ΑΥΤΟΜΑΤΗ ΕΞΑΓΩΓΗ ΓΛΩΣΣΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΣΤΗ ΝΟΣΟ ΑΛΤΣΧΆΙΜΕΡ: ΜΙΑ ΥΠΟΛΟΓΙΣΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

Vassiliki Rentoumi, George Paliouras
Inst. of Informatics and Telecommunications,
NCSR "Demokritos", Athens, Greece

Dimitra Arfani, Katerina Fragkopoulou,
Spyridoula Varlokosta,
Department of Linguistics, National and Kapodistrian University of Athens (UoA), Greece

Abstract
In the present study, we analyzed written samples obtained from Greek native speakers diagnosed with Alzheimer's in mild and moderate stages and from age-matched cognitively normal controls (NC). We adopted a computational approach for the comparison of morpho-syntactic complexity and lexical variety in the samples. We used text classification approaches to assign the samples to one of the two groups. The classifiers were tested using various morphosyntactic and lexical features. The proposed method excels in discerning AD patients in mild and moderate stages from NC leading to the in depth understanding of language deficits in this neurodegenerative disease.

1. Εισαγωγή

Οι νευροεκφυλιστικές ασθένειες, όπως η Νόσος Αλτσχάιμερ (εφεξής ΝΑ) συνδέονται με αλλαγές στον προφορικό και γραπτό λόγο, οι οποίες δεν έχουν μελετηθεί εκτενώς. Οι αλλαγές αυτές στις δύο τροπικότητες συνδέονται με προβλήματα μνήμης αλλά και γλωσσικής επεξεργασίας. Τα προβλήματα γλωσσικής επεξεργασίας είναι εμφανή από τα πρώτα στάδια της νόσου (Kempler et al. 1987, Martin & Fedio 1983). Τα γλωσσικά προβλήματα έχουν εντοπιστεί, κυρίως, στη λεξική και σημασιολογική γνώση των ατόμων με ΝΑ και ειδικότερα στην κατονομασία ρημάτων και ουσιαστικών, στην εύρεση και ανάκληση λέξεων, καθώς και στη σημασιολογική ευχέρεια, δηλαδή στην κατονομασία λέξεων που ανήκουν στην ίδια σημασιολογική κατηγορία (π.χ. ζώα, φρούτα) (Altmann et al. 2001, Chertkow et al. 1989). Έχει υποστηριχθεί ότι η *ανομία*, δηλαδή η δυσκολία εύρεσης/ανάκλησης των κατάλληλων λέξεων, αποτελεί το πιο συχνό χαρακτηριστικό της νόσου, ακόμα και σε πρώιμο στάδιο (Altmann et al. 2001). Οι ασθενείς, συχνά, υποκαθιστούν τη λέξη-στόχο με αντωνυμία ή χρησιμοποιούν σημασιολογικά σχετικές λέξεις (σημασιολογικές παραφασίες) (Tang-Wai & Graham 2008) ή ακόμα και περιφράσεις. Η δυσκολία στην εύρεση των κατάλληλων λέξεων και τα προβλήματα κατονομασίας που παρατηρούνται στα άτομα με ΝΑ έχουν ως αποτέλεσμα έναν κενό περιεχομένου λόγου χωρίς συνοχή (Ripich and Terell 1988, Kavé & Levy 2003). Συχνές είναι οι επαναλήψεις και τα μεταγλωσσικά σχόλια που δυσχεραίνουν τη συνοχή του κειμένου.

Σχετικά με το έλλειμμα στη μορφοσυντακτική γνώση, τα πορίσματα των ερευνών είναι αντιφατικά με με άλλες έρευνες να υποστηρίζουν ότι η μορφοσύνταξη είναι διατηρημένη σε σχέση με τη λεξική και σημασιολογική ικανότητα και άλλες να επισημαίνουν ορισμένες διαταραχές στη μορφοσύνταξη ακόμα και σε πρώιμο στάδιο της νόσου (π.χ. δυσκολίες στην παραγωγή λέξεων κλειστής τάξης, στη συμφωνία αριθμού και συμφωνία υποκειμένου-ρήματος).

Εξίσου αντιφατικά είναι και τα αποτελέσματα για τη συντακτική ικανότητα, με κάποιες μελέτες να υποστηρίζουν ότι τα άτομα με ΝΑ μπορούν να ερμηνεύσουν συντακτικά

πολύπλοκες δομές, παρά την ελλειμματική μνήμη εργασίας, και άλλες να εντοπίζουν προβλήματα συντακτικής κατανόησης, κυρίως, στην κατανόηση σύνθετων δομών.

Παρόλο που οι γλωσσικές ικανότητες στη ΝΑ έχουν μελετηθεί σε κάποιο βαθμό, υπάρχουν αρκετοί περιορισμοί στις έρευνες που έχουν γίνει για τη μελέτη της γλώσσας. Ένας βασικός περιορισμός είναι ότι η γλωσσική ανάλυση των δεδομένων γίνεται χειροκίνητα, γεγονός που αποτελεί μια χρονοβόρα και πολλές φορές υποκειμενική διαδικασία.

Σε αυτή τη μελέτη επιχειρείται μια αυτόματη, υπολογιστική, γλωσσική ανάλυση, με τη χρήση της μηχανικής μάθησης, σε δείγματα γραπτού λόγου φυσικών ομιλητών της Ελληνικής που βρίσκονται σε ήπιο ή και μεσαίο στάδιο άνοιας αλλά και υγιών ηλικιωμένων αντιστοιχισμένων ως προς την ηλικία και την εκπαίδευση με την πειραματική ομάδα. Με την εφαρμογή ποσοτικών μεθόδων ανάλυσης διερευνώνται οι διαφορές στα γλωσσικά χαρακτηριστικά των ατόμων με άνοια και των υγιών ηλικιωμένων. τόχος είναι η εύρεση των σημαντικότερων διαχωριστικών κριτηρίων και γλωσσικών δεικτών για τις δύο ομάδες. Με τον εντοπισμό αυτών των διακριτών, γλωσσικών χαρακτηριστικών, αλλά και κάποιων γλωσσικών δομών που ίσως αποκλίνουν από την νόρμα των υγιών, πιθανόν να μπορέσουμε να βοηθήσουμε στην πρώιμη διάγνωση της ΝΑ και των άλλων μορφών άνοιας, διευκολύνοντας, με αυτόν τον τρόπο, την κλινική διαδικασία των ιατρών.

1.1. Υπολογιστικές μέθοδοι στην ανάλυση λόγου ατόμων με Νόσο Αλτσχάιμερ (ΝΑ)

Την τελευταία πενταετία αρκετοί ερευνητές παγκοσμίως έχουν στραφεί στη μελέτη του αυθόρμητου ή συνεχούς λόγου για την ανεύρεση γλωσσικών χαρακτηριστικών που μπορούν να διακρίνουν τον λόγο ατόμων με ΝΑ, ήπια γνωστική διαταραχή ή άλλους τύπους άνοιας, από τον λόγο υγιών ομιλητών. Οι περισσότερες από αυτές τις μελέτες εφαρμόζουν υπολογιστικές μεθόδους εξαγωγής γλωσσικών χαρακτηριστικών που μπορούν να διαφοροποιήσουν τους ασθενείς από τους υγιείς ομιλητές, με στόχο την πρώιμη διάγνωση νευρολογικών διαταραχών μέσω της ανάλυσης λόγου.

Πιο συγκεκριμένα, οι de Lira et al. (2011) μελετώντας τον αφηγηματικό λόγο στη ΝΑ, έδειξαν ότι οι ασθενείς είχαν περισσότερα λεξικά λάθη σε σχέση με την ομάδα ελέγχου. Στα λεξικά λάθη υπολογίστηκαν η δυσκολία στην εύρεση λέξεων, οι επαναλήψεις, οι φωνημικές παραφασίες και οι σημασιολογικές υποκαταστάσεις. Επιπλέον, εξετάστηκε ο ρόλος της συντακτικής πολυπλοκότητας και παρατηρήθηκε ότι οι ασθενείς χρησιμοποιούν λιγότερες παρατακτικές και ελλειπτικές δομές. Η μειωμένη χρήση των ελλειπτικών δομών ήταν το χαρακτηριστικό που διαφοροποίησε τις δύο ομάδες μεταξύ τους. Ομοίως, οι Orimaye et al. (2017) παρατήρησαν ότι οι ασθενείς με ΝΑ είχαν δυσκολία στην παραγωγή συντακτικά πολύπλοκων δομών και ελλειπτικών προτάσεων. Επιπρόσθετα, σημειώθηκε σημαντική διαφορά ως προς την παραγωγή του αριθμού των κατηγορημάτων μεταξύ των ασθενών και της ομάδας ελέγχου. Ο αριθμός των κατηγορημάτων ήταν σημαντικά μικρότερος στους ασθενείς συγκριτικά με τους υγιείς. Τέλος, σε ό,τι αφορά τα λεξικά χαρακτηριστικά, παρουσιάστηκε σημαντική διαφορά στη χρήση επαναλήψεων, στην αντικατάσταση λέξεων και στην εμφάνιση μη ολοκληρωμένων λέξεων ανάμεσα στις δύο ομάδες. Η ομάδα των ασθενών χρησιμοποίησε πολύ περισσότερες επαναλήψεις, αντικαταστάσεις λέξεων και ανολοκλήρωτες λεξικές επιλογές.

Οι Roark et al. (2011) εφάρμοσαν υπολογιστικές μεθόδους ανάλυσης του προφορικού λόγου για να διακρίνουν άτομα με *Ήπιο Γνωστικό Έλλειμμα* (Mild Cognitive Impairment) από υγιείς ομιλητές. Συγκεκριμένα, μέτρησαν χαρακτηριστικά γλωσσικής πολυπλοκότητας, όπως λέξεις ανά φράση και πυκνότητα περιεχομένου, αλλά και προσωδιακά χαρακτηριστικά, όπως συχνότητα, μήκος παύσης, συνολικό χρόνο παύσεων και φώνησης. Συμπέραναν ότι ένας συνδυασμός υπολογιστικών μεθόδων μπορεί να διακρίνει τις δύο ομάδες με βάση τις μετρήσεις της γλωσσικής πολυπλοκότητας τους.

Οι Garrard et al. (2014) χρησιμοποίησαν μεθόδους μηχανικής μάθησης (όπως τους ταξινομητές Naive Bayes Gaussian (NBG) και Naive Bayes Multinomial (NBM)) για να διακρίνουν τα δείγματα λόγου ατόμων με σημασιολογική άνοια από τον υγιή πληθυσμό, βασιζόμενοι σε λεξικά χαρακτηριστικά απομαγνητοφωνημένων αφηγήσεων. Εντόπισαν ότι στο λεξιλόγιο των ατόμων με σημασιολογική άνοια κυριαρχούν γενικοί και δεικτικοί όροι, όπως «κάτι», «αυτό», καθώς και μεταφηγηματικά εκφωνήματα, όπως «γνωρίζω», «θυμάμαι». Αντιθέτως, στις περιγραφές των υγιών παρατηρούνται λέξεις χαμηλής συχνότητας με υψηλό πλούτο περιεχομένου, όπως τα ουσιαστικά «ακτή», «γρασίδι» και όχι υψηλής συχνότητας και μικρού σημασιολογικού φορτίου, όπως οι γενικοί όροι και οι αόριστες/δεικτικές αντωνυμίες.

Οι Rentoumi et al. (2014) χρησιμοποίησαν υπολογιστικές μεθόδους σε τι είδος κειμένου; για την αξιολόγηση ορισμένων λεξικών ποιοτικών και ποσοτικών χαρακτηριστικών (είδος και συχνότητα λέξεων) αλλά και της συντακτικής πολυπλοκότητας για να διακρίνουν τον λόγο ασθενών με μεικτή αγγειακή άνοια από τον λόγο ασθενών με καθαρή άνοια. Συμπέραναν ότι η ομάδα των μεικτών ανοϊκών παρουσιάζειι μειωμένη λεξική ποικιλία και συντακτική πολυπλοκότητα στον προφορικό λόγο σε σχέση με την ομάδα των καθαρά ανοϊκών.

Οι Fraser et al. (2016) εφάρμοσαν μια προσέγγιση μηχανικής μάθησης για να μελετήσουν γλωσσικά χαρακτηριστικά στη NA, όπως σημασιολογικές υποκαταστάσεις, συντακτική πολυπλοκότητα, μήκος ονοματικών, ρηματικών και επιθετικών φράσεων, ποσοστό μερών του λόγου, λεξικό πλούτο, πληροφοριακό περιεχόμενο, επαναλήψεις, ακουστικά/φωνολογικά λάθη, χαρακτηριστικά που μπορούν να εκμαιευτούν αυτόματα από ψηφιακά δείγματα συνεχούς λόγου. Τα αποτελέσματα έδειξαν ένα σημασιολογικό έλλειμμα με αυξημένη χρήση επαναλήψεων, αντωνυμιών και περιορισμένη λεξική ποικιλία. Επιπλέον, εντοπίστηκε συντακτική δυσκολία ως προς την παραγωγή βοηθητικών ρημάτων, γερουνδίων και μετοχών, ενώ στις περιγραφές των ασθενών παρουσιάστηκε χαμηλό πληροφοριακό περιεχόμενο.

Τέλος, οι Kavé & Dassa (2018) . χρησιμοποίησαν αυτόματα εργαλεία ανάλυσης κειμένου και ανέλυσαν 10 λεξικά και γραμματικά χαρακτηριστικά: τον συνολικό αριθμό λέξεων, το ποσοστό των λέξεων περιεχομένου σε σχέση με το συνολικό αριθμό λέξεων, τον λόγο των αντωνυμιών, τον λόγο άπαξ λεγομένων και λεξικών τύπων (type-token ratio), τον μέσο όρο στη συχνότητα λέξεων, το ποσοστό των ενεστωτικών ρημάτων, καθώς και των πιο συχνών ρηματικών τύπων, των προθέσεων αλλά και τους δείκτες δευτερεύουσας πρότασης (subordination markers) σε σχέση με το συνολικό αριθμό των λέξεων. Επιπλέον, αναλύθηκαν οι ενότητες πληροφοριακού περιεχομένου (information units), για παράδειγμα, οι ενέργειες στο δείγμα κειμένου που προέκυψε από την περιγραφή της εικόνας . Βρέθηκαν διαφορές μεταξύ των δύο ομάδων σε σχέση με τον συνολικό αριθμό των λέξεων που παρήχθησαν. Συγκεκριμένα, οι ασθενείς με NA παρήγαγαν πολύ περισσότερες λέξεις σε σύγκριση με τουςυγιείς, αλλά χωρίς πληροφοριακό περιεχόμενο Επίσης, σημειώθηκε υπερβολική χρήση αντωνυμιών σε σχέση με τα ουσιαστικά και οι ασθενείς εμφάνισαν μικρότερο λόγο άπαξ λεγομένων και λεξικών τύπων (type-token ratio), χρησιμοποιώντας τις πιο συχνές λέξεις.

2. Μεθοδολογία

2.1 Συμμετέχοντες

Στη μελέτη συμμετείχαν 30 ασθενείς με NA, ηλικίας 60-85 ετών σε ήπιο ή μεσαίο στάδιο της νόσου (MMSE: 10-25/30). Οι ασθενείς αξιολογήθηκαν με βάση τα διαγνωστικά κριτήρια των McKhann et al. (1984, 2011). Ο εντοπισμός και η διάγνωση του γνωστικού ελλείμματος πραγματοποιήθηκε μέσω της λήψης ιστορικού από τον ασθενή και από κάποιο άτομο του περιβάλλοντός του και μέσω αξιολόγησης, η οποία συμπεριελάμβανε αξονική τομογραφία και χρήση του Mini-Mental State Examination (MMSE) (Folstein et al.1975· Φουντουλάκης

κ.ά. 1994) για τον προσδιορισμό του σταδίου της άνοιας. Ως ομάδα ελέγχου χρησιμοποιήθηκαν 30 υγιείς ενήλικες, με την ίδια ηλικία και τα ίδια έτη εκπαίδευσης με τους ασθενείς (βλ. πίνακα 1). Δεν υπήρχε στατιστικά σημαντική διαφορά ανάμεσα στις δύο ομάδες όσον αφορά την ηλικία, την εκπαίδευση, το φύλο, ενώ σημειώθηκε σημαντική διαφορά στο MMSE ($p<0.05$). Η σύγκριση των δύο ομάδων έγινε με t-test για ανεξάρτητα δείγματα και chi-square για τους λόγους του γένους.

Πίνακας 1. Δημογραφικά Στοιχεία

Δημογραφικά	NA (n=30)	Ομάδα ελέγχου (n=30)	Στατιστική Σημαντικότητα
	Μ.Ο. (Μέσος Όρος)	Μ.Ο. (Μέσος Όρος)	* = p <0.05, μ.σ. = μη σημαντική
Ηλικία	66.48	68.03	μ.σ.
Εκπαίδευση	12	13.93	μ.σ.
Φύλο (άρρεν:θήλυ)	13:17	16:14	μ.σ.
MMSE	22.68	28.26	*

2.2 Πειραματική διαδικασία

Από τις δύο πειραματικές ομάδες συλλέχθηκαν 60 γραπτά δείγματα λόγου (30 από κάθε ομάδα) με τη χορήγηση της δοκιμασίας της Κλοπής του Μπισκότου (Cookie-Theft Picture Description Task, Goodglass & Kaplan, 1983). Κάθε υποκείμενο έβλεπε την Εικόνα 1 και έπρεπε να κάνει μια γραπτή περιγραφή. Η οδηγία ήταν η εξής: *Θέλω να γράψετε όσα βλέπετε στην εικόνα όσο το δυνατόν πιο ολοκληρωμένα.* Ο ερευνητής προσπαθούσε να εκμαιεύσει ένα αντιπροσωπευτικό δείγμα λόγου. Απ' όσο γνωρίζουμε είναι η πρώτη φορά που αναπτύσσονται υπολογιστικές μέθοδοι στην Ελλάδα για την ανίχνευση πρώιμων ενδείξεων της άνοιας μέσω της ανάλυσης του ελληνικού γραπτού λόγου.

Εικόνα 1. Η δοκιμασία της κλοπής του μπισκότου (Goodglass & Kaplan 1983)

Τα δείγματα του γραπτού λόγου συλλέχθηκαν με σκοπό α) να αναλυθούν υπολογιστικά με εργαλεία μηχανικής μάθησης και β) να ελέγξουμε κατά πόσο οι ασθενείς με ΝΑ έχουν μορφοσυντακτικά και λεξικά ελλείμματα στον γραπτό λόγο τους. Στη μηχανική μάθηση χρησιμοποιούνται αλγόριθμοι οι οποίοι τροφοδοτούνται από γλωσσικά δεδομένα και "μαθαίνουν" τη γλώσσα. Στην παρούσα έρευνα οι αλγόριθμοι ρυθμίστηκαν έτσι ώστε να εντοπίζουν τα χαρακτηριστικά της συντακτικής πολυπλοκότητας και της λεξικής ποικιλίας μέσα στα δείγματα λόγου. Απώτερος στόχος είναι η δημιουργία ενός αλγορίθμου, ο οποίος να κατηγοριοποιεί τα δεδομένα στην αντίστοιχη σωστή ομάδα, δηλαδή στην ομάδα των ασθενών με ΝΑ ή στην ομάδα ελέγχου (ΟΕ).
Ειδικότερα, η μεθοδολογία που ακολουθήθηκε αποτελούνταν από δύο συνακόλουθα στάδια: (α) εξαγωγή χαρακτηριστικών και (β) κατηγοριοποίηση με μηχανική μάθησηΤο εξαγόμενο του πρώτου σταδίου αποτελούσε το εισαγόμενο του δεύτερου βήματος..
2.2.1 Εξαγωγή χαρακτηριστικών

Στόχος της ανάλυσης ήταν να κατηγοριοποιήσουμε κάθε γραπτό δείγμα λόγου στην κατάλληλη ομάδα (ΝΑ και ΟΕ). Το πρώτο στάδιο αφορούσε την αυτόματη εξαγωγή χαρακτηριστικών (features extraction). Μία λεπτομερής περιγραφή των εξαγόμενων χαρακτηριστικών παρουσιάζεται στον Πίνακα 2. . Για την εξαγωγή χαρακτηριστικών χρησιμοποιήθηκαν ένας Part of Speech (POS) tagger, δηλαδή ένα υπολογιστικό πρόγραμμα που βάζει ετικέτες στα μέρη του λόγου και ένας NP chunker για τα ελληνικά. Οι PoS και NP chunker υλοποιήθηκαν στο πλαίσιο της υπηρεσίας ellogon και αποτελούν τμήματα αυτής (Petasis et al. 2012). Η υπηρεσία ellogon λειτουργεί ως εργαλείο για την επεξεργασία φυσικής γλώσσας. Στο αναλυόμενο εξαγόμενο των POS και NP chunker εφαρμόσαμε τον ανιχνευτή Alzheimer (Alzheimer's detector), ο οποίος σχεδιάστηκε για να εξαγάγει τιμές που αντιστοιχούν σε συγκεκριμένο εύρος χαρακτηριστικών λεξικής ποικιλίας και συντακτικής πολυπλοκότητας σε αναλυμένα ελληνικά κείμενα.

Τα χαρακτηριστικά αφορούν μετρήσεις λεξικής ποικιλίας και συντακτικής πολυπλοκότητας. Οι μετρήσεις λεξικής ποικιλίας είναι συνολικά 9 και μέσω αυτών μπορεί να ποσοτικοποιηθεί ο λεξικός πλούτος στα δείγματα λόγου ασθενών με ΝΑ και υγιών αντίστοιχων.. Συγκεκριμένα, οι μετρήσεις λεξικής ποικιλίας αφορούν 1) ένα δείκτη ποικιλίας (LV) στις λέξεις περιεχομένου, δηλαδή τον λόγο μοναδικών εμφανίσεων λέξεων προς το σύνολο των λέξεων, 2) τον διλογαριθμισμένο λόγο (Bi Logarithmic) των μοναδικών εμφανίσεων ενός λεξικού τύπου ως προς το σύνολο των λέξεων του κειμένου (Log TTR), 3) την ποικιλία στα ουσιαστικά (NV), 4) στα επίθετα (ADJV), 5) στους προσδιοριστές (MODV), 6) στα επιρρήματα (ADV), 7) στα διορθωμένα ρήματα (CVV), και 8) την λεξική μετρική Brunet, που μετράει το ποσοστό των λεξικών τύπων και του λεξιλογίου. Η ποικιλία στα ουσιαστικά, επίθετα, επιρρήματα, στα διορθωμένα ρήματα, καθώς και στους προσδιοριστές μετρήθηκε με τον λόγο των λέξεων της κάθε κατηγορίας προς το σύνολο των λέξεων περιεχομένου σε κάθε απόσπασμα λόγου. Για παράδειγμα, για την ποικιλία των ουσιαστικών μετρήθηκε ο λόγος των ουσιαστικών τύπων ως προς το σύνολο των λέξεων περιεχομένου.

Χρησιμοποιώντας όλες τις παραπάνω μετρικές λεξικής ποικιλίας σκοπός μας ήταν αφενός να μετρήσουμε το εύρος λεξιλογίου στην ομάδα ασθενών και στην ομάδα ελέγχου, όπως αυτό εμφανίζεται στα αποσπάσματα λόγου τους, αφετέρου να εντοπίσουμε το βαθμό επανάληψης σε επίπεδο λέξης/μερών του λόγου καθώς η επανάληψη λέξεων και συγκεκριμένων μερών του λόγου συνιστά κοινό χαρακτηριστικό στη ΝΑ (Tomoeda et al., 1996).

Οι συντακτικές μετρήσεις, που ποσοτικοποιούν τη συντακτική πολυπλοκότητα στα γραπτά δείγματα λόγου και για τις δύο ομάδες μας, αφορούν τον μέσο όρο του μήκους των εκφωνημάτων/προτάσεων, δηλαδή τον λόγο των λέξεων προς τον αριθμό των

εκφωνημάτων/προτάσεων (number of words/number of sentences), καθώς και τον μέσο όρο των ονοματικών φράσεων, δηλαδή τον λόγο των ονοματικών φράσεων ως προς το συνολικό αριθμό των προτάσεων σε ένα δείγμα λόγου (number of noun phrases/number of all sentences of each text).

Πίνακας 2. Σύνολο Εξαγόμενων χαρακτηριστικών

Χαρακτηριστικά Λεξικής Ποικιλίας	1. Λεξική ποικιλία (LV)	Αριθμός μοναδικών λεξικών τύπων/ σύνολο λέξεων κειμένου
	2. Διλογαριθμικός λόγος (Log TTR)	Λογάριθμος μοναδικών λεξικών τύπων/σύνολο λέξεων κειμένου
	3. Ποικιλία ουσιαστικών (NV)	Αριθμός των ουσιαστικών/ σύνολο λέξεων περιεχομένου
	4. Ποικιλία επιθέτων (ADJV)	Αριθμός επιθέτων/σύνολο λέξεων περιεχομένου
	5. Ποικιλία προσδιοριστών (MODV)	Αριθμός επιθέτων και επιρρημάτων/σύνολο λέξεων περιεχομένου
	6. Ποικιλία επιρρημάτων (ADVV)	Αριθμός επιρρημάτων/ σύνολο λέξεων περιεχομένου
	7. Ποικιλία διορθωμένων ρημάτων (CVV)	Αριθμός ρηματικών τύπων/αριθμός ρημάτων x 2
	8. Ποικιλία ρημάτων (VV)	Αριθμός ρηματικών τύπων/ σύνολο λέξεων περιεχομένου
	9. Brunet (W)	$N^{v-0.165}$ N= αριθμός λέξεων, V= λεξιλόγιο
Χαρακτηριστικά Συντακτικής πολυπλοκότητας	10. Μέσο Μήκος Πρότασης (MLS)	Αριθμός λέξεων/Αριθμός Προτάσεων

	11. Μέσος αριθμός ονοματικών φράσεων (MNP.)	Αριθμός ΟΦ/Αριθμός όλων των προτάσεων κάθε κειμένου

Οι παραπάνω μετρήσεις λεξικής ποικιλίας και συντακτικής πολυπλοκότητας συνδέονται με την παρουσία ή μη άνοιας και εντοπίζονται επηρεασμένες στα πρώιμα στάδια της άνοιας αλλά και στην εξέλιξη της νόσου.

Οι μετρήσεις αυτές δεν σχετίζονταν μόνο με λεξικά και συντακτικά χαρακτηριστικά, αλλά, επιπλέον, με κειμενικά και πιο σύνθετα υπολογιστικά χαρακτηριστικά της Θεωρίας της Πληροφορίας (Information Theory), τα οποία εκμαιεύτηκαναπό κάθε ένα σετ γλωσσικών κειμένων ξεχωριστά.

2.2.2 Κατηγοριοποίηση με μηχανική μάθηση

Στο δεύτερο στάδιο εφαρμόσαμε μια προσέγγιση κατηγοριοποίησης των χαρακτηριστικών μέσω της μηχανικής μάθησης για να προβλέψουμε την κατηγορία (ΝΑ ή ΟΕ) στην οποία ανήκει κάθε γραπτό δείγμα. Οι ταξινομητές/κατηγοριοποιητές μηχανικής μάθησης που χρησιμοποιήθηκαν είναι οι Naive Bayes και SMO στο πλαίσιο του περιβάλλοντος WEKA (Waikato Environment for Knowledge Analysis, Hall et al 2009).

Η διαδικασία της αξιολόγησης και ταξινόμησης περιγράφεται στην επόμενη υποενότητα.

2.2.3 Διαδικασία Αξιολόγησης

Για να αξιολογήσουμε την αποτελεσματικότητα και την ακρίβεια του συστήματος (αλγορίθμων και ταξινομητών), υιοθετήσαμε μια αξιολογική προσέγγιση 10 διασταυρώσεων. Η δομή των δεδομένων χωρίστηκε σε 10 υποσύνολα: τα 9 χρησιμοποιήθηκαν για εκπαίδευση (training) και το άλλο για έλεγχο (testing) του αλγορίθμου ταξινόμησης. Μετά την εκπαίδευση του ταξινομητή (classifier), ακολούθησε ο έλεγχος στο πειραματικό σύνολο (test set). Η ακρίβεια του ταξινομητή υπολογίστηκε για κάθε πλαίσιο ξεχωριστά, και μετά υπολογίστηκε η μακρομεσική (macroaverage) ακρίβεια για να υπολογιστεί η αποτελεσματικότητα του κάθε ταξινομητή σε όλα τα πλαίσια συνολικά. Σε κάθε τεστ κατηγοριοποίησης, οι μεταγραφές που χρησιμοποιήθηκαν χωρίστηκαν τυχαία σε 10 ίδιου μεγέθους, τυχαίως επιλεγμένα υποσύνολα. Αυτή η διαδικασία επαναλήφθηκε 10 φορές, χρησιμοποιώντας διαφορετικώς κατασκευασμένα υποσύνολα για έλεγχο σε κάθε δείγμα λόγου.

2.2.4 Πειράματα Κατηγοριοποίησης

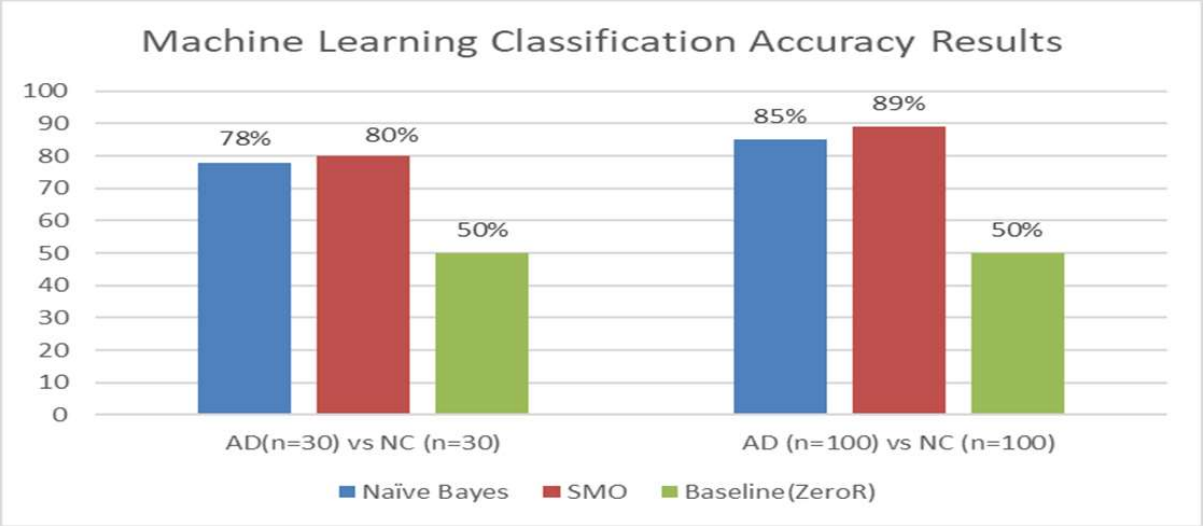
Εφαρμόστηκαν δύο πειράματα κατηγοριοποίησης. Στο πρώτο, χρησιμοποιήθηκαν 30 αληθινά δείγματα για κάθε ομάδα (ασθενείς με ΝΑ, ομάδα ελέγχου), ενώ στο δεύτερο πείραμα δημιουργήθηκε με τη βοήθεια του αλγορίθμου SMOTE (Synthetic Minority Oversampling Technique) ένα συνθετικό μεγαλύτερο δείγμα που αποτελούνταν από τα 30 αληθινά δείγματα συν τα 70 συνθετικώς παραγόμενα για κάθε ομάδα (ασθενείς με ΝΑ, ομάδα ελέγχου). Μ' αυτό τον τρόπο θέλαμε να ελέγξουμε την αποτελεσματικότητα της μεθόδου σ' ένα μεγαλύτερο δείγμα. Έτσι, δημιουργήσαμε ένα συνθετικό δείγμα 100 κειμένων για κάθε κατηγορία (30 πραγματικών και 70 συνθετικών). Η ακρίβεια των NB και SMO κατηγοριοποιητών/ταξινομητών διαφάνηκε στο συνθετικό δείγμα. Πιο συγκεκριμένα, όπως φαίνεται από τον πίνακα 3 παρακάτω, και στα δύο πειράματα κατηγοριοποίησης, οι NB και SMO ξεπέρασαν σημαντικά το επίπεδο αναφοράς (baseline condition) και για τις δύο συγκρίσεις Α και Β. Το επίπεδο αναφοράς τέθηκε με τη χρήση του ZeroR κατηγοριοποιητή

(http://chemeng.utoronto.ca/~datamining/dmc/zeror.htm), που τον παρέχει το WEKA, το οποίο προβλέπει την πλειονοτική κατηγορία.

3. Αποτελέσματα

Η σύγκριση Α αναφέρεται στο κανονικό δείγμα (αριστερό διάγραμμα στον πίνακα), ενώ η σύγκριση Β (δεξιό διάγραμμα στον πίνακα) αναφέρεται στο συνθετικό. Ενώ για τη σύγκριση Α, χρησιμοποιήσαμε το dataset των 60 γραπτών δειγμάτων, 30 για κάθε κατηγορία (NA, OE), στη σύγκριση Β χρησιμοποιήσαμε ένα συνθετικό dataset αποτελούμενο από 200 δείγματα, 100 για κάθε κατηγορία (AD, NC), για να εξετάσουμε την αποδοτικότητα της μεθόδου μας απέναντι σε ένα μεγαλύτερο dataset.

Πίνακας 3. Αποτελέσματα Ακρίβειας Κατηγοριοποίησης Μηχανικής Μάθησης



Όπως μπορεί να παρατηρηθεί από τα αποτελέσματα στον πίνακα 3, οι ταξινομητές ήταν πάρα πολύ ακριβείς: 78% ο NB, 80% ο SMO για το αληθινό δείγμα, 85% και 89 % για το συνθετικό αντίστοιχα. Η ακρίβεια των ταξινομητών προσμετράται με την απόδοση στον αριθμό των μεταγραφών σε κάθε σωστή τάξη. Χρησιμοποιήθηκαν είτε όλα τα χαρακτηριστικά είτε επιλογή μερικών για τη βελτιστοποίηση στην απόδοση των ταξινομητών. Αυτά τα αποτελέσματα υποδεικνύουν ότι η λεξική ποικιλία και η συντακτική πολυπλοκότητα αποτελούν πολύ καλούς διαφοροποιητικούς παράγοντες στη γλώσσα των ατόμων με Νόσο Αλτσχάιμερ και των υγιών ηλικιωμένων.

4. Συμπεράσματα

Από τα αποτελέσματά μας επιβεβαιώθηκε η αρχική μας υπόθεση ότι υπάρχουν γλωσσικά ελλείμματα αλλά και γνωστική έκπτωση που μπορούν να αποτυπωθούν στο γραπτό λόγο των ασθενών με ΝΑ. Πιο συγκεκριμένα, από τα αποτελέσματα που επιτεύχθηκαν μπορούμε να συμπεράνουμε ότι η γλώσσα στην ομάδα των ασθενών με ΝΑ είναι διακριτή από τη γλώσσα της ομάδας των υγιών. Επιπρόσθετα, η λεξική ποικιλία και η συντακτική πολυπλοκότητα είναι πολύ καλοί διαχωριστικοί παράγοντες όταν πρέπει να διαχωριστεί η γλώσσα των ασθενών με ΝΑ από τη γλώσσα των υγιών. Φαίνεται, δηλαδή ότι οι ασθενείς με ΝΑ έχουν

πρόβλημα με τις συντακτικές πολύπλοκες δομές και παρουσιάζουν μείωση στο λεξικό τους πλούτο, μεταξύ άλλων χαρακτηριστικών.

Υιοθετώντας μια προσέγγιση αξιολόγησης ανάμεσα σε 10 χαρακτηριστικά (10-fold), υψηλές ακρίβειες επιτεύχθηκαν και στις 2 συγκρίσεις σε όλα τα πειράματα κατηγοριοποίησης με τα 10 χαρακτηριστικά (10-fold). Υψηλή ακρίβεια μπορεί περαιτέρω να εντοπιστεί ως αποτέλεσμα μιας κατάλληλης σύνδεσης του κατηγοριοποιητή μηχανικής μάθησης με τα εφαρμόσιμα χαρακτηριστικά (συντακτική πολυπλοκότητα και λεξική ποικιλία). Αυτή η υψηλή ακρίβεια δείχνει ότι υπάρχουν συστηματικές διαφορές στη λεξική ποικιλία και στη συντακτική πολυπλοκότητα μεταξύ των ασθενών με ΝΑ και των υγιών ατόμων.

Στην παρούσα μελέτη, ένας συνδυασμός χαρακτηριστικών λεξικής ποικιλίας και συντακτικής πολυπλοκότητας ξεπέρασε μια προσέγγιση βάσης (baseline condition), εύρημα πολύ σημαντικό όσον αφορά στη χρήση υπολογιστικών μεθόδων για τη διάκριση ατόμων με ΝΑ από τους υγιείς. Επιπλέον, υπολογιστικά εργαλεία και μέθοδοι, όπως αυτά που περιγράφησαν στο παρόν άρθρο είναι φανερό πως αποτελούν πολύ χρήσιμα συμπληρωματικά εργαλεία για τη διαδικασία της κλινικής διάγνωσης της άνοιας, καθώς επίσης και της έγκαιρης αλλά και έγκυρης πρόληψης αυτής με τη δημιουργία και χρήση υπολογιστικών βιοδεικτών.

Αυτή είναι η πρώτη φορά που μέθοδοι μηχανικής μάθησης αξιολογούνται σε ελληνικά γραπτά δεδομένα από Έλληνες φυσικούς ομιλητές για τη διάγνωση του Αλτσχάιμερ. Αυτό το εγχείρημα στοχεύει στην αναγνώριση συγκεκριμένων ιδιοσυγκρασιακών γλωσσολογικών χαρακτηριστικών που είναι στενά συνδεδεμένα με το Αλτσχάιμερ. Επίσης, έχει ως στόχο να επιβεβαιώσει την παρουσία γλωσσολογικών ελλειμμάτων διαγλωσσικά εξαιτίας του Αλτσχάιμερ. Τα διαγλωσσικά ελλείμματα μπορούν να δείξουν και να ανοίξουν το δρόμο για τη δημιουργία ενός διαγλωσσικού γλωσσολογικού διαγνωστικού εργαλείου για το Αλτσχάιμερ.

Με αυτή τη μελέτη, ευελπιστούμε ότι παρέχουμε μια υπολογιστική βάση για την εκτίμηση της γνωστικής λειτουργίας και της επακόλουθης παρέμβασης στη ΝΑ με στόχο τη διατήρηση και την τελειοποίηση της ανθρώπινης νόησης μέσω ενός αποδοτικού διασυνδετικού ανθρώπινου υπολογιστή.

Βιβλιογραφικές αναφορές

Altmann, L.J.P., Kempler, D. and E.S., Andersen (2001). Speech errors in Alzheimer's disease: Reevaluating morphosyntactic preservation. *Journal of Speech, Language and Hearing Research 44*: 1069–1082.

Chertkow H., Bub, D. and M., Seidenberg (1989). Priming and semantic memory loss in Alzheimer's disease. *Brain and Language 36*: 420–446.

de Lira, J.O., Ortiz, K.Z., Campanha, A.C., Bertolucci, P.H.F. and T.S.C., Minett (2011). Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics* 23 (3): 404–412.

Fraser, K.C., Meltzer, J.A. and F., Rudzicz (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease 49* (2): 407-422.

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B. and M.L., Gorno-Tempini (2014). Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex 55*: 122-129.

Kavé, G., and A., Dassa (2018). Severity of Alzheimer's disease and language features in picture descriptions. *Aphasiology*, *32* (1), 27–40. https://doi.org/10.1080/02687038.2017.1303441

Kavé, G. and Y., Levy (2003). Sensitivity to gender, person and tense inflection by persons with Alzheimer's disease. *Brain and Language 87:* 267–277.

Kempler, D., Curtiss, S. and C., Jackson (1987). Syntactic preservation in Alzheimer's disease. *Journal of Speech and Hearing Research 30*: 343–350.

Martin, A and P., Fedio (1983). Word production and comprehension in Alzheimer's disease: the breakdown of semantic knowledge. *Brain and Language 19*: 124–141.

Orimaye, S., Kong, J.S-M, Golden, K.J., Wong, C.P. and I.N., Soyiri (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers, *BMC Bioinformatics 18 (1)*: 1-13.

Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C.A. and P., Garrard (2014). Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease 42*: S3-S17.

Ripich, D.N. and B.Y., Terell (1988). Patterns of discourse cohesion and coherence in Alzheimer's disease. *Journal of Speech and Hearing Disorders* 53 (1): 8-15.

Roark, B., Mitchell, M., Hossom, J.P., Hollingshead, K. and J. Kaye (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEE Transactions on audio, speech and language processing 19* (7): 2081-2090.

Tang-Wai, F.D. and L.N., Graham (2008). Assessment of Language Function in Dementia. *Geriatrics & Aging 11* (2): 103–110.