται αρκετά συχνά σε μελέτες εντοπισμού αγνώστου συγγραφέα και απαντάται συχνά και με τον όρο «Λειτουργική Πυκνότητα» (Functional Density) [30]. Όσο μεγαλύτερος είναι ο δείκτης τόσο μεγαλύτερο τμήμα του κειμένου αποτελείται από λέξεις περιεχομένου.

- Yule's K: Υπολογίζεται ο δείκτης K ο οποίος αποτελεί τον πιο αξιόπιστο δείκτη λεξιλογικού «πλούτου» στη σχετική βιβλιογραφία. Ο υπολογισμός του ακολουθεί τον τύπο των Tweedie & Baayen [31, 330]. Όσο πιο μικρός είναι ο δείκτης τόσο μεγαλύτερη λεξιλογική ποικιλία παρουσιάζει το κείμενο.
- Entropy (Εντροπία): Υπολογίζεται η εντροπία του κάθε κειμένου, ο βαθμός δηλαδή της οργάνωσης και της προβλεψιμότητας των λεξικών συχνοτήτων. Ο υπολογισμός της ακολουθεί τον τύπο του Oakes [32, 61]. Όσο μεγαλύτερη είναι η τιμή της εντροπίας τόσο λιγότερο αναμενόμενες, με τη στατιστική σημασία του όρου, είναι οι λέξεις που προκύπτουν στο κείμενο.
- Relative Entropy (Σχετική Εντροπία): Ο λόγος της θεωρητικά μέγιστης εντροπίας ενός κειμένου με την παρατηρηθείσα εντροπία. Μέγιστη εντροπία παρουσιάζει ένα κείμενο που κάθε λέξη που θα περιελάμβανε θα εμφανιζόταν 1 φορά και άρα στο κείμενο αυτό όλες οι λέξεις θα ήταν άπαξ λεγόμενα. Όσο μεγαλύτερη είναι η Σχετική Εντροπία, τόσο λιγότερο τυποποιημένο είναι το κείμενο, και άρα περισσότερο «πλούσιο» λεξιλογικά.

- Φάσμα Συχνότητας Μήκους Λέξεων (LW1, 2, 3 … 14): Η συχνότητα με την οποία εμφανίζονται λέξεις με 1, 2, 3 … 14 χαρακτήρες στο κείμενο.

4 Μεθοδολογία

4.1.1 Το Ηλεκτρονικό Σώμα Κειμένων εκπαίδευσης

Για να διαμορφώσουμε μια συσχέτιση υφομετρικών χαρακτηριστικών και δυσκολίας κατανόησης των κειμένων επιλέξαμε 4 φοιτητές που γνωρίζουν την Ιταλική γλώσσα σε επίπεδο διδακτικής επάρκειας και τους ζητήσαμε να αναζητήσει ο καθένας στο Διαδίκτυο και να ανακτήσει 75 αυθεντικά ιταλικά κείμενα (25 με θέμα Πολιτικά, 25 με θέμα Πολιτιστικά και 25 με θέμα Αθλητικά) με σκοπό να τα χρησιμοποιήσουν κατά τη διδασκαλία σε προχωρημένους έλληνες σπουδαστές της Ιταλικής ως ξένης γλώσσας. Στη συνέχεια ζητήθηκε από τους φοιτητές να διαβάσουν κάθε κείμενο, δίχως χρονικό περιορισμό, και να το βαθμολογήσουν σε μια κλίμακα από το 1 έως το 10 ως προς την δυσκολία κατανόησης του. Το μέγεθος του Ηλεκτρονικού Σώματος Κειμένων (ΗΣΚ) που συλλέχθηκε φαίνεται στον παρακάτω πίνακα:

Πίνακας 1: Μέγεθος του ΗΣΚ

	Λέξεις					***Κείμενα***
	Μέσος όρος (λέξεων ανά κείμενο)	Τυπική Απόκλιση	Ελάχιστο μέγεθος κειμένου	Μέγιστο μέγεθος κειμένου	Σύνολο λέξεων	Αριθμός κειμένων
Φοιτητές 1	744	395	227	2257	55.832	75
2	579	325	100	1398	43.456	75

3	664	408	69	2164	49.766	75
4	410	178	176	1049	30.758	75

Για κάθε κείμενο υπολογίστηκαν τα υφομετρικά χαρακτηριστικά που προαναφέρθησαν στην ενότητα 3.2 με χρήση ειδικευμένου λογισμικού υφομετρικής ανάλυσης [19].

Για να βεβαιωθούμε ότι οι φοιτητές θα βαθμολογούσαν με τον ίδιο τρόπο τα κείμενα τους δόθηκε αρχικά ένα κοινό σετ 10 κειμένων. Η προκαταρκτική εξέταση της κατανομής της βαθμολόγησης των κοινών κειμένων έδειξε σημαντική διυποκειμενική μεταβλητότητα σε σχέση με την κλίμακα βαθμολόγησης. Για το λόγο αυτό αποφασίστηκε η μετατροπή της κλίμακας σε δίτιμη, δηλαδή για κάθε κείμενο εκφράστηκε με «Ναι» ή «Όχι» η καταλληλότητα του για τη διδακτική αξιοποίησή του στην τάξη βάσει της προσλαμβανόμενης αναγνωστικής του ευκολίας. Επιπλέον, η στατιστική ανάλυση που χρησιμοποιήθηκε (λογιστική παλινδρόμηση) υπολογίστηκε για τα δεδομένα του κάθε φοιτητή ξεχωριστά.

4.1.2 Στατιστική Ανάλυση

Για να διερευνήσουμε τη χρησιμότητα των υφομετρικών μεταβλητών στην κατηγορική αξιολόγηση της κειμενικής δυσκολίας χρησιμοποιήσαμε την στατιστική μέθοδο της λογιστικής παλινδρόμησης (logistic regression). Η λογιστική παλινδρόμηση στην απλούστερή της μορφή, όπως και στην απλή γραμμική παλινδρόμηση, είναι ένα στατιστικό μοντέλο που χρησιμοποιεί μια εξαρτημένη και μια ανεξάρτητη μεταβλητή και παράγει μια γραμμική εξίσωση που περιλαμβάνει μια σταθερά (b_0) και τον συντελεστή της παλινδρόμησης (b_1) για την ανεξάρτητη μεταβλητή (X). Η γραμμική αυτή εξίσωση ισούται με τον φυσικό λογάριθμο των συμπληρωματικών

πιθανοτήτων (odds) του γεγονότος της εξαρτημένης μεταβλητής και μπορεί να γραφεί ως εξής:

1. $$\ln\left(\frac{\mathrm{P(A)}}{1-\mathrm{P(A)}}\right) = b_0 + b_1(\mathrm{X})$$

όπου:

- ln (P(A)/1-P(A)) είναι ο φυσικός λογάριθμος των συμπληρωματικών πιθανοτήτων ενός γεγονότος Α (στο συγκεκριμένο παράδειγμα είναι πιθανότητα ένα κείμενο να χαρακτηριστεί ως κατάλληλο για διδακτική χρήση).
- b_0 είναι η σταθερά του μοντέλου.
- b_1 είναι ο συντελεστής της παλινδρόμησης που δείχνει την ένταση με την οποία η ανεξάρτητη μεταβλητή επηρεάζει την εξαρτημένη. Χάριν παραδείγματος θα μπορούσαμε να υποθέσουμε ότι στο συγκεκριμένο μοντέλο χρησιμοποιούμε ως μοναδική ανεξάρτητη μεταβλητή το Μέσο Μήκος Πρότασης.
- Χ είναι μια οποιαδήποτε ανεξάρτητη μεταβλητή και στο απλοποιημένο αυτό παράδειγμα το Μέσο Μήκος Πρότασης.

Αν αφαιρέσουμε το φυσικό λογάριθμο, η εξίσωση **1** μετατρέπεται ως εξής:

2. $$\frac{\mathrm{P(A)}}{1-\mathrm{P(A)}} = e^{b_0 + b_1 * X}$$

Αν λύσουμε την εξίσωση ως προς το P(A), η εξίσωση **2** γίνεται:

3. $$P(\mathrm{A}) = \frac{1}{1 + e^{-(b_0 + b_1 X)}}$$

όπου

- *e* είναι η βάση των φυσικών λογαρίθμων και ισούται με 2,718 και η οποία συνδέει την πιθανότητα του γεγονότος A (την καταλληλότητα του κειμένου για διδακτική χρήση) με το μέσο μήκος της πρότασης.

Το μοντέλο επεκτείνεται εύκολα σε πολλές ανεξάρτητες μεταβλητές και τότε η εξίσωση **3** γίνεται:

4. $$P(\mathrm{A}) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 \cdots b_n X_n)}}$$

Η εξίσωση **4** είναι αυτή στην οποία στηρίζεται ο υπολογισμός της πολυπαραγοντικής λογιστικής παλινδρόμησης και μας επιτρέπει να διερευνήσουμε το βαθμό που κάθε ανεξάρτητη μεταβλητή επηρεάζει την αντίληψη δυσκολίας του κάθε κειμένου. Δεδομένου ότι ο αριθμός των ανεξάρτητων μεταβλητών είναι μεγάλος σε σχέση με το μέγεθος του δείγματος επιλέξαμε να χρησιμοποιήσουμε την αυξητική βηματική μέθοδο (forward stepwise method) για να αναπτύξουμε το λογιστικό μοντέλο. Η μέθοδος αυτή ξεκινάει το λογιστικό μοντέλο εντάσσοντας μόνο μία ανεξάρτητη μεταβλητή, αυτή που έχει τη μεγαλύτερη επίδραση στην εξαρτημένη. Στη συνέχεια, δοκιμάζει τις άλλες μεταβλητές και προσθέτει στο μοντέλο κάποια μόνο όταν η εφαρμογή του μοντέλου αυξάνεται με στατιστικά σημαντικό τρόπο. Το λογιστικό μοντέλο που προκύπτει περιλαμβάνει μόνο τις στατιστικά σημαντικές ανεξάρτητες μεταβλητές γεγονός που το κάνει καταλληλότερο για μικρότερα μεγέθη δεδομένων εκπαίδευσης.

5 Αποτελέσματα

Τα αποτελέσματα της βηματικής λογιστικής παλινδρόμησης δείχνουν ότι οι υφομετρικές μεταβλητές μπορούν να χρησιμοποιηθούν ως αξιόπιστοι

δείκτες της δυσκολίας ενός κειμένου. Βάσει τον αποτελεσμάτων της λογιστικής παλινδρόμησης η αυτόματη κατάταξη των κειμένων ανά φοιτητή φαίνεται στους παρακάτω πίνακες σύγχυσης (confusion matrix):

Πίνακας 2: Πίνακας σύγχυσης για το λογιστικό μοντέλο κειμενικής δυσκολίας που χρησιμοποιεί υφομετρικές ανεξάρτητες μεταβλητές

Παρατηρηθείσα συχνότητα			**Πρόβλεψη**		
			Εύκολο (N)	Δύσκολο (N)	% σωστής πρόβλεψης
Φοιτητής	1	Εύκολο	41	6	87,2
		Δύσκολο	8	20	71,4
		Συνολικό %			**81,3**
Φοιτητής	2	Εύκολο	48	4	92,3
		Δύσκολο	10	5	33,3
		Συνολικό %			**80,1**
Φοιτητής	3	Εύκολο	15	16	48,4
		Δύσκολο	10	30	75,0
		Συνολικό %			**63,4**
Φοιτητής	4	Εύκολο	4	8	33,3
		Δύσκολο	1	60	98,4
		Συνολικό %			**87,7**

Τα σκιασμένα φατνία του πίνακα δείχνουν τους αριθμούς των κειμένων που κατηγοριοποιήθηκαν σωστά ως «Εύκολα» και σωστά ως «Δύσκολα». Τα μη σκιασμένα φατνία δείχνουν τους αριθμούς των κειμένων που είτε ήταν «Δύσκολα» και κατηγοριοποιήθηκαν ως «Εύκολα», είτε το αντίστροφο. Όπως γίνεται φανερό τα ποσοστά σωστής πρόβλεψης είναι αρκετά ικανοποιητικά και ποικίλλουν από 87,7% έως 63,4% με μέσο όρο το 78,1%.

Ενδιαφέρον παρουσιάζει η σύνθεση του λογιστικού μοντέλου για κάθε φοιτητή:

Πίνακας 3: Το βέλτιστο λογιστικό μοντέλο για την ερμηνεία της κειμενικής δυσκολίας για κάθε φοιτητή.

Ανεξάρτητες μεταβλητές		***B***	***S.E.***	***Wald***	***df***	***Sig.***
Φοιτητής 1	SLsd	-0,256	0,089	8,371	1	0,004
	Relative Entropy	0,420	0,174	5,865	1	0,015
	9LW	0,629	0,289	4,727	1	0,030
	11LW	1,309	0,435	9,045	1	0,003
	Σταθερά	11,558	5,737	4,059	1	0,044
Φοιτητής 2	LD	-0,768	0,319	5,811	1	0,016
	11LW	0,973	0,340	8,180	1	0,004
	Σταθερά	1,123	1,842	0,372	1	0,542
Φοιτητής 3	12LW	1,104	0,434	6,470	1	0,011
	Σταθερά	-1,086	0,572	3,609	1	0,057
Φοιτητής 4	TTR	0,147	0,056	6,885	1	0,009
	9LW	0,619	0,282	4,831	1	0,028
	Σταθερά	-11,792	4,181	7,956	1	0,005

Οι αντίστοιχες εξισώσεις Αναγνωστικής Δυσκολίας για κάθε φοιτητή είναι οι ακόλουθες:

Αναγνωστική Δυσκολία[8] *(Φοιτ.1) = 11,558 - 0,256* SLsd + 0,42*Relative Entropy + 0,629*9LW + 1,309*11LW*

*Αναγνωστική Δυσκολία (Φοιτ.2) = 1,123 - 0,768*LD + 0,973*11LW*

*Αναγνωστική Δυσκολία (Φοιτ.3) = - 1,086 + 1,104*12LW*

*Αναγνωστική Δυσκολία (Φοιτ.4) = - 11,792 + 0,147*TTR + 0,619*9LW*

Όπως φαίνεται από τις παραπάνω εξισώσεις κάθε φοιτητής αντιλαμβάνεται διαφορετικά την κειμενική δυσκολία. Για το Φοιτητή 1 η κειμενική δυσκολία σχετίζεται κατά φθίνουσα σειρά σπουδαιότητας[9] με το Φάσμα Συχνότητας Μήκους Λέξεων και ειδικότερα με τη συχνότητα των λέξεων που έχουν 11 γράμματα (11LW) στο κείμενο, την Τυπική Απόκλιση του Μέσου Μήκους Πρότασης του κειμένου (SLsd), με την κειμενική Σχετική Εντροπία (Relative Entropy) και με τη συχνότητα των λέξεων με 9 γράμματα (9LW). Για το Φοιτητή 2 η κειμενική δυσκολία καθορίζεται από το Φάσμα Συχνότητας Μήκους Λέξεων και ειδικότερα από τη συχνότητα των λέξεων που έχουν 11 γράμματα (11LW) στο κείμενο και την Λεξιλογική Πυκνότητα του κειμένου (LD). Για το Φοιτητή 3 η κειμενική δυσκολία καθορίζεται μόνο από έναν παράγοντα, το Φάσμα Συχνότητας

[8] Η Αναγνωστική Δυσκολία στις συγκεκριμένες εξισώσεις αναγνωσιμότητας παίρνει τιμές από το 1 (εύκολο κείμενο κατάλληλο για διδακτική αξιοποίηση) έως το 2 (δύσκολο κείμενο ακατάλληλο για χρήση στην τάξη). Ενδιάμεσες τιμές από το 1,01 -1,50 ερμηνεύονται ως 1 και τιμές από το 1,51 έως το 1,99 ερμηνεύονται ως 2.

[9] Η επίδραση της κάθε ανεξάρτητης μεταβλητής στην εξαρτημένη (δυσκολία κειμένου) καθορίζεται από την τιμή Wald η οποία είναι ο λόγος του εκτιμητή Β με το τυπικό του λάθος (S.E.). Όσο μεγαλύτερη είναι η τιμή τόσο μεγαλύτερη η επίδραση της μεταβλητής στο λογιστικό μοντέλο.

Μήκους Λέξεων και ειδικότερα τη συχνότητα των λέξεων που έχουν 12 γράμματα (12LW). Τέλος, για το Φοιτητή 4 η κειμενική δυσκολία σχετίζεται με την υψηλό λόγο TTR (TTR) και το Φάσμα Συχνότητας Μήκους Λέξεων που στη συγκεκριμένη περίπτωση είναι η συχνότητα των λέξεων με 9 γράμματα (9LW).

Η εξέταση των ανεξάρτητων μεταβλητών που εμφανίστηκαν να επηρεάζουν με στατιστικά σημαντικό τρόπο τη δυσκολία κατανόησης των κειμένων μας αποκαλύπτει ότι όλοι οι συμμετέχοντες στην έρευνα επηρεάστηκαν από την συχνότητα μεγάλων λέξεων (>9 γράμματα) στα κείμενα, ενώ η κλασική μεταβλητή του μέσου μήκους των λέξεων στο κείμενο δεν άσκησε στατιστικά σημαντική επίδραση. Αυτό το αποτέλεσμα μας δείχνει ότι, τουλάχιστον στους μη μητρικούς ομιλητές της Ιταλικής, η ύπαρξη μεμονωμένων μεγάλων λέξεων δυσκολεύει περισσότερο την κειμενική κατανόηση, σε σχέση με την ύπαρξη ενός αυξημένου μέσου όρου του μήκους όλων των λέξεων στο κείμενο. Από τους υπόλοιπους υφομετρικούς δείκτες που εξετάστηκαν οι ακόλουθοι τέσσερεις σχετίστηκαν με την κειμενική δυσκολία:

α) Λεξιλογική Πυκνότητα για το Φοιτητή 2. Η συνάφεια της λεξιλογικής πυκνότητας στο συγκεκριμένο λογιστικό μοντέλο είναι αρνητική (B= -0,768), δηλαδή η αύξηση της λεξιλογικής πυκνότητας συνεπάγεται μείωση της κειμενικής δυσκολίας. Αυτό, εξηγείται μέσα από την πληροφοριακή δομή του κειμένου, αφού όσο μεγαλύτερη λεξιλογική πυκνότητα παρουσιάζει αυτό τόσο περισσότερες λέξεις «περιεχομένου» εμφανίζονται και επομένως το νόημά του γίνεται σαφέστερο και άρα ευκολότερο.

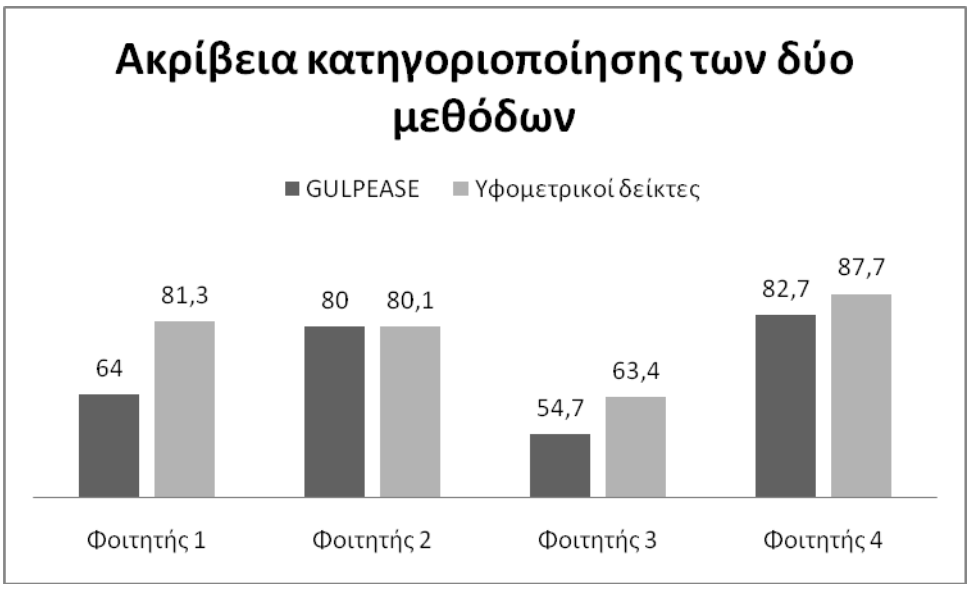
β) Σχετική Εντροπία για το Φοιτητή 1. Η συνάφεια της σχετικής εντροπίας στο συγκεκριμένο μοντέλο είναι θετική (Β= 0,42), δηλαδή η αύξηση της σχετικής εντροπίας συνιστά αύξηση της κειμενικής δυσκολίας. Εδώ η εξήγηση δίνεται μέσα από την κατανόηση της λεξιλογικής επαναληψιμότητας (lexical repeatability). Όσο μεγαλύτερη επαναληψιμότητα εμφανίζουν λέξεις ή λεκτικά σχήματα, τόσο μεγαλύτερη είναι η τυποποίηση του κειμένου και επομένως μειώνεται η πιθανότητα να συναντήσουμε νέες λέξεις που πιθανόν να είναι άγνωστες. Επομένως, όσο μεγαλώνει η σχετική εντροπία ενός κειμένου τόσο πιο «απρόβλεπτη» γίνεται η εμφάνιση των λέξεων σε αυτό με αποτέλεσμα να αυξάνεται η αίσθηση της δυσκολίας του από τον αναγνώστη.

γ) Τυπική Απόκλιση Μέσου Μήκους Προτάσεων για το Φοιτητή 1. Η συνάφεια της συγκεκριμένης μεταβλητής στο λογιστικό μοντέλο είναι αρνητική (Β= -0,256). Για το συγκεκριμένο φοιτητή η αίσθηση της δυσκολίας του κειμένου μεγαλώνει όσο μειώνεται η τυπική απόκλιση του προτασιακού μήκους στο κείμενο. Η μείωση της τυπικής απόκλισης συμβαίνει όταν τα μήκη των προτάσεων του κειμένου συσπειρώνονται κοντά στο μέσο όρο και δεν υπάρχουν μεγάλες αυξομειώσεις μήκους. Η αρνητική συσχέτιση που παρατηρείται πρέπει να σχετίζεται με το προσωπικό στυλ εκμάθησης του φοιτητή αφού στους υπόλοιπους συμμετέχοντες εμφάνισε θετική συνάφεια.

δ) TTR για το Φοιτητή 4. Η συνάφεια της λόγου Type – Token με την δυσκολία του κειμένου είναι θετική (Β= 0,147). Η αύξηση της λεξιλογικής διαφοροποίησης συνεπάγεται αύξηση της δυσκολίας του κειμένου.

Από τα παραπάνω γίνεται σαφές ότι η δυσκολία κατανόησης του κειμένου στην ξένη γλώσσα αποτελεί μια σύνθετη διαδικασία η οποία σχετίζεται άμεσα με το προσωπικό στιλ εκμάθησης. Είναι ενδεικτικό ότι κάθε φοιτητής αξιοποίησε διαφορετικό υφομετρικό δείκτη λεξιλογικού πλούτου (Λεξιλογική Πυκνότητα, Σχετική Εντροπία, TTR) και άρα προσέλαβε την λεξιλογική δυσκολία υπό διαφορετικό πρίσμα.

Η απόδοση των υφομετρικών δεικτών που προτάθηκαν είναι σημαντικά καλύτερη από το δείκτη GULPEASE που όπως είδαμε στην ενότητα 2 αποτελεί αυτή τη στιγμή τον πιο αξιόπιστο δείκτη αναγνωσιμότητας για την Ιταλική γλώσσα. Η σύγκριση της ακρίβειας ανά φοιτητή φαίνεται στο παρακάτω διάγραμμα:



Διάγραμμα 1: Συγκριτική εξέταση της ακρίβειας κατηγοριοποίησης του δείκτη GULPEASE και των υφομετρικών δεικτών της παρούσας μελέτης.

Όπως γίνεται φανερό και από το διάγραμμα ο συνδυασμός υφομετρικών δεικτών που χρησιμοποιήθηκε στην παρούσα μελέτη κατηγοριοποιεί με περισσότερη ακρίβεια τα κείμενα ως προς την αναγνωστική τους δυσκολία. Αυτό εξηγείται εν μέρει γιατί ο δείκτης GULPEASE έχει αξιολογηθεί από δείγμα μητρικών ομιλητών της Ιταλικής. Επίσης, η παρούσα μελέτη χρησιμοποιεί ευρύτερο φάσμα μεταβλητών με μεγαλύτερη έμφαση στον λεξιλογικό πλούτο, παράμετρος που είναι ιδιαίτερα σημαντική στη γλωσσική εκμάθηση.

6 Συμπεράσματα

Στην παρούσα μελέτη εξετάσαμε το θέμα της αναγνωσιμότητας ιταλικών κειμένων από έλληνες φοιτητές της Ιταλικής ως ξένης γλώσσας. Η βασική ερευνητική υπόθεση της συγκεκριμένης έρευνας ήταν ότι η μέτρηση της αναγνωσιμότητας των ιταλικών κειμένων μπορεί να βελτιωθεί σημαντικά αν αξιοποιήσουμε τις υφομετρικές μεταβλητές που έχουν χρησιμοποιηθεί μέχρι τώρα στην αυτόματη αναγνώριση του συγγραφέα ενός κειμένου. Λόγω του διαφορετικού τρόπου με τον οποίο οι φοιτητές της έρευνας αξιολόγησαν τα κείμενα αποφασίστηκε η μετατροπή της αρχικής δεκάβαθμης κλίμακας αξιολόγησης σε κατηγορική και η ανάλυση των αποτελεσμάτων για κάθε συμμετέχοντα στην έρευνα ξεχωριστά. Για την στατιστική επεξεργασία των δεδομένων χρησιμοποιήθηκε η αυξητική βηματική λογιστική παλινδρόμηση και τα βασικότερα συμπεράσματα που προέκυψαν είναι τα ακόλουθα:

- Οι υφομετρικές μεταβλητές που χρησιμοποιήθηκαν συσχετίστηκαν με τη δυσκολία κατανόησης των κειμένων και προσέφεραν ικανοποιητικά ποσοστά σωστής πρόβλεψης που κυμάνθηκαν από 63,4 – 87,7% με μέσο όρο 78,1%.

- Οι υφομετρικές μεταβλητές που εμφάνισαν συσχέτιση με την αναγνωσιμότητα είναι η Λεξιλογική Πυκνότητα, η Σχετική Εντροπία, η Τυπική Απόκλιση του Μέσου Μήκους Πρότασης του κειμένου, ο λόγος TTR και το Φάσμα Συχνότητας Μήκους Λέξεων (η συχνότητα λέξεων 9, 11 και 12 γραμμάτων).

- Από τις μεταβλητές αυτές, μόνο η Τυπική Απόκλιση του Μέσου Μήκους Πρότασης σχετίζεται έμμεσα με τη συντακτική πολυπλοκότητα του κειμένου και μάλιστα στην παρούσα έρευνα εμφάνισε αρνητική συνάφεια σε έναν μόνο φοιτητή γεγονός που σημαίνει ότι μάλλον πρόκειται για ιδιοσυγκρασιακή μαθησιακή μεταβλητή. Όλες οι υπόλοιπες σχετίζονται έμμεσα με το λεξιλόγιο και τη δυσκολία του.

- Η υψηλή συχνότητα εμφάνισης λέξεων από 9 έως 12 χαρακτήρες συσχετίστηκε με τη δυσκολία κατανόησης του κειμένου σε όλους τους συμμετέχοντες της έρευνας.

- Οι υφομετρικοί δείκτες της παρούσας έρευνας κατηγοριοποίησαν τα κείμενα ως προς την δυσκολία τους με μεγαλύτερη ακρίβεια από αυτήν του δείκτη GULPEASE που θεωρείται ο πιο αξιόπιστος δείκτης αναγνωσιμότητας στην Ιταλική γλώσσα. Αυτό το αποτέλεσμα υπογραμμίζει την αναγκαιότητα ανάπτυξης ξεχωριστών

εξισώσεων αναγνωσιμότητας για την αξιολόγηση κειμενκού υλικού στο οποίο θα εκτεθούν μη μητρικοί ομιλητές της γλώσσας.

Τα παραπάνω συμπεράσματα φωτίζουν μια νέα προσέγγιση στη μεθοδολογία υπολογισμού της αναγνωσιμότητας ιταλικών κειμένων με εφαρμογή στη διδασκαλία της Ιταλικής ως ξένης γλώσσας. Μελλοντική έρευνα θα προσανατολιστεί στην δοκιμή μεγαλύτερου αριθμού υφομετρικών δεικτών και την αξιολόγηση της αναγνωσιμότητας σε μεγαλύτερο εύρος κειμενικών ειδών και θεμάτων από περισσότερους κριτές. Τέλος, έρευνες αυτού του τύπου θα μπορούσαν να γίνουν σε παράλληλα ΗΣΚ για να διερευνηθεί η ύπαρξη καθολικών περιορισμών αναγνωσιμότητας που σχετίζονται με την κειμενική δυσκολία ανεξαρτήτου γλώσσας που έχει γραφεί ένα κείμενο.

Βιβλιογραφικές Αναφορές

1. Little, D., Sean, D., and Singleton, D. (1994) The communicative approach and authentic texts, in *Teaching modern languages* (Swarbrick, A., Ed.), pp 43-47, Routledge, London.
2. (2006) All our N-gram belong to you. Available: http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html.
3. Karpova, L., V. (1999) Consider the Following When Selecting and Using Authentic Materials, *TESOL Matters 9*, *2*, 18.
4. Dillon, A., McKnight, C., and Richardson, J. (1988) Reading from paper versus reading from screen, *The Computer Journal 31*, *5*, 457-464.
5. Min-chen, T. (2008) The difficulties that EFL learners have with reading text on the Web, *The Internet TESL Journal XIV*, *2*. Available: http://iteslj.org/Articles/Tseng-TextOnTheWeb.html.
6. Hung-Tzu, H., and Hsien-Chin, L. (2007) Vocabulary learning in an automated graded reading program, *Language Learning and Technology 11*, *3*, 64-82.
7. Klare, G. R. (1963) *The measurement of readability*, Iowa State University Press, Ames, Iowa.
8. Sherman, A. L. (1893) *Analytics of literature: a manual for the objective study of English prose and poetry*, Ginn & Co, Boston.
9. Gray, W. S., and Leary, B. (1935) *What makes a book readable*, Chicago University Press, Chicago.
10. Lorge, I. (1939) Predicting reading difficulty of selections for children, *Elementary English Review 16*, 229-233.
11. Dale, E., and Chall, J. S. (1948) A formula for predicting readability, *Educational Research Bulletin 27*, 1-20, 37-54.
12. Flesch, R. (1948) A new readability yardstick, *Journal of Applied Psychology 32*, 221-233.
13. Kincaid, P. J., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975) Derivation of new readability formulas (Automated Readability Index Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, Millington, TN.
14. Lucisano, P., and Piemontese, M. E. (1988) GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana, *Scuola e città 3*, 110-124.
15. Lucisano, P. (1992) *Misurare le parole*, Kepos, Rome.
16. Mastidoro, N., and Amizzoni, M. (1993) Linguistica applicata alla leggibilità: considerazioni teoriche e applicazioni, *Bollettino della Società*

Filosofica Italiana 149. Available: http://lgxserver.uniba.it/lei/sfi/bollettino/bollettino.htm.

17. Sebastiani, F. (2002) Machine learning in automated text categorization, *ACM Computing Surveys (CSUR) 34*, *1*, 1-47.
18. Santini, M. (2006) Some issues in Automatic Genre Classification of Web Pages, *JADT 06-Actes des 8 Journées internationales d'analyse statistiques des donnés textuelles 2*.
19. Mikros, G. K. (2006) Authorship attribution in Modern Greek newswire corpora, in *Proceedings of the SIGIR 2006 Workshop on Directions in Computational Analysis of Stylistics in Text Retrieval* (Uzuner, O., Argamon, S., and Karlgren, J., Eds.), pp 43-47, ACM, Seattle, Washington, USA.
20. Witten, I. H., and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.
21. Morton, A., Q. (1965) *The Authorship of the Pauline Epistles: A Scientific Solution*, University of Saskatchewan.
22. McMichael, G. L., and Glenn, E. M. (1962) *Shakespeare and his rivals; a casebook on the authorship controversy*, Odyssey Press, New York.
23. Solan, L. M., and Tiersma, P. M. (2004) Author identification in american courts, *Applied Linguistics 25*, *4*, 448-465.
24. Rudman, J. (1997) The state of authorship attribution studies: some problems and solutions, *Computers and the Humanities 31*, *4*, 351-365.
25. Mikros, G. K., and Argiri, E. K. (2007) Investigating topic influence in authorship attribution, in *Proceedings of the SIGIR'07 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection* (Stein, B., Koppel, M., and Stamatatos, E., Eds.), pp 29-35, CEUR, Amsterdam, Netherlands.
26. Mikros, G. K., and Carayannis, G. (2000) Modern Greek corpus taxonomy, in *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperdis, S., and Stainhaouer, G., Eds.), pp 129–134, ELRA, Athens, Greece.
27. Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G., and Tambouratzis, D. (2004) Discriminating the registers and styles in the Modern Greek language-Part 1: Diglossia in stylistic analysis, *Literary and Linguistic Computing 19*, *2*, 197-220.
28. Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G., and Tambouratzis, D. (2004) Discriminating the Registers and Styles in the Modern Greek Language-Part 2: Extending the Feature

Vector to Optimize Author Discrimination, *Literary and Linguistic Computing 19*, *2*, 221-242.
29. Hoover, D. (2003) Another perspective on vocabulary richness, *Computers and the Humanities 37*, 151-178.
30. Miranda, G., Antonio, and Calle, M., Javier. (2007) Function words in authorship attribution studies, *Literary and Linguistic Computing 22*, *1*, 49-66.
31. Tweedie, F. J., and Baayen, H. R. (1998) How variable may a constant be? Measures of lexical richness in perspective, *Computers and the Humanities 32*, *5*, 323-352.
32. Oakes, M. P. (1998) *Statistics for corpus linguistics*, Edinburgh University Press, Edinburgh.