ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΦΙΛΟΣΟΦΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΦΙΛΟΛΟΓΙΑΣ

ΤΟΜΕΑΣ ΓΛΩΣΣΟΛΟΓΙΑΣ

ΔΙΕΠΙΣΤΗΜΟΝΙΚΟ – ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΤΕΧΝΟΓΛΩΣΣΙΑ VII

Αυτόματη κατηγοριοποίηση κειμένων της ισπανικής στα επίπεδα γλωσσομάθειας του *Κοινού Ευρωπαϊκού Πλαισίου Αναφοράς για τις Γλώσσες*.

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΔΗΜΗΤΡΗΣ ΛΑΜΠΡΙΝΟΣ

ΕΠΙΒΛΕΠΩΝ

ΓΙΩΡΓΟΣ ΜΙΚΡΟΣ
ΚΑΘΗΓΗΤΗΣ
Ε.Κ.Π.Α.

ΕΠΙΤΡΟΠΗ

ΓΙΩΡΓΟΣ ΜΑΡΚΟΠΟΥΛΟΣ
ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ
Ε.Κ.Π.Α.

ΓΙΑΝΝΗΣ ΜΑΪΣΤΡΟΣ
ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ
Ε.Μ.Π.

Ευχαριστίες.

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή αυτής της διπλωματικής εργασίας Γιώργο Μικρό για την καθοδήγησή του και για τη μύησή μου στον κόσμο της υπολογιστικής γλωσσολογίας, ήδη από την εποχή των προπτυχιακών μου σπουδών.

Πολλά ευχαριστώ στη φίλη και μαθήτρια Μαρία Αναστασάτου και στον φίλο και συνάδελφο Θανάση Μπάθα για την προθυμία που έδειξαν να με βοηθήσουν με την αξιολόγηση των κειμένων του σώματος εκπαίδευσης.

Τέλος, ένα μεγάλο ευχαριστώ στη μητέρα μου Χριστίνα για την υποστήριξη της καθόλη τη διάρκεια των σπουδών μου και ιδιαίτερα κατά το τελευταίο στάδιο της εκπόνησης αυτής της εργασίας.

Περίληψη.

Σκοπός αυτής της εργασίας είναι η εκπαίδευση ενός ταξινομητή γενικής χρήσης, με τεχνικές μηχανικής μάθησης, ο οποίος να κατατάσσει κείμενα της ισπανικής, στα επίπεδα γλωσσομάθειας του *Κοινού Ευρωπαϊκού Πλαισίου για τις Γλώσσες*. Δημιουργήσαμε ένα σώμα κειμένων βαθμονομημένων σε αυτή την κλίμακα τα οποία συλλέξαμε από διαφόρων τύπων πηγές που χρησιμοποιούνται στη διδασκαλία των ισπανικών (μεθόδους, διαβαθμισμένα αναγνώσματα, μοντέλα εξετάσεων και άλλα). Κάναμε μετρήσεις στα κείμενα, καταγράφοντας τις τιμές διαφόρων γλωσσικών χαρακτηριστικών (υφομετρικών, λεξιλογικών, γραμματικών, συντακτικών και κάποιων που βασίζονταν στις αναλογίες που εμφανίζονται τα διάφορα μέρη του λόγου). Τροφοδοτήσαμε με τα αποτελέσματα των μετρήσεων έναν αλγόριθμο Μηχανών Διανυσμάτων Υποστήριξης και καταλήξαμε με δύο ταξινομητές. Έναν για τα 6 επίπεδα αναφοράς (Α1-Γ2), που πέτυχε ακρίβεια 67%, και έναν για τα 3 ευρύτερα στάδια του ΚΕΠΑΓ (Α,Β και Γ), που πέτυχε ακρίβεια 85,63%.

Abstract.

The aim of this thesis is to train a classifier that will classify Spanish texts into the levels described by the *Common European Framework of Reference for Languages*, using machine learning technics. We created a corpus of texts graded in the scale of the common reference levels and gathered from various sources that are related to the teaching of Spanish (textbooks, graded readers, sample exam papers and others). We measured the values of various types of linguistic attributes of the texts (stylometric, lexical, grammatical, syntactic and POS-based). We feeded the results of the measurments to a Support Vector Machine algorithm and obtained two classifiers. One for the six levels of reference (A1-C2) that achieved an accuracy of 67% and another for the three broad levels (A-C) whose accuracy reached 85,63%.

Περιεχόμενα.

1. Εισαγωγή.

1.1 Σύντομη διατύπωση του ερευνητικού προβλήματος και της σημασίας του.

Η χρήση αυθεντικών κειμένων στην διδασκαλία των ξένων γλωσσών απασχολεί εδώ και αρκετά χρόνια τους ειδικούς. Με τον όρο «αυθεντικό κείμενο» και γενικότερα «αυθεντικό γλωσσικό υλικό» αναφερόμαστε σε οποιοδήποτε γλωσσικό υλικο που έχει παραχθεί για επικοινωνία μεταξύ φυσικών ομιλητών της γλώσσας σε ένα πλαίσιο μη εκπαιδευτικό, που δεν έχει βαθμονομηθεί με κανένα τρόπο από γλωσσολογικής πλευράς και ούτε είναι οργανωμένο με τρόπο που να παρουσιάζει τη χρήση ενός συγκεκριμένου γραμματικού φαινομένου (Miguel García Arreza, María Dolores Zamora Navas, 1994). Πρόκειται, δηλαδή, για πραγματικά κείμενα τα οποία δεν έχουν δημιουργηθεί για σπουδαστές ξένων γλωσσών αλλά για φυσικούς ομιλητές της εκάστοτε γλώσσας (Harmer, 2003). Κάποια είδη κειμένων που θα μπορούσαμε να αναφέρουμε ενδεικτικά ως αυθεντικά, είναι ένα άρθρο εφημερίδας, ένα γράμμα, ένα διαφημιστικό φυλλάδιο, μια συνταγή φαγητού, οδηγίες χρήσης ενός προϊόντος και άλλα.

Που εντοπίζεται όμως η αξία της ενσωμάτωσης των αυθεντικών κειμένων στην διδακτική πρακτική και τι τα διαχωρίζει από τα κείμενα που έχουν γραφτεί για τα διδακτικά εγχειρίδια; Στα τελευταία, πολλές φορές η συμπερίληψη συγκεκριμένου λεξιλογίου και γραμματικών φαινομένων παραγκωνίζει την πραγματολογία και την πραγματική χρήση της γλώσσας. Επιπλέον, η επιλογή αυτού το λεξιλογίου και των γραμματικών φαινομένων που παρουσιάζονται, συχνά δεν βασίζεται σε εμπειρικά δεδομένα αλλά σε μια μορφή διαίσθησης σχετικά με την γλώσσα που θα έπρεπε να χρησιμοποιηθεί (Montalbán, 2007). Αν μάλιστα λάβουμε υπόψη το γεγονός ότι ο τελικός σκοπός των μαθημάτων ξένων γλωσσών είναι να προετοιμάσουν τους μαθητές για την αυθεντική γλώσσα του πραγματικού κόσμου, συνειδητοποιούμε την ανάγκη να τους δώσουμε την ευκαιρία να αντιμετωπίσουν αυτήν τη γλώσσα μέσα στην τάξη (Hedge, 2000).

Σήμερα, η πρόσβαση που έχουν οι καθηγητές ξένων γλωσσών σε αυθεντικό γλωσσικό υλικό, κυρίως λόγω της διάδοσης των νέων τεχνολογιών και της ανάπτυξης του Παγκόσμιου Ιστού, είναι πρακτικά απεριόριστη. Μια άστοχη επιλογή αυθεντικών κειμένων που προορίζονται για εκμετάλλευση κατά τη διαδικασία της διδασκαλίας μιας ξένης γλώσσας θα μπορούσε να δημιουργήσει προβλήματα. Έχει επισημανθεί ότι τα αυθεντικά κείμενα πολλές φορές περιλαμβάνουν αχρείαστο λεξιλόγιο και πολύπλοκες δομές που μπορούν να απογοητεύσουν και να αγχώσουν τους μαθητές (Richards, 2001). Από την άλλη, οι μαθητές που εκτίθενται σε υλικό που ανταποκρίνεται στις ανάγκες και τις ικανότητές τους, έχουν πολλά να κερδίσουν (Huizenga & Ruzic, 1994). Όσο αφορά το λεξιλόγιο, έρευνες έχουν δείξει ότι ένα μικρό ποσοστό άγνωστων λέξεων σε ένα κείμενο, έως 5%, επιτρέπει την ικανοποιητική κατανόησή του, όπως επίσης και την εφαρμογή των κατάλληλων στρατηγικών που βοηθούν τον μαθητή να συναγάγει το νόημα τους (Huang & Liou, 2007). Γενικότερα, έχει επισημανθεί (Krashen, 1982; Vygotsky, 1978) ότι η ιδανική δυσκολία ενός αναγνώσματος είναι ελάχιστα μεγαλύτερη από το τρέχον επίπεδο του μαθητή. Το βάρος της επιλογής πέφτει στον καθηγητή για τον οποίο είναι πολύ

εύκολο να αντλήσει από το διαδίκτυο ένα μεγάλο αριθμό κειμένων με το θέμα που τον ενδιαφέρει και πολύ χρονοβόρο και επίπονο να διαλέξει, μέσα από όλα αυτά, κάποιο που να ταιριάζει στο επίπεδο των μαθητών για τους οποίους προορίζεται.

Από τα παραπάνω γίνεται προφανής η χρησιμότητα ενός υπολογιστικού εργαλείου που θα συνεισέφερε στην αυτοματοποίηση της εκτίμησης της δυσκολίας ενός κειμένου. Πιο συγκεκριμένα, μας απασχολεί το πρόβλημα της αυτόματης κατάταξής ισπανικών κειμένων στο σωστό επίπεδο γλωσσομάθειας και προς αυτήν την κατεύθυνση κινείται αυτή η εργασία, το αντικείμενο της οποίας εντάσσεται στον γενικότερο κλάδο της μελέτης και ποσοτικής εκτίμησης της αναγνωσιμότητας (readability).

1.2 Η αναγνωσιμότητα των κειμένων.

Κοινό τόπο στους ορισμούς που έχουν δοθεί για την αναγνωσιμότητα αποτελεί η διαπίστωση ότι περιγράφει την ευκολία με την οποία ένα κείμενο μπορεί να διαβαστεί και να γίνει κατανοητό από τους αναγνώστες (Barzilay & Lapata, 2008; Hargis, 2000; Klare, 1963; McLaughlin, 1969). Όσο αφορά τους παράγοντες από τους οποίους εξαρτάται η αναγνωσιμότητα, οι μελετητές που έχουν επιχειρήσει να την ορίσουν επικεντρώνουν την προσοχή τους σε διαφορετικούς. Ο G. Harry McLaughlin (1969), δημιουργός της εξίσωσης SMOG για τον υπολογισμό της αναγνωσιμότητας, την όρισε ως εξής: «Ο βαθμός στον οποίο μια δεδομένη ομάδα ανθρώπων βρίσκει ένα συγκεκριμένο ανάγνωσμα ελκυστικό και κατανοητό». Σε αυτόν τον ορισμό τονίζεται η αλληλεπίδραση μεταξύ του κειμένου και μιας κατηγορίας αναγνωστών με συγκεκριμένα χαρακτηριστικά όπως η αναγνωστική τους δεξιότητα, οι προηγούμενες γνώσεις τους για το θέμα και το κίνητρό τους (DuBay, 2004). Ο Klare (1963) ενδιαφέρεται περισσότερο για το ύφος αναφερόμενος στην αναγνωσιμότητα ως την «ευκολία της κατανόησης που οφείλεται στο ύφος της γραφής». Τον πιο πλήρη ορισμό, ίσως τον δίνουν οι Dale και Chall (1948): «Το άθροισμα όλων των στοιχείων, των αλληλεπιδράσεων συμπεριλαμβανομένων, σε ένα έντυπο υλικό, που επηρεάζουν την επιτυχία που μια ομάδα αναγνωστών έχει με αυτό. Ως επιτυχία εννοείται ο βαθμός στον οποίο το καταλαβαίνουν, το διαβάζουν με μια βέλτιστη ταχύτητα και το βρίσκουν ενδιαφέρον».

Το ενδιαφέρον για μια ιστορική και στατιστική προσέγγιση της λογοτεχνίας, η ενσωμάτωση στην αμερικάνικη κοινωνία και στο αμερικάνικο εκπαιδευτικό σύστημα πληθυσμών των οποίων η μητρική γλώσσα δεν ήταν η αγγλική, η ανάγκη των εκπαιδευτικών να βρουν κατάλληλα αναγνώσματα για τους μαθητές τους, η κρισιμότητα της επαρκούς κατανόησης από τον γενικό πληθυσμό κειμένων σχετικών με ευαίσθητους τομείς όπως η Υγεία, η σημασία της σωστής τήρησης κανόνων και οδηγιών σε επαγγελματικούς χώρους όπως οι Ένοπλες Δυνάμεις και άλλοι παράγοντες, ήταν οι αιτίες που η αναγνωσιμότητα και η ποσοτική της εκτίμησή απασχόλησε τους ερευνητές κατά τη διάρκεια σχεδόν όλου του εικοστού αιώνα, με τα τελευταία χρόνια να την βλέπουν να επωφελείται από τις προόδους στην Επεξεργασία Φυσικής Γλώσσας. Το θεωρητικό υπόβαθρο και τους σημαντικότερους σταθμούς αυτής της έρευνας, θα εξετάσουμε στο κεφάλαιο που ακολουθεί.

1.3 Δομή της εργασίας.

Η εργασία αποτελείται από τρία κεφάλαια. Στο πρώτο, κάνουμε μια σύντομη ιστορική αναδρομή στις σχετικές με την αναγνωσιμότητα εργασίες για να έχουμε μία εικόνα του πως έχει προσεγγιστεί το θέμα από διάφορους ερευνητές, σε διάφορες χρονικές περιόδους. Μας δίνεται η ευκαιρία να αναφερθούμε στην κριτική που έχουν δεχθεί κάποιες από αυτές τις μεθόδους και μέσα από συγκρίσεις ανάμεσα σε σχετικές εργασίες αλλά και την παρουσίαση δύο σημαντικών πηγών —του *Κοινού Ευρωπαϊκού Πλαισίου Αναφοράς για τις Γλώσσες* και του *Προγράμματος Σπουδών του Ινστιτούτου Θερβάντες*— , να δικαιολογήσουμε τις θεωρητικές ιδιαιτεροτητες της δικής μας προσέγγισης. Στο δεύτερο κεφάλαιο, θέτουμε τους στόχους της εργασίας, περιγράφουμε τη συλλογή των δεδομένων μας και παρουσιάζουμε το σώμα κειμένων που προέκυψε, απαριθμούμε τα γλωσσικά χαρακηριστικά που χρησιμοποιήσαμε για την εκτίμηση της δυσκολίας των κειμένων και εξηγούμε τους λόγους για τους οποίους επιλέχθηκαν και τις μεθόδους που ακολουθήσαμε για τη μέτρησή τους και, τέλος, αναφέρουμε τα διάφορα εργαλεία που χρησιμοποιήσαμε. Στο τελευταίο κεφάλαιο, παρουσιάζουμε τα πειράματα που κάναμε, σχολιάζουμε τα αποτελέσματά τους, εκθέτουμε τα συμπεράσματα και τα τελικά προϊόντα της όλης διαδικασίας, και καταγράφουμε τις ιδέες μας για το ποια μπορούν να είναι τα επόμενα βήματα της έρευνάς μας.

2. Θεωρητικό υπόβαθρο και σχετικές εργασίες.

2.1 Πρώτα βήματα.

Σύμφωνα με τον DuBay (2004), θεμελιωτής του ερευνητικού κλάδου της αναγνωσιμότητας ήταν ο καθηγητής Αγγλικής Λογοτεχνίας στο Πανεπιστήμιο της Νεμπράσκα Lucius Adelno Sherman. Το 1880 υιοθέτησε μια ιστορική και στατιστική προσέγγιση στη διδασκαλία της λογοτεχνίας. Το 1983 εξέδωσε το βιβλίο του με τίτλο *Analytics of Literature, A Manual for the Objective Study of English Prose and Poetry* στο οποίο έδειξε ότι το μέσο μήκος πρότασης στην αγγλική λογότεχνία μίκραινε με τον χρόνο.

Το 1921 ο ψυχολόγος Edward Thorndike, αφού εξέτασε μια μεγάλη ποικιλία από βιβλία για ενήλικους και παιδιά, εφημερίδες και αλληλογραφία, ολοκλήρωσε την πρώτη μεγάλη καταγραφή της συχνότητας των λέξεων της αγγλικής την οποία δημοσίευσε στο βιβλίο του *The Teacher's Word Book* που περιλάμβανε 10.000 λέξεις. Παράλληλα έδειξε ότι οι συχνότερες λέξεις είναι και ευκολότερες στην κατανόηση από τις πιο σπάνιες και έθεσε τις βάσεις για τη χρησιμοποίηση λεξιλογικών χαρακτηριστικών στον υπολογισμό της αναγνωσιμότητας. Μετά από αυτό έγιναν εκτεταμένες έρευνες στο θέμα του λεξιλογίου. Ο ίδιος ο Thorndike το 1932 εξέδωσε το *A Teacher's Word Book of 20,000 Words* και το 1944 το *A Teacher's Word Book of 30,000 Words*. Κομβική ήταν και η έκδοση του *Human Behavior and The Principle of Least Effort* (1949) από τον καθηγητή του Harvard George Kingsley Zipf. Ο Zipf χρησιμοποίησε ποσοτική μεθοδολογία για να δείξει πώς η αρχή της ελάχιστης προσπάθειας λειτουργεί στον ανθρώπινο λόγο. Πιο συγκεκριμένα, μοντελοποίησε την σχέση που έχει η συχνότητα μιας λέξης με την σχετική σειρά κατάταξής της (rank), δείχνοντας ότι είναι μεγέθη αντιστρόφως ανάλογα. Παράλληλα, έδειξε ότι οι μεγαλύτερες λέξεις τείνουν να είναι πιο σπάνιες σε μια γλώσσα. Αυτή η έννοια της εξοικονόμησης ενέργειας είναι κεντρικό χαρακτηριστικό της γλώσσας και μια από τις βασικές αρχές της έρευνας της συχνότητας των λέξεων (DuBay, 2007). Ξεκινώντας από αυτήν την αφετηρία, καταλήξαμε πολύ αργότερα, στις αρχές του εικοστού πρώτου αιώνα, να αξιοποιούμε στην έρευνα της αναγνωσιμότητας ένα άλλο είδος λεξιλογικής πληροφορίας, τα στατιστικά γλωσσικά μοντέλα. Αυτά θα μας απασχολήσουν παρακάτω. Προς το παρόν μπορούμε να δούμε τον πρώτο καρπό που έδωσε η δουλειά του Thorndike. Αυτός είναι η πρώτη εξίσωση αναγνωσιμότητας.

2.2 Εξισώσεις αναγνωσιμότητας

Ο DuBay (2004) αναφέρει τους Bertha Lively και S. L. Pressey ως τους δημιουργούς της πρώτης εξίσωσης πρόβλεψης της αναγνωσιμότητας. Τους απασχολούσε το πρόβλημα της επιλογής επιστημονικών εγχειριδίων για το γυμνάσιο. Η εξίσωσή τους, δημοσιευμένη το 1923, λάμβανε υπόψη τον αριθμό διαφορετικών λέξεων ανά 1.000 λέξεις κειμένου και τον αριθμό λέξεων που δεν περιλαμβανόταν στη λίστα δέκα χιλιάδων λέξεων του Thorndike.

Ήταν η αρχή μιας μακράς διαδρομής. Μέχρι τη δεκαετία του 80 θα είχαν εμφανιστεί πάνω από 200 εξισώσεις και 1.000 μελέτες σχετικές με την αναγνωσιμότητα, με το ενδιαφέρον για τις εξισώσεις να κορυφώνεται την περίοδο 1978–1987, με μέσο όρο

17 δημοσιεύσεων κάθε χρόνο (Parker, Hasbrouck, & Weaver, 2001). Σύμφωνα με την ίδια πηγή, η τάση αυτή κάμφθηκε την τελευταία πενταετία του αιώνα με μόλις δύο, το πολύ, δημοσιεύσεις το χρόνο. Αξίζει τον κόπο να παρουσιάσουμε σύντομα κάποιες από τις πιο γνωστές και χρησιμοποιημένες εξισώσεις.

Μια εργασία ορόσημο ήταν αυτή των William S. Gray και Bernice Leary (Gray & Leary, 1935). Διερεύνησαν τους παράγοντες που κάνουν ένα βιβλίο κατάλληλο για ενήλικες με περιορισμένες αναγνωστικές δυνατότητες. Το δείγμα στο οποίο βασίστηκε η μελέτη τους περιλάμβανε 48 αποσπάσματα των εκατό λέξεων το καθένα. Ο βαθμός δυσκολίας τους καθορίστηκε μετά από ένα τεστ κατανόησης γραπτού λόγου, με τη συμμετοχή 800 ενηλίκων. Μετά, ταυτοποίησαν ένα σύνολο 228 χαρακτηριστικών τα οποία επηρεάζουν την αναγνωσιμότητα και τα κατέταξαν σε 4 κατηγορίες που αντιπροσώπευαν το περιεχόμενο, το ύφος, τη μορφοποίηση και την οργάνωση του κειμένου (κεφάλαια, επικεφαλίδες και λοιπά). Πρόκειται ίσως για την εργασία που έχει μελετήσει τα περισσότερα χαρακτηριστικά. Κατέληξαν στο ότι δεν μπορούσαν να μετρήσουν στατιστικά το περιεχόμενο, τη μορφοποίηση και την οργάνωση και κράτησαν 80 μεταβλητές σχετιζόμενες με το ύφος, από τις οποίες μπορούσαν να μετρήσουν αξιόπιστα τις 64. Έχοντας ένα μέτρο της δυσκολίας των κειμένων, έλεγξαν ποιες τιμές μεταβάλλονται όσο τα κείμενα γίνονται πιο δύσκολα και κράτησαν αυτές που παρουσίαζαν δείκτη συνάφειας μεγαλύτερο του 0,35. Δοκιμάζοντας διάφορους συνδυασμούς, κατέληξαν σε μια εξίσωση που έδινε δείκτη συνάφειας 0,645 με τις τιμές δυσκολίας που είχαν προκύψει από τα τεστ με τους ενήλικες. Αυτή η εξίσωση περιείχε τις εξής 5 μεταβλητές:

1. Μέσο μήκος πρότασης σε λέξεις.
2. Αριθμός διαφορετικών δύσκολων λέξεων[1].
3. Αριθμός αντωνυμιών πρώτου,δευτέρου και τρίτου προσώπου.
4. Ποσοστό διαφορετικών λέξεων.
5. Αριθμός προθετικών φράσεων.

Από τη δουλειά των Gray και Leary φάνηκε ότι εξισώσεις με πολλές μεταβλητές μπορεί να είναι οριακά, μόνο, πιο ακριβείς και είναι δυσανάλογα δυσκολότερες στην εφαρμογή από άλλες απλούστερες. Αυτό ώθησε τους ερευνητές να αναζητήσουν την τέλεια εξίσωση δοκιμάζοντας διαφορετικούς συνδυασμούς ολιγάριθμων μεταβλητών. Τελικά η έρευνα καταστάλαξε, σε μεγάλο βαθμό, στη χρήση μιας μεταβλητής σημασιολογικής φύσης, όπως η δυσκολία του λεξιλογίου, και άλλης μίας συντακτικής φύσης, για παράδειγμα το μέσο μήκος λέξης (DuBay, 2004).

Στα τέλη της δεκαετίας του 1940 δημοσιεύθηκε η εξίσωση Flesch Reading Ease (Flesh, 1948) η οποία έχει την παρακάτω μορφή:

Αναγνωστική Ευκολία = 206,835 – (1,015*ASL) – (0,846* ASW)

[1] Θεωρήθηκαν δύσκολες οι λέξεις που δεν ανήκουν στη λίστα των 769 «εύκολων» λέξεων που είχε συντάξει ο Dale, που με τη σειρά της προέκυψε από τον εντοπισμό των κοινών λέξεων στη λίστα λέξεων της Διεθνούς Ένωσης Νηπιαγωγίων (International Kindergarten Union) και στις 1000 συχνότερες λέξεις του Thorndike.

Όπου ASL το μέσο μήκος πρότασης μετρημένο σε λέξεις και ASW ο μέσος αριθμός συλλαβών ανά λέξη. Το αποτέλεσμα που δίνει είναι ένας αριθμός μεταξύ του 1 και του 100, με το 30 να αντιπροσωπεύει το «πολύ δύσκολο» και το 70 το «εύκολο».

Η εξίσωση του Flesch έγινε μια από τις πιο δοκιμασμένες και αξιόπιστες (Klare, 1963). Οι εκδότες ανακάλυψαν ότι μπορούσε να να αυξήσει το αναγνωστικό τους κοινό μέχρι και κάτα 60% και επίσης είχε μεγάλο αντίκτυπο στη δημοσιογραφία.

Αυτή όμως δεν ήταν η τελική μορφή της εξίσωσης. Τρία χρόνια αργότερα έγινε μια προσπάθεια απλοποίησής της (James N., James J., & Donald G., 1951) που είχε το παρακάτω αποτέλεσμα:

Αναγνωστική Ευκολία = 1.599 nosw – 1.015 sl – 31.517

Όπου nosw ο αριθμός μονοσύλλαβων λέξεων ανά 100 λέξεις και sl το μέσο μήκος πρότασης σε λέξεις.

Τέλος, το 1976, στα πλαίσια μιας μελέτης χρηματοδοτούμενης από το Ναυτικό των Η.Π.Α., η εξίσωση τροποποιήθηκε για άλλη μια φορά έτσι ώστε το αποτέλεσμά της να αντιστοιχεί στην τάξη του εκπαιδευτικού συστήματος των Η.Π.Α. που θα πρέπει να έχει τελειώσει κάποιος για να μπορεί να κατανοήσει ικανοποιητικά το κείμενο. Η νέα αυτή εξίσωση πήρε το όνομα Flesch-Kinkaid.

Την ίδια χρονία με τον Flesch, δημοσίευσαν μια εξίσωση οι Edgar Dale και Jeanne Chall (Dale & Chall, 1948), με σκοπό να διορθώσουν κάποια μειονεκτήματα της εξίσωσης του πρώτου. Ο Dale ήταν από τους πρώτους που άσκησαν κριτική στις λίστες του Thorndike και δημιούργησε τις δικές του. Η εξίσωσή Dale-Chall έχει αυτήν τη μορφή:

Τάξη = 0,1579*PDW + 0,0496*ASL + 3.6365

Όπου PDW το ποσοστό των «δύσκολων» λέξεων, δηλαδή, των λέξεων που δεν βρίσκονται στη λίστα του Dale με τις 3.000 λέξεις, το 80% των οποίων είναι γνωστές σε αναγνώστες επιπέδου τετάρτης τάξης. ASL είναι το μέσο μήκος πρότασης σε λέξεις. Το αποτέλεσμα είναι ένας αριθμός που μετατρέπεται στην τάξη ενός αναγνώστη ο οποίος μπορεί να απαντήσει σωστά τις μισές από τις ερωτήσεις κατανόησης του κειμένου[2].

Από όλες τις πρώιμες εξισώσεις, η Dale-Chall είναι αυτή που έχει δώσει τα πιο συνεπή και υψηλά αποτελέσματα, όσο αφορά τον δείκτη συνάφειάς της με τα

[2] Η τυπική διαδικασία ανάπτυξης μιας εξίσωσης αναγνωσιμότητας είναι να εργαστεί κανείς με έναν αριθμό κειμένων τα οποία έχουν ήδη βαθμονομηθεί με κάποιο τρόπο ως προς το επίπεδο δυσκολίας τους. Κάνοντας μετρήσεις πάνω σε αυτά τα κείμενα, προσδιορίζεται η επίδραση των επιλεγμένων μεταβλητών στην αναγνωσιμότητά τους. Ο τρόπος με τον οποίο γίνεται συνήθως η βαθμονόμηση είναι μέσα από τεστ κατανόησης στα οποία υποβάλλονται ομάδες ανθρώπων γνωστής αναγνωστικής ικανότητας. Οι συμμετέχοντες διαβάζουν κάθε κείμενο και απαντούν κάποιες ερωτήσεις σχετικές με αυτό. Ανάλογα με το ποσοστό σωστών απαντήσεων, καθορίζεται ο βαθμός δυσκολίας του κειμένου. Κατά την ανάπτυξη της εξίσωσης Dale-Chall χρησιμοποιήθηκαν τα ευρέως διαδεδομένα τεστ που είχαν δημιουργήσει το 1926 οι William A. McCall και Lelah Crabbs. Αυτά, αποτελούνταν από βαθμονομημένα κείμενα που τα ακολουθούσαν ερωτήσεις πολλαπλής επιλογής.

αποτελέσματα της βαθμονόμησης κειμένων μέσα από τεστ στα οποία συμμετέχουν άνθρωποι.

Ο Robert Gunning ήταν από τους πρώτους που εκμεταλλεύτηκε εμπορικά την έρευνα για την αναγνωσιμότητα. Το 1944 ίδρυσε μια εταιρεία που παρείχε συμβουλές πάνω σε θέματα αναγνωσιμότητας. Δημιούργησε την δικιά του εξίσωση με το όνομα Fog Index η οποία έγινε δημοφιλής λόγω της ευκολίας εφαρμογής της (Gunning, 1952). Χρησιμοποιεί δύο μεταβλητές, το μέσο μήκος πρότασης (ASL) και τον αριθμό λέξεων με πάνω από δύο συλλαβές ανά 100 λέξεις (HW):

Τάξη = 0,4*(ASL + HW)

Για τον προσδιορισμό της τάξης, ο Gunning έθεσε πιο αυστηρά κριτήρια από τους Dale και Chall βασιζόμενος σε ένα σκορ 90% σωστών απαντήσεων στα τεστ McCall-Crabbs.

Αυτός που βασίστηκε σε ένα ποσοστό 100% σωστών απαντήσεων στα τεστ McCall-Crabbs, ήταν ο Harry McLaughlin που ανέπτυξε την εξίσωση SMOG (McLaughlin, 1969). Συνέπεια αυτής της επιλογής ήταν, τα αποτέλεσματα που δίνει η SMOG να είναι κατά μεσό όρο δύο τάξεις ψηλότερα από αυτά που δίνει η Dale-Chall. Η εξίσωσή του, που παρουσίαζει δείκτη συνάφειας 0,88 με τα τεστ κατανόησης είναι η εξής:

SMOG = 3 + τετραγωνική ρίζα του αριθμού των λέξεων με πάνω από δύο συλλαβές ανά 30 προτάσεις.

Όλες οι εξισώσεις που παρουσιάσαμε, εκτός από την τελευταία, ανήκουν στην περίοδο που ο Dubay (2004) χαρακτηρίζει «κλασικές μελέτες αναγνωσιμότητας». Τα χρόνια που ακολούθησαν, το αντικείμενο της αναγνωσιμότητας διευρύνθηκε με μελέτες γύρω από την πρότερη γνώση, το ενδιαφέρον και το κίνητρο του αναγνώστη. Ερευνήθηκε η χρησιμότητα των εξισώσεων στη συγγραφή και προσαρμογή κειμένων, μελετήθηκε η αναγνωστική αποδοτικότητα (reading efficiency) και αναπτύχθηκαν οι δοκιμασίες συμπλήρωσης κενών (cloze tests). Οι τελευταίες αξιοποιήθηκαν στην δημιουργία των νεότερων εξισώσεων, των οποίων η βασική φιλοσοφία δεν άλλαξε. Σε μια επισκόπηση των χαρακτηριστικών που έχουν χρησιμοποιηθεί στις διάφορες εξισώσεις (Das & Roychoudhury, 2006) καταγράφηκαν τα ακόλουθα: (1) Μήκος λέξεων σε γράμματα. (2) Αριθμός λέξεων με πάνω από έξι γράμματα. (3) Αριθμός συλλαβών ανά κάποιες λέξεις. (4) Αριθμός μονοσύλλαβων λέξεων. (5) Αριθμός λέξεων με πάνω από τρεις συλλαβές. (6) Αριθμός προσφυμάτων. (7) Αριθμός λέξεων ανά πρόταση. (8) Αριθμός προτάσεων. (9) Αριθμός αντωνυμιών. (10) Αριθμός προθέσεων.

Παράλληλα με αυτές τις μελέτες, ξεκίνησε και η συζήτηση για την αποτελεσματικότητα και την αξία των εξισώσεων.

2.3 Κριτική στις εξισώσεις.

Μια έκδηλη αντίφαση των εξισώσεων αναγνωσιμότητας που εντόπισαν οι επικριτές τους, είναι τα διαφορετικά αποτελέσματα που δίνουν διαφορετικές εξισώσεις για το ίδιο κείμενο. Ενδεικτικά αναφέρουμε μια ανάλυση ενός τυχαίου κειμένου πέμπτης

τάξης κατά την οποία χρησιμοποιήθηκαν έξι εξισώσεις. Τα αποτελέσματα (σε τάξεις) ήταν τα ακόλουθα: Spache = 3,5, Corrected Dale-Chall = 7–8, Fry = 6, Raygor = 6, και Flesch = λιγότερο από 7. Ένα δεύτερο κείμενο του ίδιου επιπέδου έδωσε Spache = 3.2, Corrected Dale-Chall = 5–6, Fry = 4, Raygor = «μη έγκυρο» και Flesch = λιγότερο από 7 (Parker et al., 2001). Οι υπέρμαχοι των εξισώσεων όπως ο Dubay, θεωρούν ότι το σημαντικό δεν είναι να συμφωνούν τα αποτελέσματα σε ένα συγκεκριμένο κείμενο αλλά να υπάρχει συνέπεια στην πρόβλεψη της δυσκολίας ενός συνόλου κειμένων. Επιπλέον, απέδωσε αυτές τις διαφορές, μεταξύ άλλων, στην διαφορά των ελάχιστων ποσοστών σωστών απαντήσεων που χρησιμοποιήθηκαν στα τεστ κατανόησης, κατά την βαθμονόμηση των κειμένων αναφοράς, όταν αναπτυσόταν η κάθε εξίσωση[3]. Για παράδειγμα, η FORCAST και η Dale-Chall χρησιμοποιούν ένα ελάχιστο ποσοστό 50% ενώ η Fog 90% και η SMOG 100% (DuBay, 2004). Το επιχείρημα αυτό δεν φαίνεται πολύ πειστικό επειδή οδηγεί στην σκέψη ότι ο ορισμός αυτών των ελάχιστων ποσοστών είναι αυθαίρετος. Αυτή η παρατήρηση φέρνει στην επιφάνεια το γενικότερο πρόβλημα των κριτηρίων επικύρωσης των εξισώσεων. Γράφει, πάλι, ο Parker ότι πέρα από τα κείμενα αναφοράς, τα οποία βαθμονομήθηκαν στις περισσότερες περιπτώσεις είτε με McCall-Crabbs Standard Test ή με δοκιμασίες συμπλήρωσης κενών, σπάνια γίνεται διασταύρωση των αποτελεσμάτων των εξισώσεων με κείμενα που χρησιμοποιούνται αργότερα στην διδακτική πράξη. Επίσης, έχει σχολιαστεί (Schriven, 2000) το γεγονός ότι τα κείμενα που θα κληθεί να αξιολογήσει μια εξίσωση, μπορεί να μην έχουν καμία ομοιότητα με τα κείμενα αναφοράς που χρησιμοποιήθηκαν στην ανάπτυξή της με αποτέλεσμα οι μαθηματικές σχέσεις που αποτυπώνει να μην έχουν γενική αξία.

Έντονη κριτική έχει ασκηθεί και στα χαρακτηριστικά που χρησιμοποιούν οι εξισώσεις. Έχουν κατηγορηθεί ότι χρησιμοποιούν μονο επιφανειακά χαρακτηριστικά, όπως το μήκος λέξης και πρότασης, και αγνοούν τις βαθύτερες συντακτικές και γραμματικές δομές και το εννοιολογικό φορτίο (Contreras, García-Alonso, Echenique, & Daye-Contreras, 1999; Parker et al., 2001). Ουσιαστικά, αξιοποιούν μόνο ό,τι μπορεί να μετρηθεί εύκολα και αφήνουν στην άκρη άλλα χαρακτηριστικά, ακόμα και αν γνωρίζουν ότι αυτά επηρεάζουν την αναγνωσιμότητα (Redish, 2000). Πάνω σε αυτό, ο Schriven θεωρεί ότι τα χαρακτηριστικά που χρησιμοποιούνται δεν είναι αυτά που σχετίζονται περισσότερο με την κατανόηση του κειμένου και για αυτό, οι εξισώσεις δεν μετράν καν αυτό που υποτίθεται ότι μετράν.

Άλλος ένας προβληματικός τομέας εντοπίστηκε στις εξισώσεις που χρησιμοποιούν λίστες λέξεων. Πολλές από αυτές χρησιμοποιούν λίστες που δημιουργήθηκαν πριν από πολλά χρόνια, όπως αυτή του Dale, και για να βελτιώσουν τα αποτελέσματά τους, ενημερώνουν τις λίστες αυθαίρετα με λέξεις που θεωρούν ότι ξέρει το κοινό στο οποίο απευθύνονται. Ακόμα, αγνοούν ότι οι λέξεις μπορεί να έχουν περισσότερα από ένα νοήματα (Redish, 2000). Πέρα από αυτό, έχει γίνει η παρατήρηση ότι σε εξειδικευμένα θεματικά κείμενα, που μπορέι να έχουν εξαιρετικά απλή δομή και να θεωρούνται πολύ εύκολα από τους γνώστες του

[3] Θυμίζουμε ότι το επίπεδο κατανόησης που σηματοδοτεί την επιτυχία της ανάγνωσης, καθορίζεται από έναν ελάχιστο αριθμό σωστών απαντήσεων σε ερωτήσεις πάνω στα κείμενα αναφοράς.

θέματος, θα προκύψουν μεγάλες αποκλίσεις μεταξύ της πραγματικής αναγνωσιμότητας και τις εκτιμούμενης από εξισώσεις που χρησιμοποιούν λίστες λέξεων (Petersen & Ostendorf, 2009).

Η τελευταία διαπίστωση θίγει και το θέμα της ακαταλληλότητας των εξισώσεων όταν καλούνται να εκτιμήσουν την αναγνωσιμότητα μη παραδοσιακών ειδών κειμένων. Για παράδειγμα, οι εξισώσεις δεν μπορούν να χρησιμοποιηθούν σε φόρμες, ιστοσελίδες ή κείμενα με πολλές λίστες (Collins-Thompson & Callan, 2005; Redish, 2000). Οι Collins-Thompson και Callan (2005) εξειδικεύουν τα προβλήματα που παρουσιάζουν οι ιστοσελίδες. Θεωρούν ότι είναι χαρακτηριστικό τους οι μικρότερες προτάσεις, οι οποίες πολλές φορές είναι δυνατό να μην υπάρχουν καν με τη μορφή που τις συναντάμε σε ένα παραδοσιακό κείμενο, μιας και στις ιστοσελίδες πολλές φορές η πληροφορία παρουσιάζεται σε πίνακες και λίστες. Υπερσύνδεσμοι και διευθύνσεις ηλεκτρονικού ταχυδρομείου μπορούν να δημιουργήσουν περαιτέρω προβλήματα στον εντοπισμό των ορίων των προτάσεων. Επισημαίνουν, ακόμα, το πρόβλημα του θορύβου, λόγω των μενού πλοήγησης για παράδειγμα.

Η ίδια η κλίμακα μέτρησης της αναγνωσιμότητας που χρησιμοποιούν οι περισσότερες εξισώσεις, δηλαδή αυτή που βασίζεται στις τάξεις του εκπαιδευτικού συστήματος των Η.Π.Α., έχει γίνει αντικείμενο κριτικής. Ως προς το τι αντιπροσωπεύει, η Redish (2000) γράφει ότι οι έρευνες στις οποίες βασίστηκαν αυτές οι εξισώσεις είναι πολύ παλιές και ότι το αναγνωστικό επίπεδο των παιδιών του σχολείου, δεν είναι πλέον το ίδιο με πριν από 50 χρόνια. Επισημαίνει, επίσης, ότι οι εξισώσεις δεν διακρίνουν ανάμεσα σε διαφορετικές κατηγορίες αναγνωστών. Τι νόημα έχει άραγε η κατηγοριοποίηση σε ένα επίπεδο-τάξη του αμερικάνικου σχολείου, ενός κειμένου το οποίο προορίζεται για άτομα που μαθαίνουν τα αγγλικά ως ξένη γλώσσα; Σε τέτοιες καταστάσεις, πώς λαμβάνεται υπόψη η πολιτισμική αλληλεπίδραση μεταξύ κειμένου και αναγνώστη, τόσο σε επίπεδο περιεχομένου όσο και σε επίπεδο δομής; Για αυτόν το λόγο, έχει επικριθεί ιδιαιτέρως η επέκταση της χρήσης εξισώσεων στην ανάλυση κειμένων που προορίζονται για την διδασκαλία της γλώσσας ως ξένης (Parker et al., 2001).

Πολλά από αυτά τα ελλατώματα απαλείφθηκαν με τις νεότερες προσεγγίσεις του θέματος της αναγνωσιμότητας που προέκυψαν από τις προόδους στην Επεξεργασία Φυσικού Λόγου και άρχισαν να βλέπουν το φως με την αλλαγή του αιώνα.

2.4 Επεξεργασία Φυσικής Γλώσσας και Μηχανική Μάθηση στην αναγνωσιμότητα.

Η νέα προσέγγιση στην αναγνωσιμότητα ήρθε από τους χώρους της Επεξεργασιας Φυσικής Γλώσσας και της μηχανικής μάθησης. Το πρόβλημα της αναγνωσιμότητας άρχισε να αντιμετωπίζεται ως πρόβλημα κατηγοριοποίησης κειμένου, όπως η αναγνώριση θέματος, κειμενικού είδους ή συγγραφέα. Οι τεχνικές της μηχανικής μάθησης παραμένουν ίδιες. Αυτό που μπορεί να αλλάξει είναι τα χρησιμοποιούμενα χαρακτηριστικά (Petersen & Ostendorf, 2009).

Η πρώτη εργασία (Si & Callan, 2001) που αξιοποίησε αυτές τις τεχνικές είχε σκοπό να αντιμετωπίσει τις ιδιαιτερότητες των ιστοσελίδων που παρουσιάσαμε νωρίτερα.

Χρησιμοποιήθηκε ένα σώμα 91 κειμένων που συλλέχθηκε από ιστοσελίδες γύρω από θέματα του μαθήματος της Επιστήμης[4]. Τα κείμενα είτε είχαν γραφεί από μαθητές διαφόρων τάξεων, είτε είχαν γραφεί από ενήλικους με αναφορά όμως του επιπέδου των μαθητών στους οποίους απευθύνονταν. Τα κείμενα μοιράστηκαν σε τρεις κατηγορίες ανάλογα με τις τάξεις του αμερικάνικου σχολείου για τις οποίες ήταν γραμμένα: μέχρι δευτέρα, από τρίτη μέχρι πέμπτη, και από έκτη μέχρι ογδόη. Στην συνέχεια δημιούργησαν τρία γλωσσικά μοντέλα λέξεων (unigrams) που αντιστοιχούσαν στις προαναφερθείσες κατηγορίες. Το σώμα εκπαίδευσης των μοντέλων συλλέχθηκε από αναλυτικά προγράμματα των μαθημάτων της Επιστήμης και των Μαθηματικών. Δέκα κείμενα από κάθε κατηγορία χρησιμοποιήθηκαν για την εκπαίδευση ενός ταξινομητή με την βοήθεια του αλγόριθμου Προσδοκίας-Μεγιστοποίησης (ΕΜ). Σαν χαρακτηριστικά χρησιμοποίηθηκαν η κατηγορία που είχε προβλέψει για κάθε κείμενο ένας Naive Bayes ταξινομητής με βάση τα λεξικά μοντέλα και το μέσο μήκος πρότασης κάθε κειμένου. Το μοντέλο δοκιμάστηκε στα υπόλοιπα 61 κείμενα και πέτυχε ορθότητα της τάξης του 75,4%.

Σε μια συνέχεια αυτής της προσπάθειας, οι Collins-Thompson και Callan (2005) προσπάθησαν να κατηγοριοποιήσουν κείμενα από τον Παγκόσμιο Ιστό σε μορφή HTML, PDF και απλού κειμένου. Το corpus τους περιείχε 550 κείμενα βαθμονομημένα από τους συγγραφείς τους στην κλίμακα των τάξεων από 1-12. Η μέθοδος που ακολουθήθηκε αυτή τη φορά ήταν πολύ πιο εξελιγμένη. Εκπαίδευσαν 12 γλωσσικά μοντέλα λέξεων από το corpus τους με εξομάλυνση Good-Turing. Επιπλέον, δεν περιορίστηκαν στο να αποδώσουν ίσες πιθανότητες σε κάθε λέξη που δεν εμφανιζόταν σε κάποιο μοντέλο, αλλά πήραν πληροφορίες από τα μοντέλα των γειτονικών κλάσεων. Για την ταξινόμηση χρησιμοποίησαν μια παραλλαγή του πολυωνυμικού Naive Bayes ταξινομητή. Χώρισαν κάθε κείμενο σε αποσπάσματα των 100 λέξεων και αξιολόγησαν το καθένα ξεχωριστά. Διέταξαν τις προβλέψεις από την μικρότερη στη μεγαλύτερη και επέλεξαν για το συνολικό κείμενο αυτήν που βρισκόταν στη θέση του τρίτου τεταρτημορίου, βασιζόμενοι σε προηγούμενες μελέτες που υποστήριζαν ότι ο στόχος είναι ένας βαθμός 75% κατανόησης του κειμένου. Ακολουθώντας αυτήν τη μέθοδο, πέτυχαν συνάφεια 0,79 με τις δηλωμένες από τους συγγραφείς τάξεις των κειμένων, καλύτερα, δηλαδή, από παραδοσιακές μεθόδους όπως η εξίσωση Flesch-Kincaid που παρουσίασε συνάφεια 0,47.

Στατιστικά γλωσσικά μοντέλα υψηλότερης τάξης —διγραμμάτων και τριγραμμάτων— χρησιμοποιήθηκαν και από άλλους ερευνητές (Petersen & Ostendorf, 2009; Schwarm & Ostendorf, 2005). Οι Schwarm και Ostendorf (2005) δούλεψαν με ένα σώμα κειμένων από τη Weekly Reader, μια εκπαιδευτική εφημερίδα με εκδόσεις που στοχεύουν σε μαθητές διαφορέτικών τάξεων. Πιο συγκεκριμένα, μάζεψαν κείμενα για τις τάξεις από δευτέρα ως πέμπτη. Επιπλέον, εργάστηκαν με δύο ακόμα σώματα κειμένων, τα οποία αξιοποίησαν στην εκπαίδευση των γλωσσικών μοντέλων. Το πρώτο, το αποτελούσαν άρθρα από την κανονική έκδοση της εγκυκλοπαίδειας Britannica και τα αντίστοιχα άρθρα από την έκδοσή της για παιδιά. Το δεύτερο περιείχε ρεπορτάζ από το CNN στην πλήρη και

[4] Όπως αυτό εννοείται στα πλαίσια του εκπαιδευτικού συστήματος των Η.Π.Α. Science στα αγγλικά.

την συντομευμένη τους μορφή. Θεωρώντας ότι τα μοντέλα λέξεων δεν είναι κατάλληλα για να καταγράψουν, πέρα απο σημασιολογική πληροφορία, και συντακτική, προχώρησαν στην εκπαίδευση 12 γλωσσικών μοντέλων. Ένα μοντέλο λεξικών μονογραμμάτων, ένα διγραμμάτων και ένα τριγραμμάτων για κάθε μία από τις τέσσερεις κατηγορίες του corpus τους. Την τιμή της περιπλοκής (perplexity) κάθε κειμένου σε σχέση με κάθε μοντέλο, την χρησιμοποίησαν μαζί με άλλα χαρακτηριστικά στην εκπαίδευση ενός ταξινομητή μηχανής διανυσμάτων υποστήριξης (Support Vector Machine - SVM). Τα υπόλοιπα χαρακτηριστικά ήταν το μέσο μήκος πρότασης, ο μέσος αριθμός συλλαβών ανά λέξη, το αποτέλεσμα της εξίσωσης Flesch-Kincaid για το κείμενο, έξι μετρήσεις σχετιζόμενες με τις 100, 200 και 300 συχνότερες λέξεις του κατώτερου επιπέδου (δευτέρα τάξη) και, τέλος, τέσσερα χαρακτηριστικά που προέκυψαν από την συντακτική ανάλυση των κειμένων και την αναπαράστασή της με δέντρα συστατικής δομής (μέσο ύψος δέντρου, μέσοι αριθμοί ονοματικών φράσεων, ρηματικών φράσεων και SBAR ανά πρόταση). Οι αρμονικοί μέσοι (f measure) της ανάκλησης (recall) και της ορθότητας (precision) που πέτυχε ο ταξινομητής για κάθε επίπεδο ήταν: $2^{α}$ τάξη = 0,47, $3^{η}$ τάξη = 0,53, $4^{η}$ τάξη = 0,65 και $5^{η}$ τάξη = 0,77. Επιπλέον, μέτρησαν τα ποσοστά των λανθασμένα ταξινομημένων κειμένων σε τάξεις που να μη συνορεύουν με τις σωστές, δηλαδή τις περιπτώσεις εκείνες όπου μεταξύ της πραγματικής και της προβλεφθείσας τάξης μεσολαβεί και άλλη, και έδειξαν ότι ο ταξινομητής μηχανής διανυσμάτων υποστήριξης υπερτερεί σημαντικά σε σχέση με παραδοσιακές μεθόδους μέτρησης της αναγνωσιμότητας όπως οι εξισώσεις Flesch-Kincaid και Lexile.

Η μεθοδολογία της παραπάνω εργασίας αποτυπώνει τα τυπικά βήματα που έχουν ακολουθήσει πολλές σχετικές έρευνες (Brück, Hartrumpf, Helbig, & Hagen, 2008; Feng, Jansche, Huenerfauth, & Elhadad, 2010; Kate et al., 2010; Larsson, 2006; Pitler & Nenkova, 2008; Wang, 2006; Μικρός, n.d.) και τα οποία μπορούμε να συνοψίσουμε ως εξής:

1. Δημιουργία ενός σώματος κειμένων εκπαίδευσης, κατάλληλο για τους στόχους της έρευνας. Τα κείμενα αυτά είναι ταξινομημένα στα επίπεδα αναγνωσιμότητας που μας ενδιαφέρουν, είτε από τους συγγραφείς τους, είτε μέσα από κάποια διαδικασία στην οποία συμμετέχουν άνθρωποι ως κριτές, κατά την διάρκεια της έρευνας.
2. Επιλογή των προς διερεύνηση γλωσσικών χαρακτηριστικών του κειμένου που εικάζεται ότι επηρεάζουν την αναγνωσιμότητά του. Η ομαδοποίηση αυτών των χαρακτηριστικών διαφέρει από συγγραφέα σε συγγραφέα. Μια κατηγορία χαρακτηριστικών που αναφέρεται συχνά είναι τα συντακτικά χαρακτηριστικά, τα οποία προκύπτουν από την συντακτική επεξεργασία (parsing) του κειμένου και την απεικόνιση της δομής του με συντακτικά δέντρα. Πολλοί συγγραφείς αναφέρονται σε ρηχά (shallow) και επιφανειακά (surface) χαρακτηριστικά. Με αυτούς τους ορους, περιγράφουν χαρακτηριστικά που έχουν χρησιμοποιηθεί στις εξισώσεις αναγνωσιμότητας (μέσο μήκος πρότασης, αριθμός πολυσύλλαβων λέξεων κ.α.), αλλά και αποτελέσματα που έχει δώσει η εφαρμογή εξισώσεων στα κείμενα. Η επιλογή αυτών των όρων δικαιολογείται από το ότι για την μέτρηση αυτών των χαρακτηριστικών, δεν απαιτείται βαθύτερη μορφολογική, συντακτική ή σημασιολογική ανάλυση των κειμένων. Πιο συγκεχυμένες, και συχνά

αλληλοεπικαλυπτόμενες, κατηγορίες είναι αυτές των σημασιολογικών και λεξιλογικών χαρακτηριστικών, στις οποίες αρκετοί συγγραφείς κατατάσσουν και αυτά που προκύπτουν από τη χρήση γλωσσικών μοντέλων. Στις περισσότερες περιπτώσεις χρησιμοποίησης των τελευταίων, η εκπαίδευσή τους γίνεται με ένα σώμα κειμένων ξεχωριστό από το σώμα εκπαίδευσης του ταξινομητή.

3. Πειραματισμός με κάποιες μεθόδους παλινδρόμησης ή μηχανικής μάθησης και με διαφορετικούς συνδυασμούς χαρακτηριστικών, με σκοπό την δημιουργία ενός ταξινομητή που να δίνει ικανοποιητικά αποτελέσματα.
4. Αξιολόγηση της αποτελεσματικότητας του ή των ταξινομητών με κείμενα που δεν συμμετείχαν στην εκπαίδευσή τους ή με άλλες τεχνικές όπως η k-fold cross validation.

Θα έχουμε ξανά την ευκαιρία να αναφερθούμε σε κάποια από τα παραπάνω σημεία και σε συγκεκριμένες εργασίες, κατά την ανάπτυξη της μεθοδολογίας της παρούσας εργασίας, η οποία κινείται στα πλαίσια που μόλις παρουσιάσαμε.

2.5 Αναγνωσιμότητα και ξένες γλώσσες.

Η αναφορά που έχουμε κάνει μέχρι στιγμής στις διάφορες μεθόδους εκτίμησης της αναγνωσιμότητας δεν έχει επεκταθεί στις ιδιαιτερότητες που εμφανίζονται όταν το πρόβλημα αντιμετωπίζεται από την οπτική γωνιά της ξένης γλώσσας. Ανατρέχοντας στη σχετική βιβλιογραφία, διαπιστώνουμε, καταρχάς, ότι οι περισσότερες έρευνες, που επιπλέον θεμελίωσαν τον κλάδο, έχουν γίνει για τα αγγλικά. Η ένταξη στο εκπαιδευτικό σύστημα των Η.Π.Α. ατόμων με περιορισμένη γνώση της αγγλικής, τα τελευταία χρόνια κυρίως ισπανόφωνων, και η ανάγκη εξεύρεσης κατάλληλων αναγνωσμάτων που να βοηθούν την γλωσσική τους εξέλιξη, έχουν παρουσιαστεί ώς αφορμή για αρκετές εργασίες, ήδη από την εποχή των εξισώσεων. Στις έρευνες αυτές, όπως αυτή των Schwarm και Ostendorf (2005), που παρουσιάσαμε στην προηγούμενη ενότητα, η αντιμετώπιση του προβλήματος δεν διαφέρει από τις περιπτώσεις που τα κείμενα προς ταξινόμηση απευθύνονται σε φυσικούς ομιλητές της αγγλικής ή της εκάστοτε γλώσσας. Αυτή η μη διαφοροποίηση παρατηρείται και στα χαρακτηριστικά που χρησιμοποιούνται και στα θέματα των κειμένων αλλά και στην κλίμακα στην οποία κατατάσσονται αυτά. Αυτή η προσέγγιση δεν είναι προβληματική όταν απευθύνεται σε ένα κοινό μη φυσικών ομιλητών που όμως κινείται σε ένα περιβάλλον εμβύθισης (submersion) στη γλώσσα. Ας πάρουμε για παράδειγμα ένα παιδί ισπανόφωνων γονιών που ζει στην Αμερική, που διδάσκεται τα μαθήματα του σχολείου στα αγγλικά, που οι συμμαθητές του είναι φυσικοί ομιλητές της αγγλικής και με τους οποίους έχει κοινές κοινωνικοπολιτισμικές αναφορές και που στις καθημερινές του συναναστροφές χρειάζεται να επιτελέσει πολυάριθμες γλωσσικές λειτουργίες στα αγγλικά. Ο τρόπος που μαθαίνει τη γλώσσα, η ανάγκες του και οι ελλείψεις του είναι πολύ διαφορετικές από αυτές ενός Ισπανού ενήλικα, λόγου χάρη, που μαθαίνει αγγλικά στη χώρα του. Για αυτό το παιδί, το να βρεθεί ένα κείμενο σχετικό με θέματα βιολογίας που διδάσκεται στο σχολείο που να μπορεί να το καταλάβει ή το να έχει πρόσβαση σε λογοτεχνικά κείμενα που αντιστοιχούν στο επίπεδό του, βαθμονομημένα με τάξεις του αμερικάνικου σχολείου, και που θα βοηθήσουν την γλωσσική του εξέλιξη, έχει μεγάλη αξία. Τι αξία θα είχαν όμως για τον Ισπανό ενήλικα τέτοια κείμενα; Αποτελεί προτεραιότητά του η ανάγνωση κειμένων με θέμα τη βιολογία; Ακόμα και

λογοτεχνικά ή ενημερωτικά κείμενα, τι νόημα έχει για αυτόν να κατατάσσονται σε ταξεις του αμερικάνικου εκπαιδευτικού συστήματος; Οι γραμματικές του γνώσεις, σε μια δεδομένη στιγμή, μπορούν να είναι συγκρίσιμες με αυτές των μαθητών κάποιας τάξης του αμερικάνικου σχολείου; Διαθέτει τις πολιτισμικές αναφορές που θα του επιτρέψουν να κατανοήσει ένα κείμενο το οποίο κατά τα άλλα είναι απλό και απευθύνεται, ας πούμε, σε φυσικούς ομιλητές της 5ης τάξης; Οι δύσκολες λέξεις για τους φυσικούς ομιλητές παρουσιάζουν την ίδια δυσκολία και για αυτόν; Οι παραπάνω ερωτήσεις είναι περισσότερο ρητορικές. Ας δούμε τι έχει γραφεί σχετικά με κάποιες από αυτές.

Έχει επισημανθεί ότι στην περίπτωση της μητρικής γλώσσας, της οποία η απόκτηση αρχίζει πολύ νωρίς στη ζωή ενός ανθρώπου, οι κύριες γραμματικές δομές έχουν κατακτηθεί ήδη στην ηλικία των τεσσάρων ετών, σε παιδιά με φυσιολογική ανάπτυξη. Δηλαδή, το μεγαλύτερο μέρος της γραμματικής κατακτιέται πριν καν το παιδί πάει σχολείο. Αντιθέτως, οι σπουδαστές ξένων γλωσσών συνεχίζουν να μαθαίνουν καινούριες γραμματικές δομές της γλώσσας στόχου ακόμα και σε προχωρημένα επίπεδα (Heilman, Collins-Thompson, Callan, & Eskenazi, 2007). Επιπλέον, κατά τη διδασκαλία των ξένων γλωσσών, υπάρχει συνήθως μια σειρά με την οποία παρουσιάζονται τα διάφορα γραμματικά φαινόμενα. Σε αυτά τα δεδομένα στήριξαν οι Heilman et al. (2007) την υπόθεσή τους ότι τα γραμματικά χαρακτηριστικά μπορεί να παίζουν σημαντικό ρόλο στην αναγνωσιμότητα, όταν αυτή αφορά κείμενα που χρησιμοποιούνται στη διδασκαλία μιας ξένης γλώσσας. Δουλεύοντας με δύο σώματα κειμένων, ένα για φυσικούς ομιλητές — αυτό που είχαν χρησιμοποιήσει και οι Collins-Thompson και Callan (2005) — και ένα με κείμενα από εγχειρίδια διδασκαλίας αγγλικών, με τη βοήθεια ενός αλγόριθμου K-πλησιέστερων γειτόνων (k-nearest neighbors), πέτυχαν μεγαλύτερη συνάφεια μεταξύ προβλέψεων και πραγματικών τιμών στο δεύτερο σώμα, όταν χρησιμοποίησαν αποκλειστικά γραμματικά χαρακτηριστικά.

Σε σχέση με το λεξιλόγιο, το γεγονός ότι μια λέξη θεωρείται δύσκολη στις έρευνες για την αναγνωσιμότητα στην πρώτη γλώσσα (L1), είτε γιατί είναι πολυσύλλαβη είτε γιατί δεν συμπεριλαμβάνεται στις λίστες με τις συχνότερες λέξεις της γλώσσας, δεν σημαίνει ότι θα δυσκολέψει το ίδιο και ένα ομιλιτή της ίδιας γλώσσας ως δεύτερης (L2). Σαν παράδειγμα μπορούμε να αναφέρουμε ονόματα δεινοσαύρων (Petersen & Ostendorf, 2009) ή ιατρικούς όρους οι οποίοι συνήθως είναι παρόμοιοι σε πολλές γλώσσες και είναι εξαιρετικά πιθανό να μπορέσει να τους καταλάβει ο ομιλητής της L2 με βάση τη γνώση της μητρικής του γλώσσας.

Όσο αφορά τα θέματα και κατ΄επέκταση τα κειμενικά είδη που μπορούν να αφορούν έναν ομιλητή L2, οι φορείς που είναι αρμόδιοι για τη διδασκαλία μιας γλώσσας σαν ξένης, στα προγράμματα σπουδών που καταρτίζουν συμπεριλαμβάνουν και τα κειμενικά είδη που θα πρέπει να μπορεί να κατανοεί ή και να παράγει ο μαθητής κάθε επιπέδου (Instituto Cervantes, 2006). Για να επανέλθουμε στο παράδειγμα κειμένων σχετικών με το μάθημα της Βιολογίας, του γυμνασίου λόγου χάρη, τέτοια κείμενα θα αντιμετωπίζονταν ως εξειδικευμένα και θα απευθύνονταν σε μαθητές L2 των υψηλότερων επιπέδων.

Συνοψίζοντας τα παραπάνω, θεωρούμε ότι μια εργασία σαν την παρούσα, που αποσκοπεί στην ταξινόμηση κειμένων με βάση την αναγνωσιμότητα και αντιμετωπίζει την γλώσσα τους ως δεύτερη, πρέπει να έχει τα ακόλουθα χαρακτηριστικά:

1. Η κλίμακα αξιολόγησης που χρησιμοποιεί πρέπει να έχει νόημα για άτομα που μαθαίνουν την γλώσσα. Τα επίπεδα που την διαρθρώνουν πρέπει να αποτυπώνουν την πορεία αυτών των ατόμων προς την κατάκτηση της γλώσσας και να αντιστοιχούν σε κάποια ορόσημα αυτής της πορείας.
2. Σχετικά με τα χαρακτηριστικά που συμμετέχουν στην εκτίμηση της αναγνωσιμότητας, όταν είναι εφικτό, να χρησιμοποιούνται κάποια που να αντικατοπτρίζουν την διάρθρωση τμημάτων της διδακτέας ύλης στα επίπεδα της κλίμακας αξιολόγησης.

Οι δύο προϋποθέσεις που θέσαμε, δεν θα μπορούσαν να εκπληρωθούν στην έρευνά μας, αν δεν υπήρχαν δύο πολύ σημαντικές πηγές. Το *Κοινό Ευρωπαϊκό Πλαίσιο Αναφοράς για τις Γλώσσες* (Council of Europe, 2001) και το *Πρόγραμμα Σπουδών του Ινστιτούτου Θερβάντες* (Instituto Cervantes, 2006).

2.6 Κοινό Ευρωπαϊκό Πλαίσιο Αναφοράς για τις Γλώσσες και Κοινά Επίπεδα Αναφοράς.

Ήδη από την δεκαετία του 1970, είχε διαπιστωθεί η ανάγκη ύπαρξης ενός συστήματος διδασκαλίας, εφαρμόσιμου σε όλες τις ευρωπαϊκές γλώσσες, το οποίο να είναι βασισμένο στην ανάλυση των ατομικών αναγκών των μαθητών και σε πραγματικές επικοινωνιακές καταστάσεις. Υπό την αιγίδα του Συμβουλίου της Ευρώπης δημοσιεύθηκε το 1975 το *Threshold-Level*, στα ελληνικά Επίπεδο-Κατώφλι. Στην πρώτη του έκδοση περιέγραφε τα γλωσσικά εφόδια που θα έπρεπε να διαθέτει ο μαθητής, για να μπορεί να ανταποκριθεί, στα αγγλικά, σε καταστάσεις της καθημερινότητας οι οποίες δεν θα έκρυβαν ιδιαίτερες δυσκολίες. Κατά τη διάρκεια της ίδιας δεκαετίας και της επόμενης, το *Threshold-Level* προσαρμόστηκε και για άλλες γλώσσες και αναπτύχθηκαν περιγραφές και για άλλα επίπεδα, με τον συνολικό τους αριθμό να φτάνει τα έξι (Breakthrough, Waystage, Threshold, Vantage, Effective Operational Proficiency, Mastery). Το 1991, πραγματοποιήθηκε στο Rüschlikon της Ελβετίας ένα διακυβερνητικό συμπόσιο με τίτλο «Διαφάνεια και συνοχή στη γλωσσική εκμάθηση στην Ευρώπη: Στόχοι, αξιολόγηση, πιστοποίηση». Ανάμεσα στα συμπεράσματα του συμποσίου ήταν και η ανάγκη δημιουργίας ενός κοινού ευρωπαϊκού πλαισίου αναφοράς για τη γλωσσική εκμάθηση σε όλα τα επίπεδα που θα είχε ως στόχους «να προωθήσει και να διευκολύνει τη συνεργασία ανάμεσα στα εκπαιδευτικά ιδρύματα διαφορετικών χωρών, να αποτελέσει μια στέρεη βάση για την αμοιβαία αναγνώριση τίτλων γλωσσομάθειας και να βοηθήσει τους μαθητές, τους διδάσκοντες, τους συντάκτες προγραμμάτων σπουδών, τις εξεταστικές αρχές και τους υπεύθυνους εκπαίδευσης να προσδιορίσουν και να συντονίσουν τις προσπάθειές τους» (Council of Europe, 2001). Εκπληρώντας αυτήν τη δέσμευση, το Συμβούλιο της Ευρώπης δημοσίευσε 10 χρόνια αργότερα το *Κοινό Ευρωπαϊκό Πλαίσιο Αναφοράς για τις Γλώσσες* (ΚΕΠΑΓ).

Από τις πιο σημαντικές συνεισφορές του ΚΕΠΑΓ ήταν τα έξι Κοινά Επίπεδα Αναφοράς που ορίζουν την πρόοδο κατά την εκμάθηση μιας ξένης γλώσσας,

περιγράφοντας τι είναι ικανός να κάνει ο μαθητής μεσω της χρήσης της γλώσσας. Πρόκειται για τρία ευρύτερα στάδια Α (βασικός χρήστης),Β (ανεξάρτητος χρήστης) και Γ (ικανός χρήστης), το καθένα από τα οποία υποδιαιρείται σε δύο επίπεδα, Α1-Α2, Β1-Β2 και Γ1-Γ2. Αυτός ο ορισμός των επιπέδων αποτελεί την κάθετη διάσταση του ΚΕΠΑΓ. Στο Παράρτημα 1 παραθέτουμε έναν πίνακα με τις γενικές περιγραφές των επιπέδων (Πίνακας 6) και το σχετικό με την κατανόηση του γραπτού λόγου πλέγμα αυτοαξιολόγησης (Πίνακας 7).

Εκτός από τα Κοινά Επίπεδα Αναφοράς, το ΚΕΠΑΓ εισήγαγε και και μια σειρά από περιγραφικές κατηγορίες της χρήσης της γλώσσας και της ικανότητας του μαθητή να την χρησιμοποιήσει: το περικείμενο της χρήσης της γλώσσας, τα θέματα της επικοινωνίας, τα επικοινωνιακά καθήκοντα και τους επικοινωνιακούς σκοπούς, τις επικοινωνιακές γλωσσικές δραστηριότητες και στρατηγικές, τις επικοινωνιακές γλωσσικές διαδικασίες και, τέλος, τα κείμενα. Η κεντρική ιδέα είναι ότι ο μαθητής πρέπει πραγματοποιήσει ορισμένες γλωσσικές δραστηριότητες, οι οποίες σχετίζονται με συγκεκριμένες ικανοτητες και προϋποθέτουν τη χρήση κάποιων επικοινωνιακών στρατηγικών. Αυτές οι κατηγορίες αποτελούν την οριζόντια διάσταση του ΚΕΠΑΓ. Από τον συγκερασμό των δύο διαστάσεων, προέκυψαν κλίμακες ενδεικτικών περιγραφητών των κατηγοριών που αναφέραμε, οι οποίες διαρθρώνονται στα έξι Κοινά Επίπεδα Αναφοράς.

Από 2001 και μετά, όλοι οι αρμόδιοι φορείς για την διδασκαλία ξένων γλωσσών στην Ευρώπη, αλλά και σε άλλες χώρες όπως η Κολομβία και οι Φιλιππίνες, προσαρμόζονται σταδιακά στις επιταγές του ΚΕΠΑΓ. Η ανάπτυξη προγραμμάτων διδασκαλίας γλωσσικών μαθημάτων και οδηγιών και η σύνταξη σχετικών προγραμμάτων σπουδών, εξετάσεων και διδακτικών εγχειριδίων, έχει εναρμονιστεί με τα Κοινά Επίπεδα Αναφοράς. Το γεγονός αυτό, τα καθιστά αναπόφευκτη και ιδιαίτερα βολική επιλογή για την κλίμακα αξιολόγησης που θα χρησιμοποιήσουμε σε αυτήν την εργασία.

2.7 Το *Πρόγραμμα Σπουδών του Ινστιτούτου Θερβάντες* και τα Επίπεδα Αναφοράς για τα Ισπανικά.

Όπως είδαμε στην προηγούμενη ενότητα, τα Κοινά Επίπεδα Αναφοράς, είναι αυτό ακριβώς που δηλώνει το όνομά τους. Κοινά για όλες τις γλώσσες. Αυτό σημαίνει ότι έχουν γενική αξία, αναπτύχθηκαν ανεξάρτητα από οποιαδήποτε γλώσσα και για αυτό, για να έχουν εφαρμογή στη διδακτική πρακτική, πρέπει να εξειδικευτούν για κάθε γλώσσα και να εμπλουτιστούν με το αντίστοιχο υλικό. Για τα ισπανικά, το φορτίο αυτής της εργασίας το ανέλαβε, όπως ήταν αναμενόμενο, το Ινστιτούτο Θερβάντες.

Το Ινστιτούτο Θερβάντες είναι ένας ισπανικός δημόσιος οργανισμός που ιδρύθηκε το 1991 με δύο βασικούς σκοπούς. Τη διδασκαλία και προώθηση της ισπανικής γλώσσας και την διάδοση του ισπανικού και ισπανοαμερικάνικου πολιτισμού. Στις δραστηριότητές του περιλαμβάνονται: η οργάνωση μαθημάτων ισπανικών, η οργάνωση και διεξαγωγή των εξετάσεων για τα Διπλώματα της Ισπανικής ως Ξένης Γλώσσας (DELE), η επιμόρφωση καθηγητών της ισπανικής και η επικαιροποίηση των μεθόδων διδασκαλίας, η στήριξη στο έργο των ισπανιστών και η διατήρηση

ανοιχτών για το κοινό βιβλιοθηκών. Αυτήν τη στιγμή, διαθέτει περίπου 80 κέντρα σε όλο τον κόσμο.

Με την οριστικοποίηση του ΚΕΠΑΓ, το Ινστιτούτο Θερβάντες ανέλαβε να επικαιροποιήσει το πρόγραμμα σπουδών του, συμπεριλαμβάνοντας και τα Επίπεδα Αναφοράς για τα Ισπανικά. Η επικαιροποίηση έγινε ακολουθώντας τον *Οδηγό για την επεξεργασία περιγραφών των επιπέδων αναφοράς για τις εθνικές και τοπικές γλώσσες* του Τμήματος Γλωσσικής Πολιτικής του Συμβουλίου της Ευρώπης. Σύμφωνα με αυτόν *Οδηγό*, η περιγραφές θα έπρεπε να ακολουθούν το μοντέλο της σειράς του *Threshold-Level*, όπου η προκαθορισμένη για κάθε επίπεδο ύλη —εννοιολογική, λειτουργική, γραμματική κ.λπ.— παρουσιάζεται σε καταλόγους οργανωμένη σε ταξινομικές κατηγορίες. Οι ελάχιστοι απαιτούμενοι κατάλογοι είναι οι κατάλογοι γλωσσικών λειτουργιών, ειδικών εννοιών, γενικών έννοιών και γραμματικής. Το επικαιροποιημένο πρόγραμμα σπουδών εκδόθηκε το 2006, και σύμφωνα με τους δημιουργούς του, ήταν, τουλάχιστον μέχρι τότε, η πιο πλήρης καταγραφή επιπέδων αναφοράς για οποιαδήποτε γλώσσα. Η παρουσίαση του περιεχομένου καθενός από τους τρεις τόμους που το αποτελούν —κάθε ένας αντιστοιχεί σε ένα από τα τρία ευρύτερα στάδια Α,Β και Γ και τα επίπεδα,1 και 2, που περιλαμβάνουν— ακολουθεί το εξής σχήμα: αρχικά, αναφέρονται οι γενικοί στόχοι του επιπέδου και ακολουθούν 12 κατάλογοι με την υλη που είναι απαράιτητη για να πραγματοποιηθούν οι επικοινωνιακές δραστηριότητες που ορίζει το ΚΕΠΑΓ για το συγκεκριμένο επίπεδο. Οι κατάλογοι περιλαμβάνουν τόσο γλωσσικό, όσο και μη γλωσσικό υλικό (πολιτισμικού χαρακτήρα ή σχετιζόμενου με τη διαδικασία της μάθησης). Στο Παράρτημα 2 παραθέτουμε την οργάνωση των καταλόγων.

Από αυτούς τους καταλόγους, και πιο συγκεκριμένα τους καταλόγους της γραμματικής και των γενικών και ειδικών εννοιών, αντλήσαμε υλικό για κάποιες από τις μετρήσεις που κάναμε στα πλαίσια αυτής της εργασίας.

3. Στόχοι και μεθοδολογία.

3.1 Στόχοι της εργασίας.

Ο βασικός στόχος αυτής της εργασίας είναι η εκπαίδευση ενός ταξινομητή, με τεχνικές μηχανικής μάθησης, ο οποίος να κατατάσσει κείμενα της ισπανικής, στο επίπεδο αναφοράς του ΚΕΠΑΓ στο οποίο ανήκουν. Οι εργασίες στις οποίες έχουμε αναφερθεί μέχρι τώρα, έχουν δουλέψει με συγκεκριμένα κειμενικά είδη και συγκεκριμένα θεματικά πεδία. Στην περίπτωσή μας, αυτό που μας ενδιαφέρει είναι να μπορούμε να εκτιμούμε την αναγνωσιμότητα οποιουδήποτε κειμένου μπορεί να αφορά την εκπαιδευτική δραστηριότητα.

Για την εκπαίδευση του ταξινομητή, θέλουμε να χρησιμοποιήσουμε διάφορες ομάδες χαρακτηριστικών, υφομετρικά, γραμματικά, λεξιλογικά και άλλα, και να ελέγξουμε την επίδρασή τους στην επιτυχία του μοντέλου. Είναι επίσης σκοπός μας, να αξιοποιήσουμε στις μετρήσεις που θα κάνουμε, μέρος της ύλης που περιλαμβάνεται σε κάποιους από τους καταλόγους του *Προγράμματος Σπουδών του Ινστιτούτου Θερβάντες*, προσπαθώντας να κωδικοποιήσουμε τις πληροφορίες που περιέχουν με τρόπο που να είναι δυνατή η ανίχνευσή τους στα κείμενα του σώματος κειμένων εκπαίδευσης που συλλέξαμε.

3.2 Σώμα κειμένων.

Μια δυσκολία που, τις περισσότερες φορές, καλούνται να ξεπεράσουν οι σχετικές με την αναγνωσιμότητα έρευνες, είναι η ανυπαρξία σωμάτων κειμένων με επισημειωμένο τον βαθμό δυσκολίας, τουλάχιστον για άλλες γλώσσες εκτός από τα αγγλικά. Ακόμα και η ύπαρξη τέτοιων σωμάτων, δε σημαίνει ότι είναι κατάλληλα για χρήση στη συγκεκριμένη μελέτη που ενδιαφέρει τον κάθε ερευνητή. Είδαμε, για παράδειγμα, ότι οι Heilman et al.(2007) χρησιμοποίησαν δύο διαφορετικά σώματα για να μελετήσουν την αναγνωσιμότητα της αγγλικής ως μητρικής και ως ξένης γλώσσας. Είναι τυπικό κομμάτι των σχετικών εργασιών, η συλλογή κειμένων που εξυπηρετούν τις ανάγκες της κάθε έρευνας. Συχνή, επίσης, είναι η ανάγκη βαθμονόμησης των συλλεχθέντων κειμένων, με βάση την κλίμακα αξιολόγησης που χρησιμοποιεί η έρευνα. Η εργασία αυτή μπορεί να ανατεθεί σε ειδικούς (Kate et al., 2010; Μικρός, n.d.) —ανθρώπους που έχουν μια ιδιαίτερη σχέση με τη γλώσσα και τα κριτήρια αξιολόγησης της δυσκολίας, όπως καθηγητές, φιλόλογους, γλωσσολόγους κ.λπ.— ή σε μη ειδικούς. Στη δεύτερη περίπτωση η αξιολόγηση γίνεται είτε με απευθείας βαθμολόγηση των κειμένων από τους κριτές και με κάποια μέθοδο συμψηφισμού των διαφορετικών βαθμολογιών για κάθε κείμενο (Brück et al., 2008; Das & Roychoudhury, 2006; Kate et al., 2010), είτε, κυρίως σε παλαιότερες έρευνες, αξιολογώντας με μεθόδους όπως οι δοκιμασίες συμπλήρωσης κενών, τον βαθμό κατανόησης των κειμένων που επέδειξαν αναγνωστες γνωστού επιπέδου.

Για τις ανάγκες της εργασίας μας, χρειαζόμαστε ένα σώμα ισπανικών κειμένων κατηγοριοποιημένων στα επίπεδα αναφοράς του ΚΕΠΑΓ. Όπως έχουμε ήδη αναφέρει, ο ταξινομητής που θέλουμε να φτιάξουμε είναι γενικής χρήσης. Δεν μας ενδιαφέρει, δηλαδή, η δυνατότητα ταξινόμησης μόνο ενός συγκεκριμένου κειμενικού είδους, λόγου χάρη ειδησεογραφικών ή λογοτεχνικών κειμένων, αλλά,

η εκτίμηση της αναγνωσιμότητας οποιουδήποτε κειμένου μπορεί να εμπλακεί στην εκπαιδευτική δραστηριότητα. Θέλουμε επίσης να δουλέψουμε με όλο το φάσμα των επιπέδων αναφοράς από το Α1 μέχρι το Γ2. Αυτό από μόνο του επιβάλλει τη συμπερίληψη διαφορετικών κειμενικών ειδών στο corpus. Στον Πίνακας 8 του Παραρτήματος 2, όπου παραθέτουμε τα ονόματα των καταλόγων που περιλαμβάνονται στο *Πρόγραμμα Σπουδών του Ινστιτούτου Θερβάντες*, μπορούμε να δούμε ότι υπάρχει ένας κατάλογος με το όνομα *Κειμενικά είδη και παραγωγή λόγου*. Σε αυτόν περιλαμβάνονται, εκτός των άλλων, περιγραφές των κειμενικών ειδών τα οποία πρέπει να είναι ικανός να καταλάβει ή να παραγάγει ο μαθητής κάθε επιπέδου. Μεταφράζοντας από τον Πίνακας 9 που περιέχει τα αντίστοιχα κειμενικά είδη για τα επίπεδα Α1-Α2, τον οποίο παραθέτουμε ενδεικτικά στο Παράρτημα 2, βλέπουμε ότι στη λίστα για το Α1 συναντάμε τα τουριστικά φυλλάδια (Hojas y folletos con información turística), ότι στη λίστα του Α2 υπάρχουν οι στήλες των εφημερίδων με το πρόγραμμα των κινηματογράφων (Carteleras de espectáculos), και ότι καμία από τις δύο λίστες δεν περιλαμβάνει ακαδημαϊκές εργασίες ή σχολικά εγχειρίδια ή συμβόλαια, τα οποία συναντώνται σε υψηλότερα επίπεδα. Παρατηρούμε, ακομα, ότι το ρεπερτόριο των κειμενικών ειδών αυξάνεται όσο ανεβαίνει το επίπεδο.

Από όσο γνωρίζουμε, δεν υπάρχει κάποιο έτοιμο σώμα κειμένων που να πληροί τις προϋποθέσεις που θέσαμε στην προηγούμενη παράγραφο. Η μοναδική περίπτωση εργασίας, σχετικής με την αναγνωσιμότητα της ισπανικής ως ξένης γλώσσας, που συναντήσαμε στη βιβλιογραφία και που χρησιμοποιεί τα επίπεδα αναφοράς του ΚΕΠΑΓ, ήταν αυτή της Checa-García (2013). Σε αυτήν την εργασία ελέγχεται η επίδραση της μορφοσυντακτικής πολυπλοκότητας στην αναγνωσιμότητα κειμένων από διαβαθμισμένα βιβλία για σπουδαστές ισπανικών. Η Checa-García χρησιμοποίησε 61 αφηγηματικά αποσπάσματα 220-300 λέξεων, 44 από πρωτότυπα κείμενα και 17 από προσαρμοσμένα, τα οποία κατέταξε στα τρία ευρύτερα στάδια (Α,Β και Γ) του ΚΕΠΑΓ.

Για τη δική μας εργασία θέσαμε ως στόχο να μαζέψουμε 50-60 κείμενα για κάθε ένα από τα 6 επίπεδα. Σαν πηγές χρησιμοποιήσαμε εγχειρίδια διδασκαλίας, θέματα εξετάσεων DELE, βιβλία προετοιμασίας για εξετάσεις DELE, διαβαθμισμένα λογοτεχνικά, και όχι μόνο, βιβλία και τον Παγκόσμιο Ιστό. Προσπαθήσαμε, δηλαδή, να βρούμε κείμενα ήδη βαθμονομημένα από ειδικούς, τα οποία χρησιμοποιούνται στη διδακτική πρακτική και αποτελούν αντιπροσωπευτικο δείγμα των κειμένων με τα οποία έρχεται σε επαφή ο μαθητής, τόσο μέσα στην τάξη, όσο και έξω από αυτήν. Με αυτήν τη λογική, συμπεριλάβαμε μια μεγάλη ποικιλία κειμενικών ειδών: Ενημερωτικά (ειδήσεις, ρεπορτάζ, συνεντεύξεις), αφηγηματικά (ιστορικά, βιογραφίες, προσωπικές αφηγήσεις), περιγραφικά (περιγραφές ανθρώπων, σπιτιών, ρούχων κ.λπ.), πληροφοριακά (επιστημονικά, τεχνικά, παρουσιάσεις πόλεων, εθίμων,εταιρειών κ.λπ.), αλληλογραφία (προσωπική, επαγγελματική, διοικητική , ηλεκτρονική), επειχηρηματολογικά (άρθρα γνώμης, δοκίμια) και, τέλος, κάποια προφορικά (προσωπικοί διάλογοι, καθημερινές συνδιαλλαγές, εκφωνήσεις ραδιοφωνικών ειδήσεων κ.λπ.). Για τα τελευταία, να διευκρινίσουμε ότι παρόλο που ίσως φαίνεται αντιφατική η συμπερίληψή τους σε μια έρευνα για την αναγνωσιμότητα, που συνδέεται με την κατανόηση του γραπτού λόγου, θεωρήσαμε ότι μπορούμε να τα χρησιμοποιήσουμε για δύο λόγους. Ο πρώτος είναι

ότι τα κείμενα που χρησιμοποιήσαμε δεν αποτελούν σε καμία περίπτωση απομαγνητοφωνήσεις αυθεντικών διαλόγων ή άλλων αυθεντικών προϊόντων προφορικής επικοινωνίας. Ακόμα και οι απομαγνητοφωνήσεις δραστηριοτήτων κατανόησης προφορικού λόγου που συμπεριλάβαμε, έχουν πρώτα συνταχθεί ως γραπτό κείμενο και μετά ερμηνευθεί από ηθοποιούς στις ηχογραφήσεις. Επιπλέον, σε πολλές περιπτώσεις, δεν εντοπίζουμε καμία ποιοτική διαφορά ανάμεσα σε γραπτές ειδήσεις και στις εκφωνήσεις τους στο ραδιόφωνο. Θα μπορούσαμε ίσως να χαρακτηρίσουμε αυτά τα κείμενα ως ψευδοπροφορικά. Ο δέυτερος λόγος είναι ότι κείμενα σαν τα παραπάνω, απαντώνται έτσι και αλλιώς και σε γραπτή μορφή στα διάφορα εγχειρίδια. Για παράδειγμα, όλα τα εγχειρίδια επιπέδου Α1 περιλαμβάνουν διαλόγους μεταξύ πελατών και πωλητών σε καταστήματα. Επίσης, διαλόγους θα συναντήσει κανεις στα περισσότερα λογοτεχνικά βιβλία, τα οποία θα θέλαμε να μπορεί να κατηγοριοποιήσει ο ταξινομητής μας.

Ως προς το μέγεθος των κειμένων, προσπαθήσαμε να μην έχουν λιγότερες από 150-200 λέξεις. Πάνω σε αυτό, προέκυψαν κάποιες δυσκολίες με τα κείμενα του επιπέδου Α1 όπου είναι σύνηθες να εμφανίζονται μικρότερα κείμενα. Για να αντιμετωπίσουμε το πρόβλημα, σε ορισμένες περιπτώσεις χρειάστηκε να συνενώσουμε στο ίδιο κείμενο διάφορα ομοειδή κειμενα μικρότερης έκτασης, που όμως εμφανίζονται στην ίδια εκπαιδευτική δραστηριότητα. Για παράδειγμα, αν μια δραστηριότητα περιλάμβανε 4 κείμενα των 60 λέξεων, με θέμα άτομα που παρουσιάζουν τους εαυτούς τους, ενοποιήσαμε τα 4 μικρά κείμενα σε ένα μεγαλύτερο. Αυτό δεν έφτασε και αναγκαστήκαμε να συμπεριλάβουμε στο corpus μας κάποια, ελάχιστα, μικρότερα κείμενα.

Ένα άλλο σημαντικό κριτήριο για την επιλογή των κειμένων ήταν η χρονολογία έκδοσης των πηγών τους. Πέρα από το προφανές, ότι για να βρούμε κείμενα ταξινομημένα στα επίπεδα αναφοράς του ΚΕΠΑΓ, θα έπρεπε να ψάξουμε σε μεταγενέστερες από αυτό πηγές —θυμίζουμε ότι το ΚΕΠΑΓ δημοσιεύθηκε το 2001—, θεωρήσαμε σημαντικό οι πηγές αυτές να είναι μεταγενέστερες και του *Προγράμματος Σπουδών του Ινστιτούτου Θερβάντες* και να έχουμε ενδείξεις ότι αυτό έχει ληφθεί υπόψη κατά τη δημιουργία τους.

Σε πρώτη φάση, εντοπίσαμε κείμενα με τις προδιαγραφές που είχαμε θέσει, τα σαρώσαμε, τα περάσαμε από οπτική αναγνώριση χαρακτήρων (OCR) και τα διορθώσαμε. Η διόρθωση περιλάμβανε εκτός από τα λάθη της οπτικής αναγνώρισης και αφαίρεση τυχόν θορύβου από τα κείμενα. Σαν παράδειγμα θορύβου μπορούμε να δώσουμε τη συνεχή αναφορά του ονοματος του προσώπου που μιλάει κάθε φορά σε απομαγνητοφωνήσεις διαλόγων ή τις επιπλέον πληροφορίες που συνοδεύουν τις παρεμβάσεις σε ένα φόρουμ (όνομα χρήστη, ημερομηνία και ώρα δημοσίευσης κ.λπ.).Κατά τη διάρκεια αυτής της διαδικασίας, με την πιο προσεκτική ανάγνωση των κειμένων και την πιο συνολική εικόνα που σχηματίσαμε, συνειδητοποιήσαμε δύο πράγματα. Το πρώτο είναι ότι τα κείμενα που συμπεριλαμβάνονται σε ένα βιβλίο συγκεκριμένου επιπέδου, δεν είναι 100% σίγουρο ότι ανήκουν όλα στο επίπεδο αυτό. Τέτοιες αναντιστοιχίες είναι πολύ πιο πιθανό να συναντηθούν σε διδακτικά εγχειρίδια παρά σε βιβλία προετοιμασίας εξετάσεων ή διαβαθμισμένων αναγνωσμάτων. Υπάρχουν, για παράδειγμα, εγχειρίδια τα οποία εκτός από την προβλεπόμενη ύλη, περιλαμβάνουν,

διαχωρισμένα, κάποια επιπλέον κείμενα που προορίζονται για προαιρετική περαιτέρω πληροφόρηση γύρω από ένα θέμα και περιλαμβάνουν έξτρα λεξιλόγιο που ξεφεύγει από τις ανάγκες των μαθητών του συγκεκριμένου επιπέδου. Το δεύτερο είναι ότι ανάμεσα στους διάφορους εκδοτικούς οίκους και στις διάφορες συντακτικές ομάδες, υπάρχουν διαφορές ως προς την αντίληψη της δυσκολίας των κειμένων. Αυτό έχει επισημανθεί και από άλλους ερευνητές (Checa-García, 2013). Σε αυτό το σημείο, θεωρήσαμε ότι θα ήταν χρήσιμο να έχουμε μια εξωτερική αξιολόγηση των κειμένων. Η ιδέα ήταν να μοιράσουμε τα κείμενα του κάθε επιπέδου σε μαθητές που να έχουν ολοκληρώσει το αντίστοιχο επίπεδο χωρίς να έχουν προχωρήσει ακόμα στο επόμενο και να τους ζητήσουμε να τους δώσουν βαθμούς από το 1 μέχρι το 10, ανάλογα με τη δυσκολία τους. Δυστυχώς αυτή η ιδέα δεν καρποφόρησε λόγω μηδαμινής ανταπόκρισης από πλευράς μαθητών. Από τις ελάχιστες περιπτώσεις κειμένων για τα οποία πήραμε 3-5 αξιολογήσεις, φάνηκε να υπάρχει αρκετή απόκλιση μεταξύ τους, γεγονός που έχει παρατηρηθεί και σε άλλες έρευνες (Μικρός, n.d.). Αναγκαστήκαμε, λοιπόν, να περιοριστούμε στις δικές μας αξιολογήσεις που είχαν σαν συνέπεια, ένα μικρο ποσοστό κειμένων, γύρω στο 10%, να μεταπηδήσει σε διπλανά επίπεδα. Οι αξιολογήσεις μας στηρίχτηκαν στη συνολική εικόνα που είχαμε σχηματίσει μετά τη συγκριτική ανάγνωση πολλών κειμένων και στους κατάλογους του Ινστιτούτου Θερβάντες με τα κειμενικά είδη ανά επίπεδο. Να σημειώσουμε ότι με βάση αυτούς τους καταλόγους εντάξαμε στο corpus μας τα μη βαθμονομημένα κείμενα που αντλήσαμε από τον Παγκόσμιο Ιστό. Αυτό αφορά κείμενα του επιπέδου Γ2 και κειμενικά είδη όπως νομικά έγγραφα, επιστημονικά άρθρα, φιλοσοφικά δοκίμια και άλλα. Καταλήξαμε με ένα σώμα κειμένων του οποίου μια ποσοτική περιγραφή μπορούμε να δούμε στον παρακάτω πίνακα.

Επίπεδο	Κείμενα	Αρ.Λέξεων	Μέσος Όρος	Ελάχιστο	Μέγιστο	Τυπική Απόκλιση
Α1	63	15557	246,9	75	713	140,5
Α2	69	20638	299,1	101	1479	187,4
Β1	55	19991	363,5	116	847	112,6
Β2	59	25970	440,2	233	773	138,9
Γ1	52	31695	609,5	216	1051	195,3
Γ2	57	33677	590,8	168	1089	208,4
Σύνολο	355	147528	415,6	75	1479	216,1

Πίνακας 1. Ποσοτική περιγραφή του σώματος εκπαίδευσης.

Για την επεξεργασία των κειμένων του corpus μας, χρησιμοποιήσαμε το ανοιχτού κώδικα εργαλείο Freeling 3.1[5]. Το Freeling έχει αναπτυχθεί στο Κέντρο για τη Γλώσσα και τις Τεχνολογίες και Εφαρμογές Λόγου (TALP) του Πολυτεχνείου της Καταλονίας (UPC) και υποστηρίζει, εκτός των άλλων, όλες τις λατινογενείς γλώσσες της Ιβηρικής Χερσονήσου. Η ανάλυση των κειμένων σε λεξικές μονάδες (tokenization) έγινε με τις προτερόθετες για τα ισπανικά ρυθμίσεις. Η μορφολογική τους επισημείωση έγινε με τον επισημειωτή κρυφών μαρκοβιανών μοντέλων (HMM)

[5] http://nlp.lsi.upc.edu/freeling/

του εργαλείου[6], που αποδίδει στις λεξικές μονάδες ετικέτες από το σετ EAGLES για τα ισπανικά[7].

3.3 Κειμενικά χαρακτηριστικά.

Το επόμενο στάδιο της εργασίας ήταν να αναπαραστήσουμε κάθε κείμενο σαν μια ακολουθία ποσοτικά εκφρασμένων χαρακτηριστικών, η οποία περιέχει επιπλέον την κατηγορία, στην περίπτωσή μας το επίπεδο γλωσσομάθειας, στην οποία ανήκει το κείμενο. Με σύνολα τέτοιων αναπαραστάσεων (datasets), τροφοδοτούνται αλγόριθμοι μηχανικής μάθησης, οι οποίοι μοντελοποιούν τις τιμές των διαφόρων χαρακτηριστικών και έτσι εκπαιδεύεται ένας ταξινομητής που, με βάση το μοντέλο που έχει δημιουργηθεί, μπορεί να κατατάξει σωστά περιπτώσεις των οποίων η κατηγορία είναι άγνωστη.

Πριν περάσουμε στα κειμενικά χαρακτηριστικά που χρησιμοποιήσαμε στην εργασία μας, να κάνουμε μια μικρή αναφορά σε κάποια που δεν χρησιμοποιήσαμε αλλά θεωρούμε ότι θα μπορούσαν να είχαν βοηθήσει. Είδαμε στο προηγούμενο κεφάλαιο, ότι σε πολλές έρευνες χρησιμοποιήθηκαν με επιτυχία στατιστικά γλώσσικά μοντέλα. Στην παρούσα εργασία θα θέλαμε να συμπεριλάβουμε χαρακτηριστικά που προκύπτουν από τη χρήση τέτοιων μοντέλων, όπως η περιπλοκή, αλλά δεν ήταν δυνατό λόγω του μικρού μεγέθους του corpus μας. Το βασικότερο πρόβλημα είναι ότι θα έπρεπε να χρησιμοποιήσουμε διαφορετικά κείμενα για την ανάπτυξη των γλωσσικών μοντέλων και διαφορετικά για την εκπαίδευση του ταξινομητή, αλλιώς η χρήση του ίδιου σώματος θα απέδιδε, λανθασμένα, αυξημένες πιθανότητες στα N-γράμματα του, αλλοιώνοντας τις τιμές της περιπλοκής, με αποτέλεσμα να μην μπορούν να γενικευτούν σε άγνωστα κείμενα (Jurafsky & Martin, 2000). Στη δικιά μας περίπτωση, δεν είχαμε την πολυτέλεια να κρατήσουμε εκτός της διαδικασίας εκπαίδευσης του ταξινομητή κάποια κείμενα για ανάπτυξη γλωσσικών μοντέλων γιατί θα καταλήγαμε με πολύ λίγα δεδομένα και για τις δύο διαδικασίες. Δεν χρησιμοποιήσαμε ούτε συντακτικά χαρακτηριστικά που να προέρχονται από συντακτική επεξεργασία του κειμένου. Σε όλες τις περιπτώσεις χρήσης τέτοιων χαρακτηριστικών που συναντήσαμε στη βιβλιογραφία, η αναπαράσταση της συντακτικής δομής των κειμένων είχε γίνει με δέντρα συστατικής δομής και οι μετρήσεις είχαν βασιστεί στις ιδαιτερότητες αυτού του είδους δέντρων. Δυστυχώς δεν μπορέσαμε να βρούμε κάποιο συντακτικό αναλυτή για τα ισπανικά που να κάνει αυτού του είδους την ανάλυση.

3.3.1 Υφομετρικά χαρακτηριστικά.

Μια πρώτη ομάδα χαρακτηριστικών που μετρήσαμε ήταν υφομετρικά. Όπως αναφέρει ο Μικρός (n.d.), τέτοιου είδους χαρακτηριστικά έχουν χρησιμοποιηθει κυρίως στον εντοπισμό του συγγραφικού ύφους, αλλά έχει φανεί ότι σχετίζονται και με άλλα μετακειμενικά χαρακτηριστικά. Επιπλέον δύο κατεξοχήν υφομετρικά χαρακτηριστικά όπως το μέσο μήκος λέξης και το μέσο μήκος πρότασης, έχουν χρησιμοποιηθεί εκτεταμένα σε έρευνες αναγνωσιμότητας. Στην εργασία μας

[6]Οι επιλογές που θέσαμε για τον μορφολογικό επισημειωτή ήταν οι εξής: AffixAnalysis=yes, MultiwordsDetection=no, NumbersDetection=yes, PunctuationDetection=yes, DatesDetection=no, QuantitiesDetection=no, DictionarySearch=yes, ProbabilityAssignment=yes, OrthographicCorrection=no

[7] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

αξιοποιήσαμε τα παρακάτω υφομετρικά χαρακτηριστικά που χρησιμοποίησε ο Μικρός, με ενθαρρυντικά αποτελέσματα, στην έρευνά του για την αναγνωσιμότητα μιας συγγενικής προς την ισπανική γλώσσας, της ιταλικής:

- **Type/token ratio** (TTR): Είναι ο λόγος των μοναδιαίωνλεξιλογικών μονάδων (types) προς τον αριθμό των λέξεων (tokens) του κειμένου. Πρόκειται για ένα δείκτη λεξιλογικού πλούτου αφού όσο μεγαλύτερος είναι, τοσο περισσότερες διαφορετικές λέξεις έχει το κείμενο σε σχέση με τον συνολικό αριθμό λέξεών του.
- **Μέσο μήκος λέξης** (AWL): Το μέσο μήκος λέξεων του κειμένου υπολογισμένο σε χαρακτήρες. Ευρέως χρησιμοποιημένη μεταβλητή στις εξισώσεις αναγνωσιμότητας. Θεωρείται ότι αποτυπώνει τη λεξιλογική δυσκολία.
- **Τυπική απόκλιση του μέσου μήκους λέξης** (WLsd): Η τυπική απόκλιση του μέσου όρου του μήκους των λέξεων του κειμένου.
- **Μέσο μήκος πρότασης** (ASL): Υπολογισμένο σε λέξεις. Έχει χρησιμοποιηθεί στις περισσότερες εξισώσεις αναγνωσιμότητας ως μεταβλητή που αποτυπώνει τη συντακτική δυσκολία του κειμένου.
- **Τυπική απόκλιση του μέσου μήκους πρότασης** (SLsd).
- **Άπαξ λεγόμενα** (HapL): Το ποσοστό των λέξεων που εμφανίζονται στο κείμενο μόνο μία φορά.
- **Δις λεγόμενα** (DisL): Το ποσοστό των λέξεων που εμφανίζονται δύο φορές στο κείμενο.
- **Λόγος Δις προς Άπαξ Λεγόμενα** (Dis_HapL): Ο λόγος των δύο προηγούμενων μεταβλητών.
- **Λεξιλογική Πυκνότητα** (LD): Πρόκειται για τον λόγο του αριθμού των λέξεων περιεχομένου προς τον αριθμό των λειτουργικών λέξεων. Λειτουργικές, θεωρούνται οι λέξεις οι οποίες συνεισφέρουν γραμματική πληροφορία — στα ισπανικά, τα άρθρα, οι αντωνυμίες, οι προθέσεις, οι σύνδεσμοι, μερικά επιρρήματα και μερικά ρήματα (τα βοηθητικά)— και περιεχομένου θεωρούνται οι λέξεις που συνεισφέρουν λεξιλογική πληροφορία (Real Academia Española, 2010).
- **Yule's K (**Yule): Αυτός ο δείκτης δημιουργήθηκε το 1944 από τον Βρετανό στατιστικολόγο George Udny Yule και θεωρείται ένας από τους πιο αξιόπιστους δείκτες λεξιλογικού πλούτου. Πρόκειται για ένα μέτρο της πιθανότητας που έχουν δύο οποιεσδήποτε τυχαίες λέξεις του κειμένου, να είναι ίδιες. Για να αποφύγει πολύ μικρά νούμερα, ο Yule πολλαπλασίασε την πιθανότητα με το 10000. Ο τύπος υπολογισμού αυτού του δείκτη είναι K=10000*(M2-M1)/(M1*M1). Όπου M1 είναι ουσιαστικά ο αριθμός των tokens, και M2 είναι το άθροισμα όλων των γινομένων του αριθμού των λέξεων που έχουν ίδια συχνότητα εμφάνισης, επί το τετράγωνο της συχνότητας αυτής. Για παράδειγμα, αν ένα κείμενο 12 λέξεων έχει 2 λέξεις που εμφανίζονται από μία φορά, 2 λέξεις που εμφανίζονται από 2 και 2 που εμφανίζονται από 3, το M2 για αυτό το κείμενο είναι $M2=(2*1^2)+(2*2^2)+(2*3^2)$. Η τιμή του Yule's K μεγαλώνει όσο μεγαλώνει η λεξιλογική ομοιομορφία σε ένα κείμενο (Oakes, 1998).
- **Εντροπία** (Entr): Η εντροπία ενός κειμένου εκφράζει τον βαθμό της αβεβαιότητάς του. Ας σκεφτούμε μία πρόταση που αρχίζει με τη λέξη «Θέλω». Η πιθανότητα η επόμενη λέξη να είναι «να» είναι πολύ μεγάλη ενώ η πιθανότητα να ακολουθεί η λέξη «καλησπέρα» μηδαμινή. Η εντροπία είναι

χαμηλή σε τέτοιες περιπτώσεις, όπου οι πιθανότητες εμφάνισης διαφορετικών λέξεων είναι πολύ άνισες, δηλαδή, όταν υπάρχει μεγάλη προβλεψιμότητα. Αντίθετα, είναι υψηλή όταν όλες οι λέξεις ή, γενικότερα, όταν όλα τα σύμβολα είναι ισοπίθανα. Εμείς, μετρήσαμε την εντροπία (H) σε επίπεδο λέξης, χρησιμοποιώντας τον τύπο που πήραμε πάλι από τον Oakes:

$$H=-[p_1\log_2(p_1)+ p_2\log_2(p_2)+\ldots+ p_n\log_2(p_n)]$$

Όπου p_1, p_2,..., p_n, η πιθανότητα εμφάνισης για κάθε λέξη του κειμένου. Το μείον χρησιμοποιείται για να είναι η εντροπία θετικός αριθμός αφού οι λογάριθμοι των πιθανοτήτων με βάση το 2 είναι αρνητικοί αριθμοί.

- **Σχετική εντροπία** (RelEntr): Ο λόγος της θεωρητικά μέγιστης εντροπίας ενός κειμένου προς την πραγματική εντροπία του. Μέγιστη θα ήταν η εντροπία ενός κειμένου αν όλες οι λέξεις του είχαν συχνότητα 1, αν ήταν δηλαδή όλες άπαξ λεγόμενα.
- **Φάσμα συχνότητας μήκους λέξεων** (LW1…LW14): Το ποσοστά των λέξεων του κειμένου που αποτελούνται από 1,2,3...14 γράμματα.

Να σημειώσουμε εδώ ότι οι μετρήσεις των παραπάνω χαρακτηριστικών έγιναν με τα κείμενα χωρισμένα σε λεξιλογικές μονάδες (tokenized) και τους αριθμούς αντικατεστημένους με μια κοινή ετικέτα. Κάναμε αυτήν την επιλογή με το σκεπτικό ότι οι αριθμοί, σημειωμένοι με ψηφία, δεν παρουσιάζουν κάποια δυσκολία για τους μαθητές μιας και αποτελούν διεθνή σύμβολα. Αν τους αφήναμε ως είχαν, θα επηρέαζαν τους δείκτες λεξιλογικού πλούτου, αποδίδοντας μια πλασματικά μεγαλύτερη λεξιλογική πολυπλοκότητα στο κείμενο. Θεωρούμε, δηλαδή, ότι ένα κείμενο που περιλαμβάνει πολλούς αριθμούς, δεν κρύβει, από αυτό και μόνο, περισσότερες δυσκολίες για ένα μαθητή ξένης γλώσσας. Αντιθέτως, μπορεί να του φαίνεται και πιο εύκολο.

3.3.2 Λεξιλογικά χαρακτηριστικά.

Στη συνέχεια μετρήσαμε κάποια λεξιλογικά χαρακτηριστικά. Έχουμε δει ότι ήδη από τις πρώτες μελέτες της αναγνωσιμότητας, έγιναν προσπάθειες να εκτιμηθεί η εννοιολογική δυσκολία των κειμένων. Εκτός από μεταβλητές όπως το μέσο μήκος λέξης, η χρήση της οποίας πολλές φορές στηρίχθηκε στην υπόθεση ότι οι μεγαλύτερες λέξεις είναι και πιο δύσκολες, χρησιμοποιήθηκαν και άλλες, που προέκυψαν από μετρήσεις με βάση λίστες λέξεων. Στο προηγούμενο κεφάλαιο, αναφερθήκαμε στις λίστες του Thorndike και του Chall με τις συχνότερες λέξεις των αγγλικών. Είδαμε, επίσης, ότι τέτοιες λίστες έχουν χρησιμοποιηθεί ευρέως σε πολλές έρευνες αναγνωσιμότητας. Στον πυρήνα αυτής της προσέγγισης, υπάρχει η ιδέα ότι η δυσκολία μιας λέξης συνδέεται με το πόσο οικεία είναι, που με τη σειρά του σχετίζεται με τη συχνότητα χρήσης της λέξης. Οι συχνότερες λέξεις, συνήθως διαβάζονται πιο γρήγορα και γίνονται καλύτερα κατανοητές από τις πιο σπάνιες (Graesser, McNamara, Louwerse, & Cai, 2004). Η συνήθης εφαρμογή των παραπάνω στις μελέτες της αναγνωσιμότητας είναι η μέτρηση του ποσοστού των λέξεων των κειμένων που ανήκουν, ή δεν ανήκουν, στις λίστες των Χ συχνότερων λέξεων της γλώσσας. Σε μια πρώτη φάση, εργαστήκαμε και εμείς με τον παραπάνω τρόπο. Την λίστα με τις συχνότερες λέξεις των ισπανικών την πήραμε από το *Σώμα Κειμένων Αναφοράς για τη Σύγχρονη Ισπανική* (CREA) της Ισπανικής Βασιλικής Ακαδημίας (RAE).

Το CREA, στην τρέχουσα έκδοσή του, αυτή του Ιουνίου του 2008, περιέχει πάνω από 150 εκατομμύρια λέξεις το 10% των οποίων προέρχεται από προφορικές πηγές. Καλύπτει τη χρονική περίοδο από το 1975 μέχρι 2004 και περιλαμβάνει κείμενα από πάνω από 100 θεματικά πεδία των οποίων η γεωγραφική κατανομή είναι 50% από την Ισπανία και 50% από τις ισπανόφωνες χώρες της Αμερικής[8].

Για τους σκοπούς της εργασίας μας, πήραμε τις 5000 συχνότερες λέξεις του CREA και τις χωρίσαμε σε 5 λίστες, ανά χιλιάδα. Καταλήξαμε δηλαδή με μία λίστα με τις 1000 συχνότερες λέξεις, μία με τις λέξεις των θέσεων 1001-2000 στη γενική κατάταξη και ούτω καθεξής. Αυτή η κατανομή των λέξεων σε χιλιάδες δεν είναι ασυνήθιστη στη βιβλιογραφία (Criado & Sánchez, 2009; Liontou, 2012). Να σημειώσουμε ότι όταν λέμε λέξεις εννοούμε γλωσσικούς τύπους και όχι λήμματα. Ακολούθως, μετρήσαμε το ποσοστό των λέξεων (Κ1-Κ5) κάθε κειμένου που εντοπίζονται σε καθεμία από αυτές τις 5 λίστες.

Σε μια δεύτερη φάση, εργαστήκαμε με τους καταλόγους των γενικών και ειδικών εννοιών που περιλαμβάνει το *Πρόγραμμα Σπουδών του Ινστιτούτου Θερβάντες* και που αποτελούν την εννοιολογική συνιστώσα του. Με τον όρο γενικές έννοιες αναφερόμαστε σε αυτές που ο ομιλητής μπορεί να χρειαστεί, οποιοδήποτε και αν είναι το θέμα της επικοινωνιακής κατάστασης. Σε μεγάλο βαθμό πρόκειται για αφηρημένες έννοιες. Αντίθετα, οι ειδικές έννοιες σχετίζονται με συγκεκριμένες επικοινωνιακές καταστάσεις και καθορίζονται άμεσα από την επιλογή του θέματος. Για παράδειγμα, γενικές είναι οι έννοιες της ύπαρξης, της ποιότητας, της ποσότητας, του χώρου ή του χρόνου, που μπορούν να εκφραστούν με λέξεις όπως «υπάρχω», «απουσία», «αύριο» κ.λπ., και τις οποίες μπορεί να χρειαστούμε σε οποιαδήποτε επικοινωνιακή κατάσταση. Από την άλλη, μια ειδική έννοια όπως η «μακαρονάδα» θα εκφραστεί μόνο σε σχέση με το συγκεκριμένο θέμα της διατροφής. Το περιεχόμενο αυτών των καταλόγων, δεν απότελείται μόνο από λέξεις, με τη στενή έννοια του όρου, αλλά και από ευρύτερες λεξιλογικές μονάδες όπως λεξιλογικές συνάψεις, διαφόρων τύπων εκφράσεις κ.λπ. Στο Παράρτημα 2, Πίνακας 10, παραθέτουμε τις γενικότερες σημασιολογικές κατηγορίες στις οποίες είναι οργανωμένοι οι δύο αυτοί κατάλογοι.

Θα πρέπει να αναφέρουμε ότι οι συγγραφείς του Προγράμματος Σπουδών δεν προβάλλουν αξιώσεις πληρότητας γι' αυτούς τους καταλόγους. Κάτι τέτοιο είναι ανέφικτο για πολλούς λόγους. Ο πρώτος έχει να κάνει με το περιβάλλον της μάθησης και την ποικιλία της γλώσσας που χρησιμοποιείται στη γεωγραφική περιοχή όπου εκτυλίσσεται αυτή. Από τη μία, δεν γίνεται να συμπεριληφθούν στους καταλόγους όλες οι διαφορετικές λέξεις που δηλώνουν την ίδια έννοια σε διαφορετικές γεωγραφικές ποικιλίες της γλώσσας. Έτσι, έχει επιλεγεί να αποτυπωθεί στους καταλόγους η γλωσσική ποικιλία της κεντρικής και βόρειας Ισπανίας. Από την άλλη, εμπόδιο αποτελούν και οι πολιτισμικές διαφορές. Για παράδειγμα, στο επίπεδο Α2, στην κατηγορία τρόφιμα, συμπεριλαμβάνονται η «μπανάνα», το «μήλο» και το «πορτοκάλι», τα οποία είναι φρούτα ευρείας κατανάλωσης στην Ισπανία, και όχι κάποια άλλα εξωτικά φρούτα που ίσως είναι τα

[8] Για περισσότερες πληροφορίες για τη σύσταση του CREA βλ.
http://www.rae.es/recursos/banco-de-datos/crea-escrito για το γραπτό,
http://www.rae.es/recursos/banco-de-datos/crea-oral για το προφορικό.

πιο συνηθισμένα στον Παναμά ή στην Παραγουάη. Σε πολλές περιπτώσεις, η πληρότητα δεν ήταν καν στις προθέσεις των συγγραφέων. Κάποιες έννοιες έχουν συμπεριληφθεί ενδεικτικά. Στο επίπεδο B2, λόγου χάρη, στην κατηγορία για τα εργαλεία, τα ρούχα και τα μέρη εργασίας, έχουν δοθεί μόνο τα του ηλεκτρολόγου, χωρίς αυτό να σημαίνει ότι αυτά είναι τα μόνα που πρέπει να γνωρίζει ο μαθητής αυτού του επιπέδου. Απλά έχουν δοθεί σαν παράδειγμα οργάνωσης των σχετικών εννοιών που ο κάθε ενδιαφερόμενος θα πρέπει να το αναπαραγάγει ανάλογα με το επάγγελμα που τον απασχολεί. Τέλος, ένα άλλο θέμα, σχετικό με την πληρότητα των καταλόγων, είναι αυτό των συνδυασμών στις διάφορες εκφράσεις. Ας σκεφτούμε ένα παράδειγμα στα ελληνικά. Θα μπορούσαμε σε κάποιο επίπεδο να βρούμε την καταχώριση «χαλάει μια συσκευή» και σε κάποιο άλλο, μάλλον ανώτερο, το «χαλάει ένα φαγητό». Είναι προφανές ότι δεν μπορούν να συμπεριληφθούν στον κατάλογο όλα τα υπώνυμα των εννοιών «συσκευή» και «φαγητό». Αυτό επηρεάζει και τις μετρήσεις μας, μιας και αυτές έγιναν σε μορφολογικά επισημειωμένα κείμενα, χωρίς κάποια μορφή σημασιολογικής ανάλυσης, με αποτέλεσμα να μη μπορούμε να διαχωρίσουμε χρήσεις όπως οι παραπάνω. Ανοίξαμε αυτήν την παρένθεση, για να δείξουμε ότι σε καμία περίπτωση δεν καλύπτουν αυτοί οι κατάλογοι το σύνολο του λεξιλογίου που θα πρέπει να γνωρίζει μαθητής με την ολοκλήρωση του αντίστοιχου επιπέδου, ούτε τα κείμενα, με τα οποία δουλέψαμε, αποτελούνται αποκλειστικά από τις έννοιες που περιλαμβάνουν οι κατάλογοι. Επίσης θέλαμε να θίξουμε κάποιους περιορισμούς που επηρέασαν κάποιες μεθοδολογικές αποφάσεις, σχετικές με τις μετρήσεις.

Αυτό που μετρήσαμε είναι το ποσοστό των λέξεων του κειμένου που μπορεί να αποδοθεί στο ένα η το άλλο επίπεδο με βάση τους καταλόγους[9]. Για να γίνει αυτό, χρειάστηκε να μετατραπεί το περιεχόμενο των καταλόγων σε κανονικές εκφράσεις (regular expressions) που να ανιχνεύουν τις αντίστοιχες λέξεις, ή γενικότερα δομές, στα μορφολογικά επισημειωμένα από το Freeling κείμενα. Οι αναζητήσεις έπρεπε να γίνουν σε επισημειωμένα κείμενα, γιατί σε πολλές περιπτώσεις, ήταν απαραίτητο να πραγματοποιηθούν με βάση το λήμμα ή τη μορφολογική ετικέτα της έννοιας και όχι με την ακριβή καταχώριση στον κατάλογο. Για παράδειγμα, για την καταχώριση «λαμβάνω ένα φαξ», θα έπρεπε να μπορούμε να εντοπίζουμε και το «έλαβε ένα φαξ», και την καταχώριση «είμαι Χ χρονών» έπρεπε να την εντοπίζουμε ανεξαρτήτως αριθμού. Επειδή κάποιες λέξεις εμφανίζονται στον ιδιο κατάλογο και σαν ανεξάρτητες έννοιες και σαν τμήμα άλλων εκφράσεων, έπρεπε να φροντίσουμε να μη μετρηθούν περισσότερες από μία φορά. Για να το πετύχουμε αυτό, αποφασίσαμε να αφαιρούμε κάθε ανεύρεση από το κείμενο πριν συνεχίσουμε με τις υπόλοιπες αναζητήσεις. Στο παράδειγμα με το «φαξ», που υπάρχει στον ίδιο κατάλογο (ειδικές έννοιες A2) με το «λαμβάνω ένα φαξ», αν δεν γινόταν η αφαίρεση της πρώτης ανεύρεσης, η λέξη «φαξ» θα μετριόταν δύο φορές. Αυτή η προσέγγιση προϋποθέτει ότι θα αναζητηθεί και θα αφαιρεθεί πρώτα η έκφραση «λαμβάνω ένα φαξ». Σε αντίθετη περίπτωση, με την αφαίρεση του σκέτου «φαξ», δε θα ήταν δυνατή η ανεύρεσή της μεγαλύτερης έκφρασης. Λαμβάνοντας το

[9] Οι κατάλογοι μπορούν να βρεθούν στις παρακάτω διευθύνσεις:
http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/niveles/09_nociones_especificas_inventario_a1-a2.htm για τις ειδικές έννοιες
http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/niveles/08_nociones_generales_inventario_a1-a2.htm για τις γενικές.

τελευταίο υπόψη, οργανώσαμε τις λίστες των κανονικών εκφράσεων με τρόπο που οι αναζητήσεις μεγαλύτερων κομματιών λόγου να γίνονται πρώτες. Στο τέλος των αναζητήσεων για κάθε επίπεδο, το κείμενο αποκαθίσταται στην αρχική του μορφή για να συνεχίσει η διαδικασία με την επόμενη λίστα. Επίσης, αφού δεν είχε προηγηθεί συντακτική ανάλυση των κειμένων, σε πολλές περιπτώσεις, χρειάστηκε να συμπεριλάβουμε περισσότερες από μια παραλλαγές κάποιων κανονικών εκφράσεων, με στόχο να μπορέσουμε να ανιχνεύσουμε διαφορετικές περιπτώσεις σύνταξης της ίδια δομής. Ένα τελευταίο πρόβλημα που έπρεπε να αντιμετωπίσουμε, προερχόταν από την έλλειψη σημασιολογικής ανάλυσης. Πολλές λέξεις εμφανίζονταν στους καταλόγους διαφορετικών επιπέδων με άλλες έννοιες, χωρίς να έχουμε εμείς τη δυνατότητα να διακρίνουμε μεταξύ των σημασιών. Σε αυτές τις περιπτώσεις, αποφασίσαμε να προσμετρήσουμε την ανεύρεση στο χαμηλότερο από τα επίπεδα. Για να γίνει αυτό, χρειάστηκε να εντοπίσουμε τις όμοιες ανευρέσεις, να εντοπίσουμε όσες ήταν προϊόν διαφορετικών κανονικών εκφράσεων από διαφορετικά επίπεδα, και να τροποποιήσουμε ανάλογα τις λίστες. Αυτή η διαδικασία έδωσε και ένα ενδιαφέρον υποπροϊόν που θα μπορούσε να αξιοποιηθεί σε εργασίες προσαρμογής κειμένων σε ένα συγκεκριμένο επίπεδο, αφού για να διεκπεραιωθεί χρειάστηκε να καταγράψουμε σε ξεχωριστά αρχεία όλες τις ανευρέσεις, συνοδευόμενες από την ένδειξη του επιπέδου της λίστας από την οποία προέρχονταν.

Καταλήξαμε σε 12 λίστες, που περιλάμβαναν συνολικά πάνω από 15000 κανονικές εκφράσεις, και υπολογίσαμε ισάριθμα χαρακτηριστικά —6 για τις γενικές έννοιες (NG_A1, NG_A2, NG_B1, NG_B2, NG_C1, NG_C2) και έξι για τις ειδικές (NE_A1, NE_A2, NE_B1, NE_B2, NE_C1, NE_C2)— που χρησιμοποιήσαμε στην εκπαίδευση του ταξινομητή μας, θεωρώντας ότι οι δυσκολίες και οι περιορισμοί που περιγράψαμε στην προηγούμενη παράγραφο, επηρεάζουν στον ίδιο βαθμό τις μετρήσεις για όλα τα επίπεδα και δεν αλλοιώνουν τις πραγματικές αναλογίες των εννοιών κάθε επιπέδου που υπάρχουν σε κάθε κείμενο.

Μιας και δεν γνωρίζουμε κάποια άλλη εργασία που να έχει ασχοληθεί με την ποσοτική μελέτη των γενικών και ειδικών εννοιών, θα ήταν ίσως χρήσιμο να κάνουμε μία σύντομη παρουσίαση των αποτελεσμάτων των μετρήσεών μας, παρόλο που αυτού του είδους ο σχολιασμός δεν εντάσσεται άμεσα στους στόχους αυτής της εργασίας. Για τον λόγο αυτό, παραθέτουμε στο Παράρτημα 3 (Εικόνα 8 και Εικόνα 9) τα γραφήματα που παρουσιάζουν την κατανομή των τιμών των σχετικών μεταβλητών στα διάφορα επίπεδα. Στα ίδια γραφήματα, σημειώνεται επίσης και ο μέσος όρος για κάθε επίπεδο. Υπενθυμίζουμε ότι οι τιμές στον κάθετο άξονα αντιστοιχούν σε ποσοστό επί τοις εκατό. Επίσης, χωρίς να κάνουμε εκτεταμένη αναφορά σε στατιστικές αναλύσεις, να πούμε ότι μονοπαραγοντικές αναλύσεις διακύμανσης (one-way ANOVA), έδειξαν ότι οι διαφορές των μέσων όρων, για όλες τις μεταβλητές εκτός από την NE_C2, είναι στατιστικά σημαντικές, με επίπεδο σημαντικότητας 0,05.

Σε πρώτη φάση, παρατηρώντας τις καμπύλες των μέσων όρων, διαπιστώνουμε ότι η υπόθεση που θα μπορούσε να κάνει κανείς, ότι οι έννοιες ενός ορισμένου επιπέδου απαντώνται σε μεγαλύτερο ποσοστό στα κείμενα του ίδιου επιπέδου, δε φαίνεται να ισχύει πλήρως. Για παράδειγμα, όσο αφορά την NE_C2, βλέπουμε ότι ο

μέσος όρος της στα κείμενα του Β2 είναι μεγαλύτερος από τον αντίστοιχο στα κείμενα του Γ2. Παρόλα αυτά, τα αποτελέσματα των γενικών εννοιών, δείχνουν να προσαρμόζονται καλύτερα σε αυτό το μοντέλο. Αυτό εξηγείται πιθανώς από τη φύση των γενικών εννοιών, οι οποίες, όπως έχουμε πει, μπορούν να χρησιμοποιηθούν σε οποιαδήποτε συζήτηση. Αυτό έχει σαν αποτέλεσμα αφενός να μην επηρεάζεται ο αριθμός των ανευρέσεων από θέματα που δεν έχουν προβλεφθεί στο *Πρόγραμμα Σπουδών του Ινστιτούτου Θερβάντες* και αφετέρου να είναι πιθανότερο οι κατάλογοί τους να είναι πιο πλήρεις. Το θέμα της πληρότητας, είναι πιθανό να σχετίζεται και με το γεγονός ότι, όπως βλέπουμε και στον Πίνακα 2, οι ανευρέσεις των γενικών εννοιών είναι περισσότερες από αυτές των ειδικών.

	A1		A2		B1		B2		Γ1		Γ2	
	NG_A1	NE_A1	NG_A2	NE_A2	NG_B1	NE_B1	NG_B2	NE_B2	NG_C1	NE_C1	NG_C2	NE_C2
A1	30,90	15,46	5,69	5,70	5,49	3,13	2,96	2,33	1,73	1,13	0,63	0,66
A2	27,56	10,05	6,54	5,05	6,26	3,80	4,11	3,11	1,85	1,39	0,50	0,59
B1	26,03	9,31	6,41	4,52	6,96	4,00	4,64	3,12	2,89	1,21	0,55	0,63
B2	23,86	7,32	5,20	3,75	6,53	3,83	6,02	3,72	3,26	1,77	0,85	0,85
Γ1	22,82	6,04	4,91	3,18	5,89	3,60	6,11	3,71	4,26	1,89	0,91	0,76
Γ2	22,17	4,37	3,65	2,47	5,40	2,93	6,22	3,96	4,08	2,12	1,43	0,70
Σύνολο	25,56	8,76	5,40	4,11	6,09	3,55	5,01	3,32	3,01	1,58	0,81	0,70

Πίνακας 2. Συνολικά και ανά επίπεδο ποσοστά ανευρέσεων ειδικών και γενικών εννοιών.

NG_A1	215	NE_A1	416
NG_A2	223	NE_A2	632
NG_B1	672	NE_B1	1425
NG_B2	1242	NE_B2	2351
NG_C1	1716	NE_C1	2424
NG_C2	1592	NE_C2	2323
Σύνολο	5660	Σύνολο	9571

Πίνακας 3. Αριθμός κανονικών εκφράσεων ανά λίστα.

Τέλος, έχει ενδιαφέρον να παρατηρήσουμε, σε συνδυασμό με τον Πίνακα 3, ότι αυτός ο μεγαλύτερος αριθμός ανευρέσεων γενικών εννοιών, προκύπτει από μικρότερες λίστες κανονικών εκφράσεων. Επίσης, και για τις ειδικές έννοιες και για τις γενικές, οι ολιγομελείς λίστες των χαμηλών επιπέδων παράγουν πολύ περισσότερες ανευρέσεις από τις λίστες των υψηλότερων επιπέδων, όπου η τάξη μεγέθους του αριθμού των κανονικών εκφράσεων, έχει αλλάξει.

3.3.3 Γραμματικά χαρακτηριστικά.

Ακολούθως, συνεχίσαμε να δουλεύουμε με το *Πρόγραμμα Σπουδών του Ινστιτούτου Θερβάντες* και πιο συγκεκριμένα, με τους καταλόγους που περιέχουν την ύλη της γραμματικής[10]. Σε αυτούς τους καταλόγους η ύλη ειναι οργανωμένη, σε πρώτη φάση, σε επίπεδο λέξης, ανάλογα με το μέρος του λόγου, και έπειτα, σε συντακτικό επίπεδο, σε μεγαλύτερες δομές όπως φράσεις και προτάσεις.

10 http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/niveles/02_gramatica_inventario_a1-a2.htm

Αυτό που κάναμε αρχικά, ήταν να διαλέξουμε μέσα από τους καταλόγους, δομές που να μπορούμε να εντοπίσουμε στα μορφολογικά επισημειωμένα κείμενά μας. Αυτή τη φορά, αντιμετωπίσαμε πολύ περισσότερους περιορισμούς ως προς το τι είναι δυνατό να μετρηθεί από όταν δουλεύαμε με τις γενικές και ειδικές έννοιες. Κάποιοι από αυτούς τους περιορισμούς, οφείλονται στη φύση ορισμένων κανόνων των καταλόγων. Ας πάρουμε τον κανόνα που υπάρχει στο επίπεδο Β1 και λέει ότι δεν μπορούν όλα τα επίθετα να σχηματίσουν επιρρήματα με την κατάληξη *-mente*. Αναφέρεται δηλαδή σε κάτι που θα ήταν λάθος. Είναι προφανές ότι δεν μπορούμε να μετρήσουμε κάτι που δεν θα συναντήσουμε στα κείμενα. Συνέπεια αυτού είναι ότι δεν μπορούμε εντοπίσουμε ούτε σωστές εφαρμογές κανόνων που σχετίζονται με την απουσία κάποιων στοιχείων. Σαν παράδειγμα θα μπορούσαμε να δώσουμε έναν κανόνα που να ορίζει σε ποιες περιπτώσεις δεν χρησιμοποιείται οριστικό άρθρο με το ουσιαστικό. Άλλοι περιορισμοί ήρθαν σαν αποτέλεσμα της έλλειψης συντακτικής και σημασιολογικής ανάλυσης. Ουσιαστικά, αυτό που μπορούσαμε να μετρήσουμε ήταν κάποιες συγκεκριμένες δομές οι οποίες μπορούν να εμφανιστούν σε ένα κείμενο με κάποιες προβλέψιμες μορφές, έτσι ώστε να μπορούν να περιγραφούν από κανονικές εκφράσεις. Ενδεικτικά μπορούμε να αναφέρουμε ότι συμπεριλάβαμε ρηματικές εκφράσεις στις οποίες εμπλέκεται και κάποια απρόσωπη μορφή ρήματος, κειμενικούς δείκτες που συντάσσονται με το ρήμα σε συγκεκριμένο χρόνο ή έγκλιση, κ.ά. Λόγω αυτού του συσχετισμού μεταξύ κειμενικών δεικτών και συγκεκριμένων τρόπων σύνταξης, αποφασίσαμε να μετρήσουμε μαζί με τα γραμματικά χαρακτηριστικά και τους κειμενικούς δείκτες που δεν περιλαμβάνονται στους καταλόγους της γραμματικής, αλλά σε αυτούς των πραγματολογικών τακτικών και στρατηγικών[11].

Φτιάξαμε λοιπόν 5 λίστες με κανονικές εκφράσεις που αντιστοιχούσαν στο υλικό που είχαμε συγκεντρώσει από τους καταλόγους των επιπέδων από Α2-Γ2, οι οποίες περιγράφουν 51, 72, 104, 90 και 48 γραμματικές δομές αντίστοιχα. Το επίπεδο Α1 το αφήσαμε έξω για δύο λόγους. Ο πρώτος είναι ότι ανάμεσα στις ολιγάριθμες καταχωρίσεις του, δεν περιείχε κάτι που μπορούσε να μετρηθεί. Ο δεύτερος, οτί ακόμα και να μπορούσαμε να μετρήσουμε κάτι, επειδή θα ήταν απολύτως βασικό, ήταν βέβαιο ότι θα απαντάται —και πιθανότατα στον ίδιο βαθμό— στα κείμενα όλων των επιπέδων. Δεν θα μας παρείχε, δηλαδή, κάποια επιπλέον πληροφορία που να μας βοηθά να διακρίνουμε μεταξύ των επιπέδων, αντίθετα από αυτό που συνέβη στις περιπτώσεις των ειδικών και γενικών εννοιών, όπου η μετρήσεις για το Α1 ήταν χρήσιμες. Η διαφορά είναι ότι με τις μεν έννοιες, είχαμε εκτεταμένες λίστες, που κάλυπταν μεγάλο μέρος της ύλης, και μετρήσαμε το ποσοστό των λέξεων του κειμένου που εμπίπτει σε αυτές, στη δε γραμματική, μπορούμε μόνο να αναζητήσουμε επιλεκτικά κάποια μικρά κομμάτια της συνολικής ύλης και μετράμε κάτι διαφορετικο από ποσοστά. Μετράμε σε κάθε κείμενο τον αριθμό των ανευρέσεων γραμματικών φαινομένων κάθε επιπέδου (G_A2, G_B1, G_B2, G_C1, G_C2), ανά 100 λέξεις κειμένου. Να σημειωθεί ότι δόθηκε έμφαση στο να αποφευχθούν οι ψευδείς ανιχνεύσεις και για αυτό χρειάστηκε να συνοδεύσουμε κάποιες κανονικές εκφράσεις από άλλες, δευτερεύουσες, οι οποίες έλεγχαν την

11 http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/niveles/06_tacticas_pragmaticas_inventario_a1-a2.htm

εγκυρότητα της ανίχνευσης. Για να επιτευχθεί αυτό, χρειάστηκε να εξαγάγουμε σε ξεχωριστά αρχεία και να ελέγξουμε όλες τις ανιχνέυσεις. Όπως και στην περίπτωση των γενικών και ειδικών εννοιών, τέτοιου είδους καταγραφή θα μπορούσε να είναι χρήσιμη κατά τη διαδικασία προσαρμογής κειμένων. Στο Παράρτημα 3(Εικόνα 10), παραθετουμε επίσης διαγράμματα με τα αποτελέσματα των μετρήσεων και για αυτές τις μεταβλητές.

Στα γραμματικά χαρακτηριστικά που μετρήσαμε με τις παραπάνω λίστες κανονικών εκφράσεων, δεν συμπεριλαμβάνονται οι χρόνοι και οι εγκλίσεις των ρημάτων. Θα μπορούσαμε να πούμε ότι η εκμάθηση καθενός από αυτά αποτελεί ορόσημο στην πορεία του μαθητή προς την κατάκτηση της γλώσσας. Είναι σύνηθες, μαθητές ισπανικών που ερωτώνται σε τι επίπεδο βρίσκονται, να δίνουν απαντήσεις όπως «έχω κάνει τον αόριστο» ή «μπήκα στην υποτακτική». Γενικά, υπάρχει μια αρκετά σταθερή σειρά με την οποία γίνεται η παρουσίαση των διάφορων χρόνων και εγκλίσεων στους μαθητές. Το γεγονός αυτό έχει επισημανθεί και από την Checa-García (2013), η οποία μέτρησε για την έρευνά της τον αριθμό των διαφορετικών ρηματικών τύπων σε κάθε κείμενο. Εμείς αποφασίσαμε να προχωρήσουμε σε μια καταγραφή όλων των διαφορετικών χρόνων και εγκλίσεων που εμφανίζονται στα κείμενά μας. Μετρήσαμε λοιπόν τα παρακάτω 22 χαρακτηριστικά, των οποίων οι τιμές αντιπροσωπεύουν τα επί τοις εκατό ποσοστά επί του συνολικού αριθμού των ρημάτων του κειμένου:

- Ενεστώτας – Οριστική (Prind)
- Ενεστώτας – Υποτακτική (Prsub)
- Παρατατικός – Οριστική (Plind)
- Παρατατικός – Υποτακτική (Plsub)
- Παρακείμενος – Οριστική (PPind)
- Παρακείμενος – Υποτακτική (PPsub)
- Υπερσυντέλικος – Οριστική (Plind)
- Υπερσυντέλικος – Υποτακτική (Plsub)
- Μέλλοντας – Οριστική (Futind)
- Μέλλοντας – Υποτακτική (Futsub)
- Πρότερος Αόριστος – Οριστική (Paind)
- Συντελεσμένος Μέλλοντας – Οριστική (FPind)
- Απλός Δυνητικός – Οριστική (CS)
- Σύνθετος Δυνητικός – Οριστική (CC)
- Αόριστος – Οριστική (Indef)
- Σύνθετο Απαρέμφατο (Cinf)
- Σύνθετο Γερούνδιο (Cger)
- Ρηματική έκφραση με το ρήμα Estar + Γερούνδιο (EstGER)
- Προστακτική (Imp)
- Απαρέμφατο (Inf)
- Γερούνδιο (Ger)
- Παθητική Μετοχή (Par)

3.3.4 Χαρακτηριστικά με βάση τα μέρη του λόγου (POS-based).

Η επίδραση των αναλογιών των λέξεων που κατατάσσονται στα διάφορα μέρη του λόγου στην αναγνωσιμότητα έχει μελετηθεί από διάφορους ερευνητές (Feng et al.,

2010; Heilman et al., 2007; Kane, Carthy, & Dunnion, 2006). Στη δικιά μας περίπτωση, μετρήσαμε τα παρακάτω χαρακτηριστικά, τα οποία, εκτός από το τελευταίο, εκφράζονται σε ποσοστά επί τοις εκατό, επί του συνόλου των λέξεων:

- Αντωνυμίες (Pron)
- Ρήματα (Verb)
- Επιρρήματα (Adv)
- Επίθετα (Adj)
- Προσδιοριστές (Det)
- Ουσιαστικα (Nouns)
- Ονοματικές Οντότητες (NE)
- Συμπλεκτικοί Σύνδεσμοι (ConC)
- Υποτελείς Σύνδεσμοι (ConS)
- Επιφωνήματα (Inter)
- Προθέσεις (Prep)
- Προσωπικές αντωνυμίες ανά ρήμα (PPpV): Ο λόγος του αριθμού των προσωπικών αντωνυμιών προς τον αριθμό των ρημάτων του κειμένου.

3.3.5 Συντακτικά χαρακτηριστικά.

Όπως έχουμε πει ήδη, τα κείμενά μας δεν υποβλήθηκαν σε συντακτική επεξεργασία και έτσι δεν έγιναν μετρήσεις πάνω σε συντακτικά δέντρα. Προσπαθήσαμε όμως να αποτυπώσουμε τη συντακτικη πολυπλοκότητα τους δουλεύοντας πάλι με τους κειμενικούς δείκτες που είχαμε συμπεριλάβει και στις μετρήσεις των γραμματικών χαρακτηριστικών, στηριζόμενοι στην ιδέα ότι όσο περισσότεροι κειμενικοί δείκτες υπάρχουν σε ένα κείμενο, τόσο μεγαλύτερη θα είναι και η συντακτική του πολυπλοκότητα. Αυτή τη φορά, φτιάξαμε μία μοναδική λίστα η οποία περιλαμβάνει όλες τις σχετικές με κειμενικούς δείκτες κανονικές εκφράσεις, που ήταν πριν διεσπαρμένες στις λίστες των διαφόρων επιπέδων, και μετρήσαμε τον αριθμό των ανευρέσεών τους ανά 100 λέξεις κειμένου (Mark).

Στον Πίνακας 11 του Παραρτήματος 3, καταγράφονται όλα τα χαρακτηριστικά, στο σύνολό τους 81, που χρησιμοποιήσαμε.

3.4 Περιβάλλον εκπαίδευσης και αλγόριθμος.

Ο αλγόριθμος που επιλέξαμε για την εκπαίδευση του ταξινομητή μας ήταν οι Μηχανές Διανυσμάτων Υποστήριξης (SVM). Οι Μηχανές Διανυσμάτων Υποστήριξης έχουν χρησιμοποιηθεί με επιτυχία σε πληθώρα εργασιών κατηγοριοποίησης κειμένου, γενικότερα, και αναγνωσιμότητας ειδικότερα (Larsson, 2006; Liu, Croft, Oh, & Hart, 2004; Petersen & Ostendorf, 2009; Pitler & Nenkova, 2008; Schwarm & Ostendorf, 2005; Tanaka-Ishii, Tezuka, & Terada, 2010; Wang, 2006). Η συγκεκριμένη υλοποίηση που χρησιμοποιήσαμε στη δική μας περίπτωση, είναι αυτή του αλγορίθμου Σειριακής Ελάχιστης Βελτιστοποίησης (SMO) που περιλαμβάνεται στην πλατφόρμα μηχανικής μάθησης ανοιχτού κώδικα WEKA[12]. Όσο αφορά τις παραμέτρους του αλγόριθμου, αφήσαμε τις προτερόθετες ρυθμίσεις μιας και ήταν αυτές που έδιναν τα καλύτερα αποτελέσματα. Η αξιολόγηση των μοντέλων που παρήγαγαν τα πειράματα που θα περιγράψουμε στο επόμενο κεφάλαιο, έγινε με

[12] http://www.cs.waikato.ac.nz/ml/weka/

την τεχνική της δεκάπτυχης διασταυρούμενης επικύρωσης (10-fold cross validation).

4. Αποτελέσματα, συμπεράσματα και μελλοντική δουλειά.

4.1 Αποτελέσματα των πειραμάτων και σχολιασμός τους.

Πριν περάσουμε στον σχολιασμό των αποτελεσμάτων, να κάνουμε μιά αναφορά στα μεγέθη που χρησιμοποιήσαμε για την αξιολόγησή τους:

- Ακρίβεια (accuracy): Το ποσοστό των κειμένων που κατέταξε ο ταξινομητής στο σωστό επίπεδο.
- Ορθότητα (precision): Ο λόγος των σωστά ταξινομημένων σε κάθε επίπεδο κειμένων προς το σύνολο των ταξινομημένων στο ίδιο επίπεδο κειμένων.
- Ανάκληση (recall): Ο λόγος των σωστά ταξινομημένων σε κάθε επίπεδο κειμένων προς το σύνολο των κειμένων που ανήκουν στο επίπεδο αυτό.
- F-measure: Ο αρμονικός μέσος της ορθότητας και της ανάκλησης, δίνεται από τον τύπο:

  F-measure = 2*(Ορθότητα * Ανάκληση)/ (Ορθότητα + Ανάκληση)

  Χρησιμοποιούμε αυτήν την εκδοχή του τύπου, με τον συντελεστή 2, επειδή θέλουμε να αποδώσουμε την ίδια σημασία στην Ορθότητα και την Ανάκληση κατά τον υπολογισμό του F-measure.

Τα πρώτα βήματα που ακολουθήσαμε κατά την πειραματική διαδικασία, ήταν να εκπαιδεύσουμε κάποιους ταξινομητές χρησιμοποιώντας μεμονωμένες ομάδες χαρακτηριστικών, για να ελέγξουμε κατά πόσο επηρεάζουν την επιτυχία της ταξινόμησης. Στη συνέχεια προχωρήσαμε στην εκπαίδευση μοντέλων με τη συμμετοχή συδυασμών ομάδων χαρακτηριστικών, προσπαθώντας να καταλήξουμε στον ταξινομητή που κατηγοριοποιεί καλύτερα τα κείμενα μας.

Η πρώτη ομάδα χαρακτηριστικών που δοκιμάσαμε ήταν τα υφομετρικά. Πήραμε τα εξής αποτελέσματα:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.556	0.068	0.636	0.556	0.593	0.888	A1
	0.449	0.192	0.36	0.449	0.4	0.722	A2
	0.309	0.16	0.262	0.309	0.283	0.615	B1
	0.271	0.101	0.348	0.271	0.305	0.656	B2
	0.231	0.059	0.4	0.231	0.293	0.801	C1
	0.544	0.141	0.425	0.544	0.477	0.82	C2
Weighted Avg.	0.4	0.122	0.408	0.4	0.397	0.751	

```
=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 35 19  9  0  0  0 |  a = A1
 16 31 13  5  1  3 |  b = A2
  3 17 17 10  1  7 |  c = B1
  0 12 17 16  5  9 |  d = B2
  0  4  7  6 12 23 |  e = C1
  1  3  2  9 11 31 |  f = C2
```

Εικόνα 1. Αποτελέσματα με χρήση μόνο υφομετρικών χαρακτηριστικών (όλα τα κείμενα).

Η συνολική ακρίβεια του μοντέλου είναι 40% και για τα μισά επίπεδα, το F-measure γύρω στο 0,3. Βλέπουμε ότι δε θα μπορούσαμε να βασιστούμε μόνο στα υφομετρικά

χαρακτηριστικά για να φτιάξουμε έναν αποτελεσματικό ταξινομητή. Σε αυτό το σημείο, θελήσαμε να ελέγξουμε την επιρροή άλλων παραγόντων στα υφομετρικά χαρακτηριστικά, εκτός από το επίπεδο γλωσσομάθειας. Όπως έχουμε δει και στην περιγραφή του corpus, τα κείμενα μας παρουσιάζουν μια μεγάλη ετερογενεια ως προς τα είδη τους. Οι διακυμάνσεις στις τιμές των υφομετρικών χαρακτηριστικών, ίσως να οφείλονται περισσότερο στην ετερογένεια αυτή παρά σε διαφορές των επιπέδων, με αποτέλεσμα να μην μπορούν να αποτυπώσουν τις τελευταίες. Είναι λογικό, για παράδειγμα, να υποθέσουμε ότι υπάρχουν έντονες υφολογικές διαφορές μεταξύ ενός προσωπικού διαλόγου και ενός άρθρου γνώμης. Αποφασίσαμε, λοιπόν, να ελέγξουμε τη συμπεριφορά των υφομετρικών χαρακτηριστικών έχοντας προσπαθήσει να εξαλείψουμε την παραπάνω ετερογένεια. Για τον λόγο αυτόν, δουλέψαμε με ένα υποσύνολο του corpus στο οποίο συμπεριλάβαμε μόνο ενημερωτικά και πληροφοριακά κείμενα, 138 τον αριθμό, και πήραμε τα παρακάτω αποτελέσματα:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.45	0.068	0.529	0.45	0.486	0.888	A1
	0.48	0.159	0.4	0.48	0.436	0.792	A2
	0.25	0.11	0.278	0.25	0.263	0.704	B1
	0.4	0.142	0.385	0.4	0.392	0.647	B2
	0.273	0.069	0.429	0.273	0.333	0.814	C1
	0.385	0.205	0.303	0.385	0.339	0.768	C2
Weighted Avg.	0.377	0.13	0.385	0.377	0.376	0.766	

```
=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
  9  9  2  0  0  0 |  a = A1
  5 12  4  4  0  0 |  b = A2
  2  5  5  4  0  4 |  c = B1
  1  4  2 10  1  7 |  d = B2
  0  0  1  3  6 12 |  e = C1
  0  0  4  5  7 10 |  f = C2
```

Εικόνα 2. Αποτελέσματα με χρήση μόνο υφομετρικών χαρακτηριστικών (ενημερωτικά και πληροφοριακά κείμενα).

Τα αποτελέσματα είναι ελαφρώς χειρότερα, γεγονός που δείχνει ότι η αποτελεσματικότητα των υφομετρικών χαρακτηριστικών δεν περιορίζεται από την ποικιλία στα κειμενικά είδη του corpus μας.

Σε μια προσπάθεια να βελτιώσουμε τα αποτελέσματα που παίρνουμε από τα υφομετρικά χαρακτηριστικά και λόγω του ότι τα μισά και πλέον από αυτά σχετίζονται με το φάσμα συχνοτήτων μήκους λέξεων, αποφασίσαμε να ελέγξουμε τις τιμές των τελευταίων για να δούμε αν η κατανομές τους στα διάφορα επίπεδα αποτυπώνουν κάποιο μοτίβο. Παρατηρώντας την καμπύλη των μέσων όρων ανά επίπεδο για καθένα από αυτά τα 14 χαρακτηριστικά, προσέξαμε ότι σε τρεις περιπτώσεις, για τα LW_4, LW_5, LW_12, η σχέση μέσων όρων – επιπέδων ήταν μονοτονική, φθίνουσα για τα πρώτα δύο και αύξουσα για το τρίτο (Παράρτημα 3, Εικόνα 11). Επιπλέον, και για τις τρεις αυτές μεταβλητές, όπως έδειξαν οι σχετικές μονοπαραγοντικές αναλύσεις διακύμανσης, οι διαφορές είναι στατιστικά σημαντικές με επίπεδο σημαντικότητας 0,001. Αφαιρέσαμε όλες τις υπόλοιπες μεταβλητές που

σχετίζονται με το φάσμα συχνότητας μήκους λέξεων και εκπαιδεύσαμε ένα μοντέλο που παρουσίαζε 41,41% ακρίβεια. Στα πειράματα που ακολουθούν, όπου εμπλέκονται υφομετρικά χαρακτηριστικά, η σύνθεση τους θα είναι αυτή, με τρεις μεταβλητές από το φάσμα συχνότητας μήκους λέξεων αντί για 14.

Το μοντέλο που προέκυψε από τη χρήση αποκλειστικά λεξιλογικών χαρακτηριστικών έδωσε συνολική ακρίβεια 42.82% και τα εξής αποτελέσματα:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.508	0.062	0.64	0.508	0.566	0.867	A1
	0.58	0.22	0.388	0.58	0.465	0.763	A2
	0.109	0.053	0.273	0.109	0.156	0.633	B1
	0.424	0.186	0.313	0.424	0.36	0.693	B2
	0.269	0.069	0.4	0.269	0.322	0.764	C1
	0.614	0.101	0.538	0.614	0.574	0.888	C2
Weighted Avg.	0.428	0.119	0.428	0.428	0.414	0.77	

=== Confusion Matrix ===

a	b	c	d	e	f	<-- classified as
32	21	2	5	3	0	a = A1
11	40	8	8	0	2	b = A2
6	22	6	15	4	2	c = B1
1	15	4	25	6	8	d = B2
0	4	2	14	14	18	e = C1
0	1	0	13	8	35	f = C2

Εικόνα 3. Αποτελέσματα με τη χρήση μόνο λεξιλογικών χαρακτηριστικών.

Ούτε αυτή τη φορά μπορούμε να πούμε ότι προκύπτει ένας αξιοπρεπής ταξινομητής. Τα επίπεδα Α1 και Α2 βρίσκονται σταθερά ανάμεσα σε αυτά που καλύτερα αναγνωρίζονται αλλά αυτή τη φορά έχουμε ικανοποιητικά αποτελέσματα και για το Γ2. Θα μπορούσαμε να πούμε ότι ο συγκεκριμένος ταξινομητής, έχει την τάση να αποδίδει, σε μεγάλο βαθμό, κείμενα στο επίπεδο Α2 και σε λιγότερο μεγάλο βαθμό στο Γ2, αφού και τα δύο αυτά επίπεδα παρουσιάζουν μεγάλη ανάκληση και αρκετά μικρότερη ορθότητα.

Στο επόμενο πείραμα χρησιμοποιήθηκαν τα γραμματικά χαρακτηριστικά μαζί με το μοναδικό συντακτικό, γιατί αφενός παρουσιάζουν κάποια συνάφεια και αφετέρου δεν είχε νόημα να εκπαιδεύσουμε έναν ταξινομητή με ένα μόνο χαρακτηριστικό. Τα αποτελέσματα που πήραμε είχαν την παρακάτω εικόνα:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.778	0.065	0.721	0.778	0.748	0.943	A1
	0.725	0.112	0.61	0.725	0.662	0.877	A2
	0.436	0.077	0.511	0.436	0.471	0.79	B1
	0.508	0.145	0.411	0.508	0.455	0.755	B2
	0.308	0.043	0.552	0.308	0.395	0.791	C1
	0.456	0.101	0.464	0.456	0.46	0.789	C2
Weighted Avg.	0.549	0.092	0.549	0.549	0.542	0.828	

```
=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 49 12  0  1  0  1 |  a = A1
 13 50  2  2  0  2 |  b = A2
  2  9 24 10  3  7 |  c = B1
  0  5 10 30  4 10 |  d = B2
  1  2  6 17 16 10 |  e = C1
  3  4  5 13  6 26 |  f = C2
```

Εικόνα 4. Αποτελέσματα από τη χρήση του συντακτικού και των γραμματικών χαρακτηριστικών.

Η βελτίωση είναι αισθητή, όπως μαρτυρά η αυξημένη ακρίβεια 54,93% αλλά και η κάπως ομαλότερη κατανομή των λάνθασμένων προβλέψεων στα επίπεδα γύρω από τα σωστά. Το F-measure είναι ιδιαίτερα ψηλό για τα δύο χαμηλότερα επίπεδα, τα οποία δείχνουν και σε αυτήν την περίπτωση να είναι αυτά που διαχωρίζονται καλύτερα. Φαίνεται ότι οι διαφορές στην γραμματική πολυπλοκότητα είναι πιο αδρές μεταξύ των μεγαλυτέρων επιπέδων. Στην ερμηνεία αυτών των αποτελεσμάτων, θα πρέπει να λάβουμε υπόψη και τη φύση των χαρακτηριστικών που μετρήσαμε. Τα περισσότερα από αυτά σχετίζονται με τους διάφορους χρόνους και εγκλίσεις των ρημάτων. Η παρουσίαση αυτών των φαινομένων, σε μια τυπική οργάνωση της ύλης, έχει ολοκληρωθεί μέχρι τα αρχικά στάδια του επιπέδου Β2. Θα μπορούσαμε να πούμε ότι οι γραμματικές γνώσεις των μαθητών αυξάνονται με μεγαλύτερα βήματα όσο διανύουν τα πρώτα επίπεδα και αυτό έχει σαν αποτέλεσμα οι μεταβάσεις μεταξύ αυτών των επιπέδων να διακρίνονται πιο εύκολα από τον συγκεκριμένο ταξινομητή.

Η τελευταία μεμονωμένη ομάδα χαρακτηριστικών με την οποία δημιουργήσαμε ένα μοντέλο, ήταν αυτή που σχετίζεται με τις αναλογίες των μερών του λόγου στα κείμενα. Αν και η ακρίβεια που παρουσίασε αυτός ο ταξινομητής είναι συγκρίσιμη με αυτή των δύο πρώτων μοντέλων, τα αποτελέσματα στην πραγματικότητα ήταν αρκετά χειρότερα. Το F-measure ήταν 0,338 αλλά ο ταξινομητής δεν κατέταξε κανένα κείμενο στο επίπεδο Γ1 και στο Β1 κατέταξε μόλις 9, εκ των οποίων σωστά τα 2. Είναι προφανές ότι το να χρησιμοποιηθούν αυτά τα χαρακτηριστικά μόνα τους, δεν έχει κάποια αξία.

Μετά από αυτά τα πειράματα, εκπαιδεύσαμε ένα μοντέλο χρησιμοποιώντας όλες τις ομάδες χαρακτηριστικών το οποίο και κρατήσαμε ως τον τελικό ταξινομητή μας. Καταλήξαμε σε αυτόν, γιατί παρά τις δοκιμές που ακολούθησαν, οι οποίες συμπεριλάμβαναν συνδυασμούς διαφόρων ομάδων χαρακτηριστικών, αφαίρεση μεμονωμένων χαρακτηριστικών και πειραματισμούς με την κατάταξη των χαρακτηριστικών με βάση το πληροφοριακό τους κέρδος, δεν προέκυψε κάποιο

καλύτερο μοντέλο. Η συνολική ακρίβεια του ταξινομητή ήταν 65,07% και τα αναλυτικά αποτελέσματα φαίνονται στην παρακάτω εικόνα:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.81	0.034	0.836	0.81	0.823	0.966	A1
	0.754	0.08	0.693	0.754	0.722	0.9	A2
	0.527	0.083	0.537	0.527	0.532	0.793	B1
	0.542	0.101	0.516	0.542	0.529	0.819	B2
	0.481	0.059	0.581	0.481	0.526	0.836	C1
	0.737	0.06	0.7	0.737	0.718	0.908	C2
Weighted Avg.	0.651	0.07	0.65	0.651	0.649	0.873	

```
=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 51 11  1  0  0  0 |  a = A1
  8 52  8  1  0  0 |  b = A2
  2 10 29 10  1  3 |  c = B1
  0  1 11 32 10  5 |  d = B2
  0  0  3 14 25 10 |  e = C1
  0  1  2  5  7 42 |  f = C2
```

Εικόνα 5. Επιδόσεις του ταξινομητή, που εκπαιδεύτηκε με χρήση όλων των ομάδων χαρακτηριστικών.

Πέρα από την ικανοποιητική γενική ακρίβεια του ταξινομητή, θεωρούμε πολύ ενθαρρυντική την κατανομή των προβλέψεών του. Παρατηρούμε ότι το μεγαλύτερο ποσοστό λάθος προβλέψεων αφορά επίπεδα διπλανά του σωστού. Έτσι, για το επίπεδο Α1 το 91,66% των λάθος προβλέψεων εντοπίζεται στο Α2, για το Α2 το 94,12% εντοπίζεται στα Α1 και Β1 και αντίστοιχα για το Β1 το 76,92%, για το Β2 το 77,77%, για το Γ1 το 88,88% και για το Γ2 το 46,66%. Παρατηρούμε επίσης ότι, εκτός από την περίπτωση του Β1 του οποίου 1 κείμενο έχει ταξινομηθεί στο Γ1 και 3 στο Γ2, όσο απομακρυνόμαστε από το σωστό επίπεδο, ο αριθμός των λάθος προβλέψεων μειώνεται. Δεν πρέπει να ξεχνάμε ότι δεν υπάρχουν σαφείς διαχωριστικές γραμμές μεταξύ των επιπέδων και ότι το φάσμα της δυσκολίας είναι συνεχές. Για αυτό και η ταξινόμηση των κειμένων σε επίπεδα δυσκολίας δεν είναι εύκολη ούτε για ανθρώπους. Αυτό ήταν κάτι που διαπιστώσαμε από πρώτο χέρι κατά την εκπόνηση αυτής της εργασίας, τόσο από την προσωπική μας εμπειρία όσο και από τις αποκλίσεις στις ελάχιστες αξιολογήσεις κειμένων που καταφέραμε να συλλέξουμε από μαθητές και τις αναφορές παρόμοιων περιπτώσεων στη βιβλιογραφία. Όλα αυτά μας κάνουν να πιστεύουμε ότι τα χαρακτηριστικά που επιλέξαμε για την εκπαίδευση του ταξινομητή μας, όντως αποτυπώνουν την σχετική με τη δυσκολία των κειμένων πραγματικότητα.

Ένα άλλο πράγμα που αξίζει να σχολιαστεί σε σχέση με τα αποτελέσματα, είναι οι διαφορές στην απόδοση του ταξινομητή που παρατηρούνται μεταξύ των επιπέδων. Παρατηρούμε ότι για τα επίπεδα Α1, Α2 και Γ2 το F-measure είναι σημαντικά μεγαλύτερο από ότι στα υπόλοιπα, όπου κυμαίνεται γύρω στο 0,53. Την μεγαλύτερη επιτυχία στα επίπεδα Α1 και Γ2 μπορούμε να την αποδώσουμε εν μέρει στο ότι, σαν ακραία επίπεδα που οριοθετούν το φάσμα της δυσκολίας, έχουν απώλειες μόνο προς τη μία πλευρά του φάσματος, σε αντίθεση με τα ενδιάμεσα επίπεδα τα οποία χάνουν προβλέψεις και προς τα χαμηλότερα και προς τα

υψηλότερα επίπεδα. Οι χαμηλότερες επιδόσεις στα επίπεδα Β1-Γ1, μπορούν να αποδοθούν, όπως εξηγήσαμε και λίγο παραπάνω, στις πιο αδρές διαφορές που υπάρχουν μεταξύ τους, τουλάχιστον ως προς το είδος των χαρακτηριστικών που λάβαμε υπόψη στην έρευνά μας.

Με αφορμή τις παρατηρήσεις που κάναμε παραπάνω σχετικά με την κατανομή των λάθος προβλέψεων στα διπλανά επίπεδα, αποφασίσαμε ότι θα άξιζε τον κόπο να εκπαιδεύσουμε έναν δεύτερο ταξινομητή ο οποίος να κατατάσσει τα κείμενα στα τρία ευρύτερα σταδια του ΚΕΠΑΓ, δηλαδή στα Α,Β και Γ, ελπίζοντας ότι θα πετύχουμε πολύ καλύτερα αποτελέσματα. Κάτι τέτοιο θα είχε αξία για δύο λόγους. Ο πρώτος είναι ότι, σε πολλές περιπτώσεις, δεν γίνεται η διάκριση στα επιμέρους επίπεδα 1 και 2 κάθε σταδίου ούτε και από τις μεθόδους διδασκαλίας ή άλλα βιβλία που αξιοποιούνται στην εκμάθηση των ισπανικών. Υπάρχουν δηλαδή ενιαία εγχειρίδια για τα επίπεδα Α1-Α2 ή Β1-Β2, χωρίς να υπάρχει μέσα σε αυτά ένα σαφές σημείο στο οποίο να γινέται η μετάβαση από το ένα επίπεδο στο άλλο. Το ίδιο ισχύει και για βιβλία με διαβαθμισμένα αναγνώσματα, γραμματικής και άλλα. Το ίδιο το ΚΕΠΑΓ αναφέρει ότι ο καθορισμός οριακών σημείων μεταξύ των επιπέδων είναι μια υποκειμενική διαδικασία, και για αυτό δίνει παραδείγματα περιπτώσεων που θα μπορούσαν να ακολουθηθούν διαφορετικοί τρόποι υποδιαίρεσης των ευρύτερων σταδίων. Ο δεύτερος λόγος είναι ότι αν θεωρήσουμε ότι η βασική χρησιμότητα ενός ταξινομητή, όπως την περιγράψαμε στην εισαγωγή, είναι ο εντοπισμός κειμένων που θα χρησιμοποιηθούν συμπληρωματικά κατά τη διδακτική διαδικασία, ο μαθητής μπορεί να επωφεληθεί από αυτά έστω και αν είναι λίγο δυσκολότερα ή ευκολότερα από το ακριβές του επίπεδο, αν θεωρήσουμε ότι αυτό μπορεί να εκτιμηθεί με ακρίβεια.

Ακολουθώντας, για τους πειραματισμούς μας, την ίδια τακτική με πριν, εκπαιδεύσαμε μοντέλα με μεμονωμένες ομάδες χαρακτηριστικών. Για να έχουμε απλά μια εικόνα των επιδόσεών τους, παραθέτουμε τις γενικές ακρίβειες που πέτυχαν.

Ομάδα χαρακτηριστικών	**Συνολική ακρίβεια**
Υφομετρικά	67,04%
Υφομετρικα -ΦΣΜΛ	69,01%
Λεξιλογικά	63,94%
Γραμματικά +Συντακτικό	73,24%
Μέρη του λόγου	62,53%

Πίνακας 4. Ταξινομητής 3 κατηγοριών. Ακρίβειες μοντέλων μεμονωμένων ομάδων χαρακτηριστικών.

Επαναλαμβάνοντας την διερεύνηση της επίδρασης του φάσματος συχνοτήτων μήκους λέξεων που κάναμε και στην περίπτωση του ταξινομητή 6 επιπέδων, καταλήξαμε ότι στην περίπτωση των 3 επιπέδων παίρνουμε καλύτερα αποτελέσματα αν το αφαιρέσουμε τελείως. Με αυτήν την επέμβαση, βλέπουμε ότι η επίδοση των υφομετρικών χαρακτηριστικών είναι συγκρίσιμη με αυτή των γραμματικών, στα οποία προστέθηκαν για άλλη μια φορά και οι κειμενικοί δείκτες. Η χρήση, δηλαδή, των υφομετρικών χαρακτηριστικών είναι πολύ πιο αποδοτική,

όταν δουλεύουμε με τρία επίπεδα. Πολύ μεγάλη βελτίωση παρατηρούμε και με την ομάδα χαρακτηριστικών των μερών του λόγου, η οποία, ενώ πριν δεν ήταν αξιοποιήσιμη από μόνη της, τώρα έχει συγκρίσιμα αποτελέσματα με τα λεξιλογικά χαρακτηριστικά. Φαίνεται ότι οι δύο ομάδες που παρουσίασαν τη βελτίωση, παρόλο που δεν μπορούσαν να αποτυπώσουν τις λεπτότερες διαφορές μεταξύ των 6 επιπέδων, τα καταφέρνουν καλύτερα με τις πιο απότομες μεταβάσεις μεταξύ 3 επιπέδων.

Το επόμενο βήμα ήταν, πάλι, να δοκιμάσουμε συνδυασμούς ομάδων χαρακτηριστικών. Τα καλύτερα αποτελέσματα τα πήραμε με συνδυασμό όλων των ομάδων —χωρίς το φάσμα συχνοτήτων μήκους λέξεων— που έδωσε συνολική ακρίβεια 83,38%. Με αφετηρία αυτό το μοντέλο, δοκιμάσαμε να αφαιρέσουμε διάφορα χαρακτηριστικά με βάση το πληροφοριακό τους κέρδος, με τη βοήθεια του σχετικού φίλτρου του WEKA, χωρίς να πετύχουμε κάποια βελτίωση. Έπειτα, προχωρήσαμε σε δοκιμές με διαδοχικές αφαιρέσεις όλων των χαρακτηριστικών και καταλήξαμε στο τελικό μας μοντέλο, το οποίο δεν περιλαμβάνει τα χαρακτηριστικά WLsd (3), SLsd (5), Dis_HapL (8), LD (9), Yule (10), NE_C2 (42), PPsub (52), Futsub (56), Paind (57), Cger (63), ConC (76), Inter (78). Η συνολική του ακρίβεια είναι 85,63% και οι αναλυτικές του επιδόσεις οι εξής:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.947	0.031	0.947	0.947	0.947	0.972	A
	0.798	0.112	0.771	0.798	0.784	0.843	B
	0.807	0.069	0.838	0.807	0.822	0.915	C
Weighted Avg.	0.856	0.069	0.857	0.856	0.857	0.913	

=== Confusion Matrix ===

a	b	c	<-- classified as
125	7	0	a = A
6	91	17	b = B
1	20	88	c = C

Εικόνα 6. Επιδόσεις τελικού ταξινομητή για τρία επίπεδα.

Μπορούμε να πούμε ότι τα αποτελέσματα είναι πολύ ικανοποιητικά. Ακόμα και αν λάβουμε υπόψη την αναμενόμενη αύξηση της ακρίβειας λόγω της μείωσης των επιπέδων, ο καινούριος ταξινομητής είναι πιο επιτυχημένος από τον παλιό. Κάνοντας τις απαράιτητες αναγωγές σε 3 επίπεδα, τα αποτελέσματα του ταξινομητή 6 επιπέδων θα ήταν τα ακόλουθα:

Α	**Β**	**Γ**	**<-Ταξινομημένα ως**	**Ορθότητα**	**Ανάκληση**	**F-measure**
122	10	0	**Α**	0,897	0,924	0,910
13	82	19	**Β**	0,707	0,719	0,713
1	24	84	**Γ**	0,816	0,771	0,792
			Μεσοσταθμικές τιμές	0,811	0,811	0,811

Πίνακας 5. Αποτελέσματα ταξινομητή 6 επιπέδων με αναγωγές σε 3 επίπεδα.

Παρατηρώντας τα αποτελέσματα βλέπουμε ότι και σε αυτή την περίπτωση το πρώτο επίπεδο παρουσιάζει τις καλύτερες επιδόσεις. Πιστεύουμε ότι κάτι τέτοιο, ιδίως συνοδευόμενο από μια εξαιρετική τιμή ορθότητας, έχει ιδιαίτερη αξία γιατί τα αυθεντικά κείμενα που ανήκουν στα χαμηλά επίπεδα είναι πολύ πιο δυσεύρετα. Επειδή, ακριβώς, τα αυθεντικά κείμενα απευθύνονται σε φυσικούς ομιλητές, είναι λογικό τις περισσότερες φορές να είναι πολύ δύσκολα για μαθητές των χαμηλών επιπέδων. Αντίθετα, είναι πολύ πιο εύκολο να βρει ένας καθηγητής αυθεντικά κείμενα για μαθητές ψηλότερων επιπέδων, καταφεύγοντας στον τύπο, σε ιστολόγια και αλλού. Ένας ταξινομητής, σαν τον παραπάνω, ο οποίος να αναγνωρίζει με μεγάλη ορθότητα κείμενα χαμηλών επιπέδων, είναι ιδιαίτερα χρήσιμος.

Βλεποντας την επιτυχία αυτού του μοντέλου και θέλοντας να εκμεταλλευτούμε τις βελτιωμένες του επιδόσεις, αποφασίσαμε να εκπαιδεύσουμε έναν τελευταίο ταξινομητή για 6 επίπεδα, προσθέτοντας ως χαρακτηριστικό τις προβλέψεις του ταξινομητή τριών επιπέδων. Να σημειώσουμε εδώ, ότι οι προβλέψεις που χρησιμοποιήσαμε δεν ήταν ακριβώς αποτέλεσμα του τελικού ταξινομητή των τριών επιπέδων, αλλά των 10 επιμέρους μοντέλων που παρήχθησαν κατά τη διαδικασία της δεκάπτυχης διασταυρούμενης επικύρωσης. Αν χρησιμοποιούσαμε τον τελικό ταξινομητή για να πάρουμε προβλέψεις, αυτές θα παρουσίαζαν την πλασματικά μεγαλύτερη ακρίβεια που εμφανίζεται όταν αξιολογείται ένα μοντέλο πάνω στο σώμα εκπαίδευσής του. Για αυτόν τον λόγο, δεν θα ήταν αξιοποιήσιμες οι προβλέψεις στην εκπαίδευση του άλλου ταξινομητή. Έτσι λοιπόν, οι προβλέψεις που πάρθηκαν για κάθε δέκατο μέρος του σώματος κειμένων προέρχονταν από διαφορετικό μοντέλο, στην εκπαίδευση του οποίου δεν συμμετείχε το αντίστοιχο μέρος του corpus.

Η ιδέα αυτή φανηκε να δίνει καρπούς, αφού η συνολική ακρίβεια του ταξινομητή βελτιώθηκε κατά 2,25% και ανέβηκε στο 67,32%, με τα αναλυτικά αποτελέσματα να έχουν την παρακάτω μορφή:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.825	0.031	0.852	0.825	0.839	0.968	A1
	0.783	0.063	0.75	0.783	0.766	0.903	A2
	0.6	0.08	0.579	0.6	0.589	0.821	B1
	0.559	0.095	0.541	0.559	0.55	0.853	B2
	0.442	0.053	0.59	0.442	0.505	0.832	C1
	0.772	0.07	0.677	0.772	0.721	0.912	C2
Weighted Avg.	0.673	0.065	0.672	0.673	0.67	0.885	

```
=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 52 11  0  0  0  0 |  a = A1
  8 54  6  1  0  0 |  b = A2
  1  6 33 11  1  3 |  c = B1
  0  0 14 33  7  5 |  d = B2
  0  0  3 13 23 13 |  e = C1
  0  1  1  3  8 44 |  f = C2
```

Εικόνα 7. Αποτελέσματα με την προσθήκη των προβλέψεων του ταξινομητή τριών επιπέδων.

Σημειώθηκε άυξηση στο F-measure, από 0,03 έως 0,057, για όλα τα επίπεδα εκτός από το Γ1 για το οποίο μειώθηκε κατά 0,021, συνοδευόμενο από μια τάση να κατατάσσονται περισσότερα κείμενα αυτού του επιπέδου, στο Γ2. Το τελευταίο αντικατοπτρίζει μια γενικότερη, αναμενόμενη, μετατόπιση λανθασμένων προβλέψεων, από εκτός του σωστού ευρύτερου σταδίου —Α,Β ή Γ— προς αυτό, ασχέτα με το αν μετουσιώθηκαν σε προβλέψεις του σωστού επιπέδου.

Κλείνοντας αυτήν την ενότητα, θα ήταν καλό να συνοψίσουμε κάποια συμπεράσματα που βγήκαν κατά την εξέλιξη των πειραμάτων. Το πρώτο είναι ότι η μεγάλη ποικιλία στα κειμενικά είδη που παρουσιάζει το corpus μας, δεν φάνηκε να επηρεάζει την απόδοση των υφομετρικών χαρακτηριστικών, όπως διαπιστώσαμε από το πείραμα με το υποσύνολο του corpus που περιλάμβανε μόνο πληροφοριακά και ενημερωτικά κείμενα. Ένα δεύτερο συμπέρασμα είναι ότι η ομάδα χαρακτηριστικών που από μόνη της έδωσε τα καλύτερα αποτελέσματα, ήταν αυτή των γραμματικών χαρακτηριστικών. Αυτό σχετίζεται με την σημασία της σειράς παρουσίασης των γραμματικών φαινομένων στη διάρθρωση της ύλης των προγραμμάτων σπουδών. Μια άλλη παρατήρηση είναι ότι για τις ακραίες βαθμίδες της κλίμακας δυσκολίας, δηλαδή για τα επίπεδα Α1 και Γ2, οι ταξινομητές παρουσιασαν αυξημένη ακρίβεια, ένα τμήμα της οποίας μπορεί να εξηγηθεί από το ότι αυτά τα επίπεδα συνορεύουν μόνο με ένα επίπεδο το καθένα και έτσι έχουν απώλειες μόνο προς αυτό. Το ότι οι ταξινομητές έχουν μεγαλύτερη επιτυχία με τα χαμηλοτερα επίπεδα, μπορεί ίσως να εξηγηθεί από τη μη γραμμικότητα της κλίμακας δυσκολίας. Όπως επισημαίνεται και στο ΚΕΠΑΓ, οι αποστάσεις μεταξύ των επιπέδων δεν είναι ίσες αλλά αυξάνονται όσο αυτά μεγαλώνουν. Ο χρόνος, δηλαδή, που χρειάζεται ένας μαθητής για να πάει από το Α1 στο Α2 είναι πολύ λιγότερος από το χρόνο της μετάβασης από το Γ1 στο Γ2. Αντίστοιχα, αυξάνεται και η ύλη κάθε επιπεδου. Αυτό απότυπώνεται και στους καταλόγους του Προγράμματος Σπουδών του Ινστιτούτου Θερβάντες. Θεωρούμε ότι όσο πιο περιορισμένη είναι η ύλη των επιπέδων τόσο πιο ευδιάκριτες είναι οι διαχωριστικές γραμμές μεταξύ τους, γιατί τα ίδια τα επίπεδα μπορούν να οριοθετηθούν πιο εύκολα. Στο πρακτικο μέρος, που μας αφορά, η πιο περιορισμένη ύλη είναι και πιο εύκολα μετρήσιμη. Για παράδειγμα, όσο λιγότερα γραμματικά φαινόμενα περιέχει η ύλη ενός επιπέδου, τόσο μεγαλύτερο ποσοστό τους μπορούμε να κωδικοποιήσουμε και να μετρήσουμε. Έτσι, αφενός έχουμε μια πληρέστερη περιγραφή του επιπέδου και αφετέρου η καταγραφή τους είναι πιο συνεπής, δεν είναι σποραδική.

4.2 Μελλοντική έρευνα.

Οι δύο βασικοί άξονες στους οποίους θεωρούμε ότι πρέπει να κινηθούν οι μελλοντικές προσπάθειες για βελτίωση των ταξινομητών, είναι η προσθήκη επιπλέον χαρακτηριστικών στην εκπαίδευσή τους και η διεύρυνση του σώματος κειμένων, με τον πρώτο στόχο να εξαρτάται ως ένα βαθμό από τον δεύτερο.

Ως προς τον πρώτο άξονα, είδαμε ότι η δουλειά μας με το *Πρόγραμμα Σπουδών του Ινστιτούτου Θερβάντες* είχε θετικά αποτελέσματα, και σκεφτόμαστε ότι θα μπορούσε να επεκταθεί και σε άλλους καταλόγους του, με πρώτο υποψήφιο αυτόν των γλωσσικών λειτουργιών. Πιστεύουμε επίσης ότι η αναζήτηση νέων χαρακτηριστικών θα πρέπει να στοχεύει, κυρίως, στην βελτίωση της απόδοσης του ταξινομητή στα πιο προβληματικά επίπεδα. Να γίνει, δηλαδή, μια διερεύνηση του τι

μπορεί να βοηθήσει στον καλύτερο διαχωρισμό των κειμένων του Γ1, μιας και αυτό είναι το επίπεδο με το χαμηλότερο F-measure, από τα κείμενα των γειτονικών του επιπέδων. Αυτή η διερεύνηση θα μπορούσε, εκτός από τα νέα χαρακτηριστικά, να αφορά και αυτά που έχουμε ήδη χρησιμοποιήσει. Πιο συγκεκριμένα, η ιδέα είναι να ξεχωρίσουμε από τις υπάρχουσες λίστες κάποιες έννοιες ή γραμματικά φαινόμενα τα οποία είτε θα εντοπίζονται αποκλειστικά σε κείμενα επιπέδου ίσου ή μεγαλύτερου από αυτό στου οποίου τις λίστες ανήκουν, είτε θα καταγράφουν μια στατιστικά σημαντική διαφορά μεταξύ των εμφανίσεών τους σε διαφορετικά επίπεδα, και να τα ομαδοποιήσουμε σε νέες λίστες ή, ακόμα, κάποια από αυτά να τα χρησιμοποιήσουμε ως αυτόνομα χαρακτηριστικά. Το πρόβλημα είναι ότι ακόμα και αν υπάρχουν τέτοια χαρακτηριστικά, είναι δύσκολο να τα εντοπίσουμε δουλεύοντας με ένα σώμα κειμένων του μεγέθους του δικού μας. Ας δώσουμε ένα παράδειγμα με τα γραμματικά φαινόμενα του επιπέδου Γ2. Οι ανιχνεύσεις τους, όπως μπορούμε να διαπιστώσουμε και στο αντίστοιχο γράφημα της Εικόνα 10 του Παραρτήματος 3, είναι πολύ αραιές. Ακόμα και ανάμεσα στα κείμενα του επιπέδου Γ2, μπορούμε να βρούμε μολίς 11 στα οποία να έχει ανιχνευτεί έστω ένα φαινόμενο από τη λίστα. Ας υποθέσουμε ότι ανάμεσα σε αυτά τα φαινόμενα, υπάρχουν κάποια που αποκλείεται να εντοπισθούν σε κείμενα μικρότερων επιπέδων και τα οποία θα μπορούσαν να χρησιμοποιηθούν ως μεμονωμένα χαρακτηριστικά, οριοθετώντας, έτσι, καλύτερα το επίπεδο Γ2. Στο σύνολο 33.677 λέξεων των κειμένων του επιπέδου Γ2 του corpus μας, οι ανιχνεύσεις τους θα ήταν από ελάχιστες έως ανύπαρκτες. Αυτό σημαίνει, από τη μία, ότι από τα δεδομένα μας, δεν θα μπορούσαμε να αναγνωρίσουμε πια είναι αυτά τα σημαντικά γραμματικά φαινόμενα, και από την άλλη, ότι ακόμα και αν με κάποιο τρόπο τα γνωρίζαμε, δεν θα μπορούσαμε να τα αξιοποιήσουμε σαν χαρακτηριστικά στην εκπαίδευση του ταξινομητή, αφού οι τιμές τους για όλα τα κείμενα, θα ήταν πιθανότατα 0. Αυτό ήταν ένα σενάριο που όντως αντιμετωπίσαμε κατά τις μετρήσεις μας. Μπορούμε να αναφέρουμε την περίπτωση του πρότερου αορίστου, ο οποίος είναι ένας χρόνος που έχει πέσει σε αχρηστία στα σύγχρονα ισπανικά, που απαντάνται μόνο σε λόγια γραπτά κείμενα, που συμπεριλαμβάνεται στην ύλη του Γ2 και που η παρουσία του θα μπορούσε να αποτελεί μια ισχυρή ένδειξη ότι ένα κείμενο ανήκει στο επίπεδο αυτό. Δεν ανιχεύτηκε ούτε μία φορά στο corpus μας.

Άλλος ένας περιορισμός που μας επιβλήθηκε από το μικρό μέγεθος του σώματος κειμένων, είναι ότι δεν μπορέσαμε να χρησιμοποιήσουμε στατιστικά γλωσσικά μοντέλα. Τους λόγους τους έχουμε εξηγήσει ήδη, στην αρχή της ενότητας των Κειμενικών Χαρακτηριστικών. Είναι κάτι που πιστεύουμε ότι θα βελτιώσει πολύ τους ταξινομητές μας και που μας ενδιαφέρει έντονα να κάνουμε στο μέλλον, με την προϋπόθεση ότι θα έχουμε στη διάθεσή μας περισσότερα δεδομένα.

Με βάση τα παραπάνω, αυτό που βλέπουμε στην παρούσα φάση ως επόμενο βήμα, είναι να βγάλουμε την μέχρι τώρα δουλειά μας προς τα έξω, με τη μορφή μιας online εφαρμογής. Αυτό αποβλέπει σε δύο πράγματα. Πρώτο, θεωρούμε σημαντικό να αξιολογηθούν οι ταξινομητές μας σε επιπλέον κείμενα. Δεύτερο, πιστεύουμε ότι μπορεί να εξελιχθεί σε έναν αποτελεσματικο τρόπο να διευρύνουμε χωρίς πολύ κόπο το σώμα κειμένων μας. Αυτό μπορεί να επιτευχθεί με τη συνεργασία χρηστών που θα έχουν σχέση με την διδασκαλία των ισπανικών ως ξένη γλώσσα και θα είναι διατεθειμένοι να σχολιάσουν τα αποτελέσματα των ταξινομητών για τα κείμενα που

θα υποβάλλουν για εκτίμηση, δηλώνοντας την προέλευση του κειμένου, το κειμενικό είδος στο οποίο ανήκει και, το βασικότερο, το επίπεδο στο οποίο πιστεύουν αυτοί ότι ανήκει. Θα ήταν πολύ χρήσιμο να υπάρχουν χρήστες που να υποβάλλουν κείμενα των οποίων το επίπεδο γνωρίζουν με σιγουριά, απλά και μόνο για να βοηθήσουν στην αξιολόγηση των ταξινομητών και στον εμπλουτισμό του corpus.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Barzilay, R., & Lapata, M. (2008). Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*. doi:10.1162/coli.2008.34.1.1

Brück, T. Der, Hartrumpf, S., Helbig, H., & Hagen, F. (2008). A Readability Checker with Supervised Learning Using Deep Indicators. *Informatica*, *32*(4), 429–435.

Checa-García, I. (2013). Complejidad gramatical y niveles de dificultad en lecturas de ELE adaptadas y originales. *Revista de Lingüística Teórica Y Aplicada*, *51*(2), 49–72. Retrieved from http://www.scielo.cl/pdf/rla/v51n2/art_04.pdf

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, *56*(13), 1448–1462. doi:10.1002/asi.20243

Contreras, a, García-Alonso, R., Echenique, M., & Daye-Contreras, F. (1999). The SOL formulas for converting SMOG readability scores between health education materials written in Spanish, English, and French. *Journal of Health Communication*, *4*(1), 21–9. doi:10.1080/108107399127066

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. *Council of Europe* (pp. 1–273). Retrieved from http://www.coe.int/t/dg4/linguistic/CADRE1_EN.asp#TopOfPage

Criado, R. (University of M.), & Sánchez, A. (University of M. (2009). Vocabulary in EFL Textbooks. A Contrastive Analysis against Three Corpus-Based Word Ranges. In *A survey on corpus-based research/panorama de investigaciones basadas en corpus* (pp. 862–875). Murcia: Editum. Retrieved from http://digitum.um.es/xmlui/handle/10201/13848

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, *27*(1), 11–28 CR – Copyright © 1948 Taylor & Francis. doi:10.2307/1473169

Das, S., & Roychoudhury, R. (2006). Readability modelling and comparison of one and two parametric fit: A case study in Bangla*. *Journal of Quantitative Linguistics*, *13*(1), 17–34. doi:10.1080/09296170500500843

DuBay, W. (2004). The principles of readability. 2004. *Costa Mesa: Impact Information*, (949). Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Principles+of+Readability#1

DuBay, W. (2007). The Classic Readability Studies. *Online Submission*. Retrieved from http://eric.ed.gov/?id=ED506404

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (Vol. Poster Vol, pp. 276–284). Retrieved from http://dl.acm.org/citation.cfm?id=1944598

Flesh, R. (1948). A new readability yardstick. *The Journal of Applied Psychology*, *32*, 221–233. doi:10.1037/h0057532

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc*, *36*(2), 193–202. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15354684

Gray, W. S., & Leary, B. E. (1935). What makes a book readable, with special reference to adults of limited reading ability. Chicago, Ill.: The University of Chicago press. Retrieved from http://catalog.hathitrust.org/Record/001176549

Gunning, R. (1952). *The Technique of Clear Writing*. New York: Mcgraw-Hill.

Hargis, G. (2000). Readability and computer documentation. *ACM Journal of Computer Documentation*. doi:10.1145/344599.344634

Harmer, J. (2003). The Practice of English Language Teaching. *TESOL Quarterly*, *37*(1), 179–181. Retrieved from http://www.jstor.org/discover/10.2307/3588472?uid=2129&uid=2134&uid=2&uid=70&uid=4&sid=21104739903593

Hedge, T. (2000). *Teaching and learning in the language classroom*. *Language Teaching* (Vol. 56). Oxford University Press. doi:10.1093/elt/56.3.337

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Computational Linguistics*, (April), 460–467. Retrieved from http://acl.ldc.upenn.edu/N/N07/N07-1058.pdf

Huang, H., & Liou, H. (2007). Vocabulary learning in an automated graded reading program. *Language Learning & Technology*, *11*(3), 64–82.

Huizenga, J., & Ruzic, M. T. (1994). *Reading Workout*. Heinle & Heinle Publishers.

Instituto Cervantes- Biblioteca nueva. (2006). *Plan curricular*. Madrid.

James N., F., James J., J., & Donald G., P. (1951). Simplification of Flesch Reading Ease Formula. *Journal of Applied Psychology*, *35*(5), 333–337. doi:10.1037/h0062427

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (1st ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

Kane, L., Carthy, J., & Dunnion, J. (2006). *Readability Applied to Information Retrieval* (pp. 523–526). Springer Berlin Heidelberg. doi:10.1007/11735106_56

Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., … Welty, C. (2010). Learning to Predict Readability using Diverse Linguistic Features. *Computational Linguistics*, (August), 546–554.

Klare, G. R. (1963). The measurement of readability. Ames: Iowa State University Press. Retrieved from http://catalog.hathitrust.org/Record/001111077

Krashen, S. D. (1982). *Principles and Practice in Second Language Acquisition. The Modern Language Journal* (Vol. 67, p. 168). doi:10.2307/328293

Larsson, P. (2006). *Classification into Readability Levels*. Uppsala Universitet.

Liontou, J. (2012). Examining text difficulty through automated textual analysis tools and readers' beliefs : the case of the Greek State Certificate of English Language Proficiency exam. *Research Papers in Language Teaching and Learning*, *3*(1), 64–77. Retrieved from http://rpltl.eap.gr

Liu, X., Croft, W. B., Oh, P., & Hart, D. (2004). Automatic Recognition of Reading Levels from User Queries. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 548–549). New York.

McLaughlin, G. (1969). SMOG grading: A new readability formula. *Journal of Reading*, 639–646. Retrieved from http://www.jstor.org/stable/40011226

Miguel García Arreza, María Dolores Zamora Navas, J. J. S. B. (1994). *La lengua inglesa en la educación primaria*. Málaga : Aljibe: Archidona.

Montalbán, F. Á. (2007). El uso de material auténtico en la enseñanza de ELE. In *FIAPE. II Congreso internacional: Una lengua, muchas culturas.* (pp. 26–29). Granada,. Retrieved from http://www.mecd.gob.es/dctm/redele/Material-RedEle/Numeros Especiales/2007_ESP_12_II Congreso FIAPE/Talleres/2007_ESP_12_13Alvarez.pdf?documentId=0901e72b80e67299&ei=Q2rrU_2hFqrMyAPPhYKACQ&usg=AFQjCNFmuXq8m7HfQm_J10YDzAWPP9d8MA&sig2=qlnZA3EzaMo-Wu_8NcmnEw&bvm=bv.72938740,d.bGQ&cad=rja

Oakes, M. (1998). Statistics for corpus linguistics. Retrieved from http://eprints.lancs.ac.uk/id/eprint/11573

Parker, R. I., Hasbrouck, J. E., & Weaver, L. (2001). Spanish Readability Formulas for Elementary-Level Texts: a Validation Study. *Reading & Writing Quarterly*, *17*(4), 307–322. doi:10.1080/105735601317095052

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language*, *23*(1), 89–106. doi:10.1016/j.csl.2008.04.003

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. *Proceedings of the Conference on Empirical …*, (October), 186–195. Retrieved from http://dl.acm.org/citation.cfm?id=1613742

Real Academia Española. (2010). *Nueva gramática de la lengua española. Nueva gramática de la lengua española*. Madrid: Espasa Libros.

Redish, J. (2000). Readability formulas have even more limitations than Klare discusses. *ACM Journal of Computer Documentation (JCD)*, *24*(3), 132–137. Retrieved from http://dl.acm.org/citation.cfm?id=344637

Richards, J. C. (2001). *Curriculum Development in Language Teaching*. Cambridge University Press. doi:http://dx.doi.org/10.1017/CBO9780511667220

Schriven, K. A. (2000). Readability Formulas in the New Millennium : What ' s the Use ? *ACM Journal of Computer Documentation (JCD)*, *24*(3), 138–140. doi:10.1145/344599.344638

Schwarm, S. E., & Ostendorf, M. (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 523–530). Stroudsburg, PA, USA. doi:10.3115/1219840.1219905

Si, L., & Callan, J. (2001). A statistical model for scientific readability. *Proceedings of the Tenth International Conference on Information and Knowledge Management - CIKM'01*, 574. Retrieved from http://portal.acm.org/citation.cfm?doid=502585.502695

Tanaka-Ishii, K., Tezuka, S., & Terada, H. (2010). Sorting Texts by Readability. *Computational Linguistics*, *36*(2), 203–227. doi:10.1162/coli.2010.09-036-R2-08-050

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes. Mind in Society The Development of Higher Psychological Processes* (Vol. Mind in So, p. 159). doi:10.1007/978-3-540-92784-6

Wang, Y. (2006). Automatic recognition of text difficulty from consumers health information. In *19th IEEE International Symposium on Computer-Based Medical Systems* (Vol. 2006, pp. 131–136). Salt Lake City, UT: IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1647558

Μικρός, Γ. Κ. (n.d.). *Δυσκολία κατανόησης του ξενόγλωσσου κειμένου και υφομετρία . Μια νέα προσέγγιση στην αναγνωσιμότητα κειμένων από έλληνες που μαθαίνουν την Ιταλική ως ξένη γλώσσα* . Retrieved from http://users.uoa.gr/~gmikros/Pdf/Readability (final).pdf

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα 1 – *Κοινό Ευρωπαϊκό Πλαίσιο Αναφοράς για τις Γλώσσες.*

Βασικός χρήστης	Α1	Μπορεί να κατανοήσει και να χρησιμοποιήσει καθημερινές εκφράσεις που του είναι οικείες και πολύ βασικές φράσεις που έχουν στόχο την ικανοποίηση συγκεκριμένων αναγκών. Μπορεί να συστηθεί και να συστήσει άλλους και μπορεί να ρωτήσει και να απαντήσει ερωτήσεις που αφορούν προσωπικά στοιχεία, όπως το πού μένει, τα άτομα που γνωρίζει και τα πράγματα που κατέχει. Μπορεί να συνδιαλλαγεί με απλό τρόπο υπό την προϋπόθεση ότι ο συνομιλητής του μιλάει αργά και καθαρά και είναι διατεθειμένος να βοηθήσει.
	Α2	Μπορεί να κατανοήσει προτάσεις και εκφράσεις που χρησιμοποιούνται συχνά και που σχετίζονται με περιοχές που είναι άμεσα συναφείς (π.χ. πολύ βασικές ατομικές και οικογενειακές πληροφορίες, αγορές, τοπική γεωγραφία, εργασία). Μπορεί να επικοινωνήσει σε απλά και συνηθισμένα καθήκοντα που απαιτούν απλή και απευθείας ανταλλαγή πληροφοριών για θέματα που του είναι οικεία και για θέματα ρουτίνας. Μπορεί να περιγράψει με απλά λόγια πτυχές του ιστορικού του, του άμεσου περιβάλλοντός του καθώς και θέματα άμεσης ανάγκης.
Ανεξάρτητος χρήστης	Β1	Μπορεί να κατανοήσει τα κύρια σημεία που του παρουσιάζονται με σαφήνεια και χωρίς αποκλίσεις από τον κοινό γλωσσικό τύπο και που αφορούν θέματα που συναντώνται τακτικά στη δουλειά, στο σχολείο, στον ελεύθερο χρόνο, κτλ. Μπορεί να χειριστεί καταστάσεις που είναι πιθανό να προκύψουν στη διάρκεια ενός ταξιδιού σε μια περιοχή όπου ομιλείται η γλώσσα. Μπορεί να παραγάγει απλό κείμενο σχετικό με θέματα που γνωρίζει ή που τον αφορούν προσωπικά. Μπορεί να περιγράψει εμπειρίες και γεγονότα, όνειρα, ελπίδες και φιλοδοξίες και να δώσει συνοπτικά λόγους και εξηγήσεις για τις γνώμες και τα σχέδιά του.
	Β2	Μπορεί να κατανοήσει τις κύριες ιδέες ενός σύνθετου κειμένου, τόσο για συγκεκριμένα, όσο και για αφηρημένα θέματα, συμπεριλαμβανομένων συζητήσεων πάνω σε τεχνικά ζητήματα της ειδικότητάς του. Μπορεί να συνδιαλλαγεί με κάποια άνεση και αυθορμητισμό που καθιστούν δυνατή τη συνήθη επικοινωνία με φυσικούς ομιλητές της γλώσσας χωρίς επιβάρυνση για κανένα από τα δύο μέρη. Μπορεί να παραγάγει σαφές, λεπτομερές κείμενο για ένα ευρύ φάσμα θεμάτων και να εξηγήσει μια άποψη πάνω σε ένα κεντρικό ζήτημα, δίνοντας τα πλεονεκτήματα και τα μειονεκτήματα των διαφόρων επιλογών.
Ικανός χρήστης	Γ1	Μπορεί να κατανοήσει ένα ευρύ φάσμα απαιτητικών, μακροσκελών κειμένων και να αναγνωρίσει σημασίες που υπονοούνται. Μπορεί να εκφραστεί άνετα και αυθόρμητα χωρίς να φαίνεται συχνά πως αναζητά εκφράσεις. Μπορεί να χρησιμοποιεί τη γλώσσα ευέλικτα και αποτελεσματικά για κοινωνικούς, ακαδημαϊκούς και επαγγελματικούς σκοπούς. Μπορεί να παραγάγει σαφή, καλά διαρθρωμένα, λεπτομερή κείμενα για σύνθετα θέματα, επιδεικνύοντας ελεγχόμενη χρήση οργανωτικών σχημάτων, συνδετικών στοιχείων και μηχανισμών συνοχής.

	Γ2	Μπορεί να κατανοήσει με ευκολία σχεδόν όλα όσα ακούει ή διαβάζει. Μπορεί να κάνει περιλήψεις με βάση πληροφορίες που προέρχονται από διαφορετικές προφορικές ή γραπτές πηγές, ανασυνθέτοντας επιχειρήματα και περιγραφές σε μια συνεκτική παρουσίαση. Μπορεί να εκφραστεί αυθόρμητα, με μεγάλη άνεση και ακρίβεια, διαχωρίζοντας λεπτές σημασιολογικές αποχρώσεις ακόμα και σε ιδιαίτερα σύνθετες περιστάσεις.

Πίνακας 6. Κοινά Επίπεδα Αναφοράς: σφαιρική κλίμακα.

A1	A2	B1
Μπορώ να αναγνωρίσω γνωστές λέξεις και πολύ στοιχειώδεις φράσεις που αφορούν εμένα, την οικογένεια μου και το άμεσο συγκεκριμένο περιβάλλον, όταν οι άνθρωποι μιλούν αργά και καθαρά	Μπορώ να κατανοήσω φράσεις και λέξεις υψηλής συχνότητας που σχετίζονται με περιοχές άμεσης προσωπικής συνάφειας (π.χ. πολύ στοιχειώδεις προσωπικές και οικογενειακές πληροφορίες, αγορές, τοπική γεωγραφία, εργασία). Μπορώ να συλλάβω την κεντρική ιδέα σε περιπτώσεις σύντομων, σαφών, απλών μηνυμάτων και ανακοινώσεων.	Μπορώ να κατανοήσω τα κύρια σημεία μιας σαφούς ομιλίας χωρίς ιδιωματικά χαρακτηριστικά σε γνωστά θέματα που συναντώ τακτικά στη δουλειά, στο σχολείο, στον ελεύθερο χρόνο, κτλ. Μπορώ να κατανοήσω τα κύρια σημεία πολλών ραδιοφωνικών ή τηλεοπτικών ενημερωτικών εκπομπών σχετικών με την τρέχουσα επικαιρότητα ή εκπομπών με θέματα προσωπικού ή επαγγελματικο ύ ενδιαφέροντος όταν η παρουσίαση είναι σχετικά αργή και ξεκάθαρη.
B2	**Γ1**	**Γ2**
Μπορώ να κατανοήσω εκτεταμένο προφορικό λόγο και διαλέξεις και να παρακολουθήσω ακόμα και σύνθετα επιχειρήματα, υπό την προϋπόθεση ότι το θέμα δεν είναι ιδιαίτερα άγνωστο. Μπορώ να καταλάβω τις περισσότερες ειδησεογραφικές και ενημερωτικές εκπομπές της τηλεόρασης. Μπορώ να καταλάβω τις περισσότερες ταινίες στην καθιερωμένη διάλεκτο.	Μπορώ να κατανοήσω εκτεταμένο προφορικό λόγο ακόμα και όταν δεν είναι δομημένος ξεκάθαρα ή και όταν οι σχέσεις μεταξύ των λεγομένων απλώς υπονοούνται αντί να δηλώνονται ρητά. Μπορώ να καταλάβω τηλεοπτικές εκπομπές και ταινίες χωρίς ιδιαίτερη προσπάθεια.	Δεν έχω δυσκολία να κατανοήσω οποιοδήποτε είδος προφορικού λόγου, είτε ζωντανά είτε σε εκπομπή, ακόμα και όταν εκφωνείται με μεγάλη ταχύτητα φυσικού ομιλητή, υπό την προϋπόθεση ότι έχω λίγο χρόνο να εξοικειωθώ με την προφορά.

Πίνακας 7. Κοινά Επίπεδα Αναφοράς: πλέγμα αυτοαξιολόγησης κατανόησης γραπτού λόγου.

Παράρτημα 2 – *Πρόγραμμα Σπουδών του Ινστιτούτου Θερβάντες.*

1. Κατάλογοι γραμματικού περιεχομένου:
   1.1. Γραμματική.
   1.2. Προφορά και προσωδία.
   1.3. Ορθογραφία.
2. Κατάλογοι πραγματολογικού-κειμενικού περιεχομένου:
   2.1. Γλωσσικές λειτουργίες.
   2.2. Πραγματολογικές τακτικές και στρατηγικές.
   2.3. Κειμενικά είδη και παραγωγή λόγου.
3. Κατάλογοι εννοιολογικού περιεχομένου.
   3.1. Γενικές έννοιες.
   3.2. Ειδικές έννοιες.
4. Κατάλογοι πολιτισμικού περιεχομένου.
   4.1. Πολιτισμικά σημεία αναφοράς.
   4.2. Κοινωνικοπολιτισμικές γνώσεις και συμπεριφορές.
   4.3. Διαπολιτισμικές δεξιότητες και στάσεις.
5. Κατάλογοι μαθησιακού περιεχομένου:
   5.1. Μαθησιακές διαδικασίες.

1. Componente gramatical:
   1.1. Gramática.
   1.2. Pronunciación y prosodia
   1.3. Ortografía.
2. Componente pragmático-discursivo:
   2.1. Funciones.
   2.2. Tácticas y estrategias pragmáticas.
   2.3. Géneros discursivos y productos textuales.
3. Componente nocional:
   3.1. Nociones generales.
   3.2. Nociones específicas.
4. Componente cultural:
   4.1. Referentes culturales.
   4.2. Saberes y comportamientos socioculturales.
   4.3. Habilidades y actitudes interculturales.
5. Componente de aprendizaje:
   5.1. Procedimientos de aprendizaje.

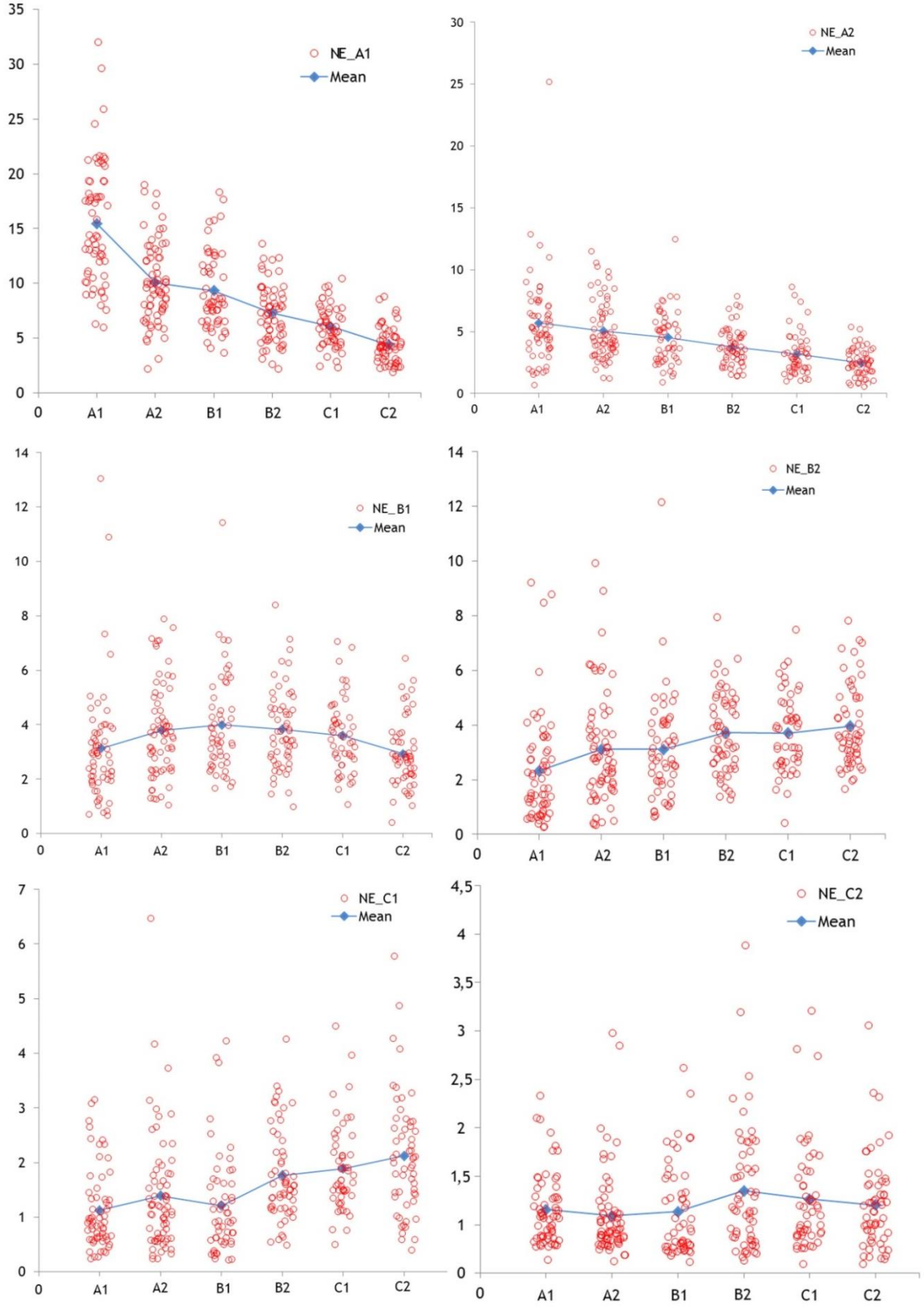
Πίνακας 8. Οργάνωση των καταλόγων του Προγράμματος Σπουδών του Ινστιτούτου Θερβάντες.

1.3. Géneros de transmisión escrita (R) (P): recepción y producción; (R): solo recepción; (P): solo producción	
A1	**A2**
Billetes (de transporte, de banco) (R) Carteles en hoteles, tiendas, supermercados, mercados; directorios de centros comerciales. (R) Diccionarios bilingües (R) Formularios (datos personales) (R) (P) Hojas y folletos con información turística (R) Horarios de establecimientos y transporte público (R) Menús del día, turísticos o en establecimientos de comida rápida (R) Notas muy breves y sencillas (R) Postales y mensajes electrónicos, breves y sencillos (R) (P)	Anuncios publicitarios relacionados con alojamiento, establecimientos hoteleros y viajes (R) Biografías breves y sencillas (R) (P) Cartas de restaurantes y menús (R) Carteleras de espectáculos (R) Carteles en restaurantes, estaciones de tren, lugares de trabajo. (R) Cuentos breves en versión simplificada (R) Diarios breves y pautados (R) (P) Etiquetas de productos y embalajes (R) Formularios (inscripciones, matrículas.)(R) (P) Hojas y folletos informativos y publicitarios sencillos (R) Horóscopos (R) Informaciones meteorológicas (R) Informes breves y predecibles sobre temas familiares (R) Listas de precios y productos (R) Notas y mensajes breves (R) Notas y mensajes muy breves y sencillos sobre áreas de necesidad inmediata (P) Noticias de actualidad altamente predecibles sobre temas conocidos (R) Ofertas de trabajo (R) Postales, cartas y mensajes electrónicos, personales y breves, de presentación, agradecimiento, excusa, invitación... (R) (P) Programación de radio y televisión (R) Recetas de cocina breves y sencillas (R)

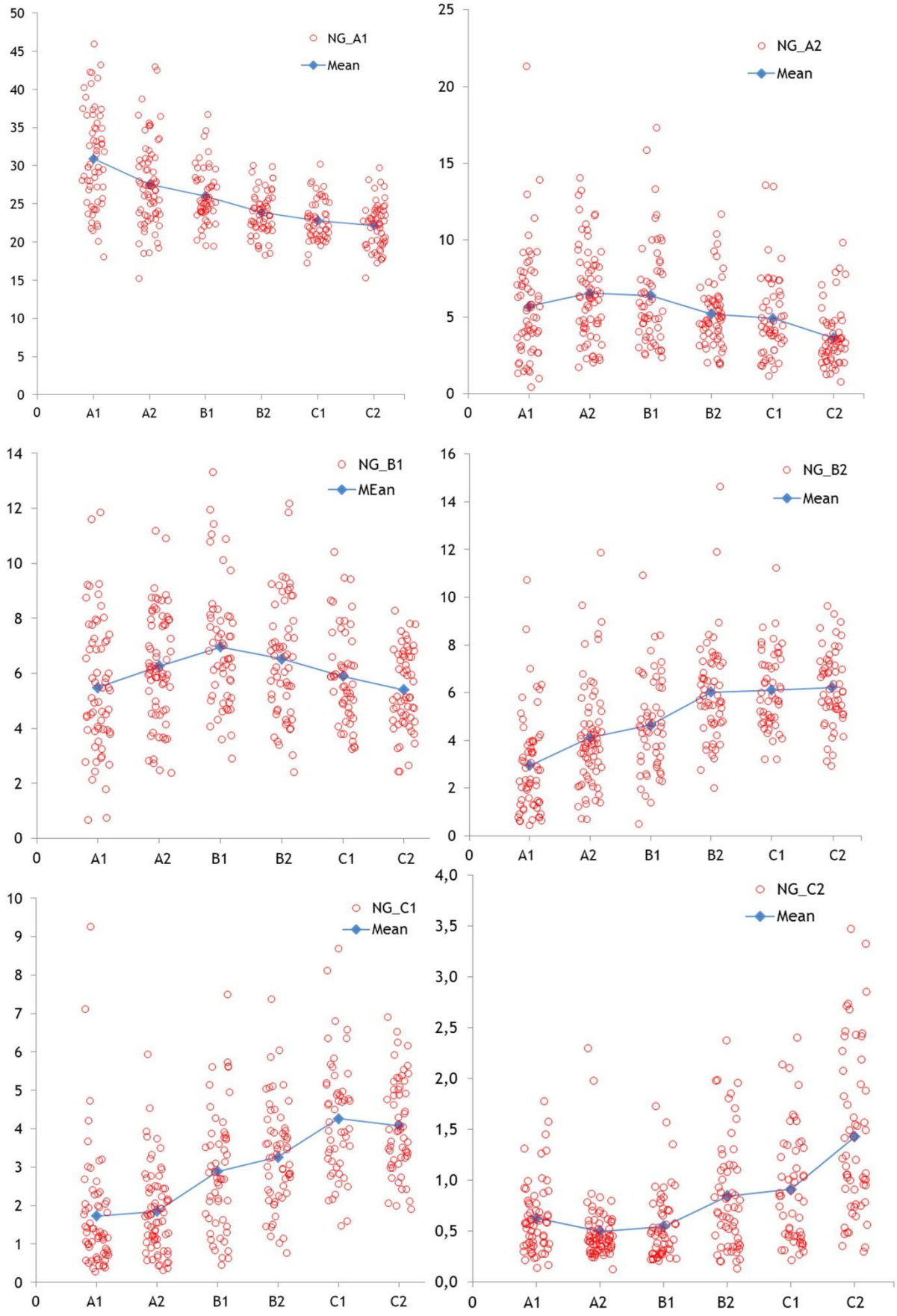
Πίνακας 9. Γραπτά κειμενικά είδη, επίπεδα Α1-Α2.

ΓΕΝΙΚΕΣ ΕΝΝΟΙΕΣ	ΕΙΔΙΚΕΣ ΕΝΝΟΙΕΣ
• Υπαρξιακές • Ποσοτικές • Του Χώρου • Χρονικές • Ποιοτικές • Αξιολογικές • Νοητικές	• Άτομο: Φυσική υπόσταση • Άτομο: Ψυχική υπόσταση και αντίληψη • Προσωπική ταυτότητα • Προσωπικές σχέσεις • Διατροφή • Εκπαίδευση • Εργασία • Ελεύθερος χρόνος • Πληροφόρηση και μέσα ενημέρωσης • Κατοικία • Υπηρεσίες • Αγορές και καταστήματα • Υγεία και υγιεινή • Ταξίδια, κατάλυματα και συγκοινωνίες • Οικονομία και παραγωγή • Επιστήμη και τεχνολογία • Κυβέρνηση, πολιτική και κοινωνία • Καλλιτεχνικές δραστηριότητες • Θρησκεία και φιλοσοφία • Γεωγραφία και φύση

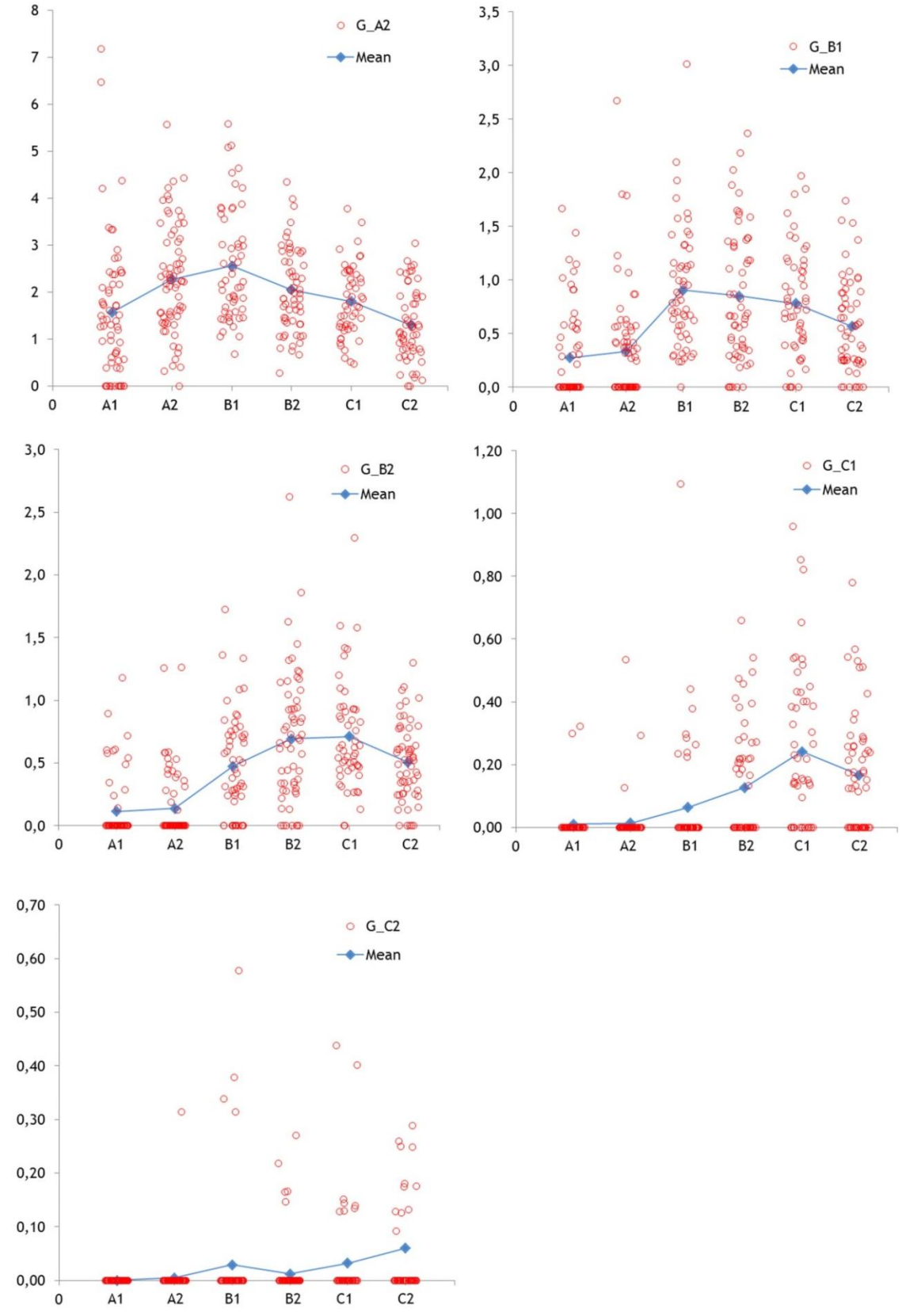
Πίνακας 10. Οργάνωση των καταλόγων των γενικών και ειδικών εννοιών του Προγράμματος Σπουδών του Ινστιτούτου Θερβάντες.



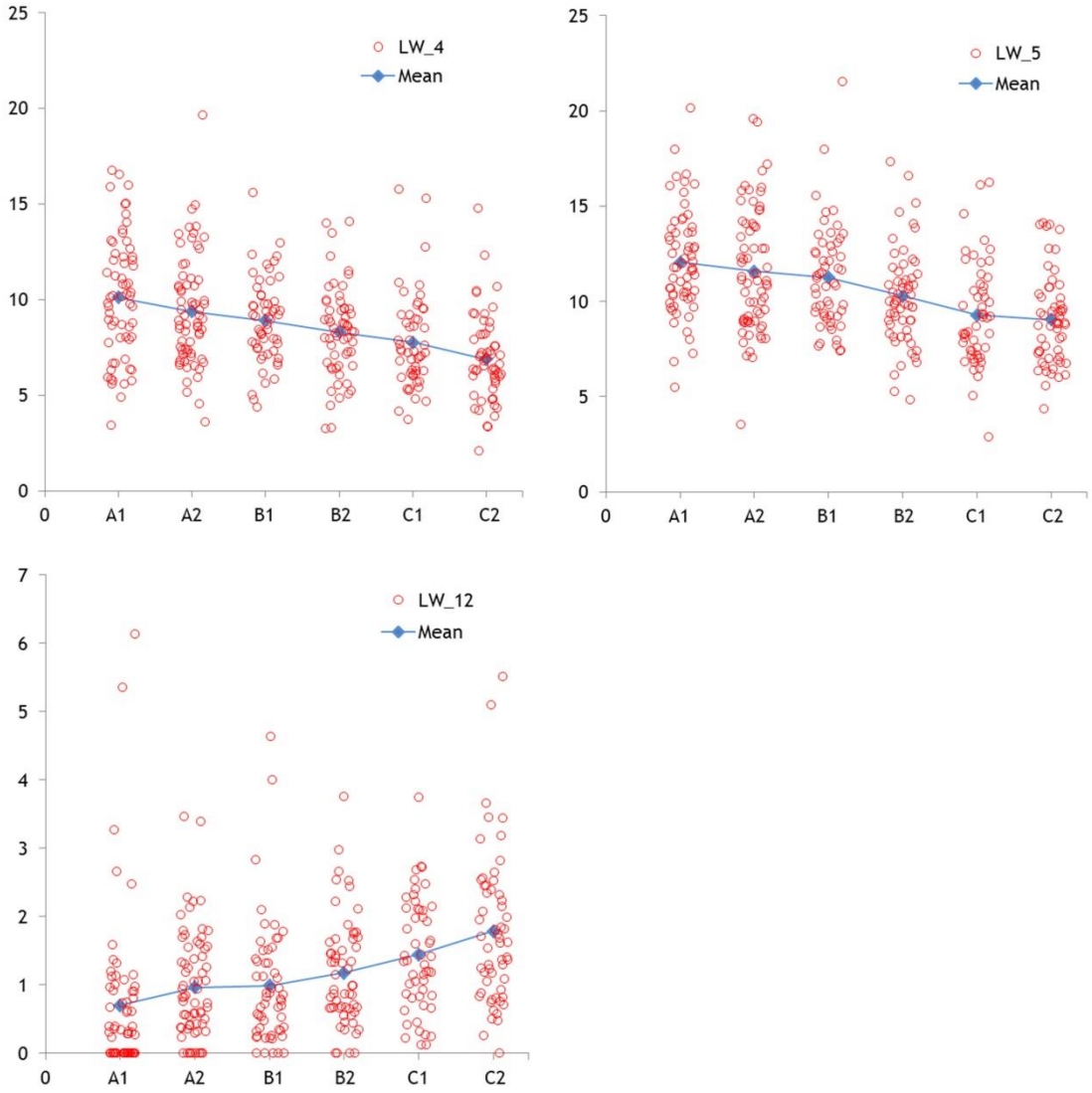
Εικόνα 8. Μετρήσεις και μέσοι όροι (σε ποσοστά %) για τις ειδικές έννοιες.



Εικόνα 9. Μετρήσεις και μέσοι όροι (σε ποσοστά %) για τις γενικές έννοιες.



Εικόνα 10. Μετρήσεις και μέσοι όροι (εμφανίσεις ανά 100 λέξεις) για τις γραμματικές δομές.



Εικόνα 11. Τιμές και μέσοι όροι (σε ποσοστά %) για τα χαρακτηριστικά LW_4, LW_5 και LW_12.

	Υφομετρικά Χαρακτηριστικά	
1	Type/token ratio	TTR
2	Μέσο μήκος λέξης	AWL
3	Τυπική απόκλιση του μέσου μήκους λέξης	WLsd
4	Μέσο μήκος πρότασης	ASL
5	Τυπική απόκλιση του μέσου μήκους πρότασης	SLsd
6	Άπαξ λεγόμενα	HapL
7	Δις λεγόμενα	DisL
8	Λόγος Δις προς Άπαξ Λεγόμενα	Dis_HapL
9	Λεξιλογική Πυκνότητα	LD
10	Yule's K	Yule
11	Εντροπία	Entr
12	Σχετική εντροπία	RelEntr
13-26	Φάσμα συχνότητας μήκους λέξεων	LW1…LW14
	Λεξιλογικά Χαρακτηριστικά	
27-31	Λέξεις που ανήκουν στις πρώτες 5 χιλιάδες των πιο συχνών λέξεων της ισπανικής	K1…K5
32-37	Γενικές Έννοιες Α1-Γ2	NG_A1, NG_A2, NG_B1, NG_B2, NG_C1, NG_C2
38-42	Ειδικές έννοιες Α1-Γ2	NE_A1, NE_A2, NE_B1, NE_B2, NE_C1, NE_C2
	Γραμματικά Χαρακτηριστικά	
42-46	Γραμματικές δομές Α2-Γ2	G_A2, G_B1, G_B2, G_C1, G_C2
47	Ενεστώτας - Οριστική	Prind
48	Ενεστώτας - Υποτακτική	Prsub
49	Παρατατικός - Οριστική	Plind
50	Παρατατικός - Υποτακτική	Plsub
51	Παρακείμενος - Οριστική	PPind
52	Παρακείμενος - Υποτακτική	PPsub
53	Υπερσυντέλικος - Οριστική	Plind
54	Υπερσυντέλικος - Υποτακτική	Plsub
55	Μέλλοντας - Οριστική	Futind
56	Μέλλοντας - Υποτακτική	Futsub
57	Πρότερος Αόριστος - Οριστική	Paind
58	Συντελεσμένος Μέλλοντας - Οριστική	FPind
59	Απλός Δυνητικός - Οριστική	CS
60	Σύνθετος Δυνητικός - Οριστική	CC
61	Αόριστος - Οριστική	Indef
62	Σύνθετο Απαρέμφατο	Cinf
63	Σύνθετο Γερούνδιο	Cger
64	Ρηματική έκφραση με το ρήμα Estar + Γερούνδιο	EstGER
65	Προστακτική	Imp

66	Απαρέμφατο	Inf
67	Γερούνδιο	Ger
68	Παθητική Μετοχή	Par
	Μέρη του λόγου	
69	Αντωνυμίες	Pron
70	Ρήματα	Verb
71	Επιρρήματα	Adv
72	Επίθετα	Adj
73	Προσδιοριστές	Det
74	Ουσιαστικα	Nouns
75	Ονοματικές Οντότητες	NE
76	Συμπλεκτικοί Σύνδεσμοι	ConC
77	Υποτελείς Σύνδεσμοι	ConS
78	Επιφωνήματα	Inter
79	Προθέσεις	Prep
80	Προσωπικές αντωνυμίες ανά ρήμα	PPpV
	Συντακτικά Χαρακτηριστικά	
81	Κειμενικοί δείκτες	Mark

Πίνακας 11. Σύνολο χαρακτηριστικών που μετρήθηκαν.