Προβλέποντας το φύλο του συγγραφέα: Μία υφομετρική ανάλυση σε κείμενα εφημερίδων

Γεώργιος Κ. Μικρός

1 Εισαγωγή[1]

Γυναίκες και άνδρες μιλάνε διαφορετικά; Η συγκεκριμένη ερώτηση φαίνεται ότι απασχολεί όλο και περισσότερο ένα ευρύ φάσμα επιστημών από τη νευροφυσιολογία και τη γνωσιακή επιστήμη έως την κοινωνιογλωσσολογία και την ανθρωπολογία. Η διαφυλική επικοινωνία εξελίσσεται τα τελευταία 20 χρόνια σε ένα εξαιρετικά ενδιαφέροντα διεπιστημονικό τομέα έρευνας με ευρήματα που αναμορφώνουν τις απόψεις που έχουμε για την αλληλεπίδραση του κοινωνικού γένους (gender) με τη γλωσσική χρήση.

Η παρούσα έρευνα έχει οργανωθεί σε τρία διαφορετικά αλλά απόλυτα συσχετισμένα στάδια μεταξύ τους τα οποία αναφέρονται επιγραμματικά παρακάτω:

1. Θεωρητικές προσεγγίσεις στις διαφορές των δύο φύλων σχετικά με τη γλωσσική ικανότητα και τη γλωσσική επιτέλεση.
2. Πειραματικός προσδιορισμός των υφομετρικών χαρακτηριστικών που διαφέρουν στατιστικά σημαντικά μεταξύ ανδρών και γυναικών συγγραφέων σε Ηλεκτρονικό Σώμα Κειμένων (ΗΣΚ) εφημερίδας
3. Ανάπτυξη Τεχνητού Νευρωνικού Δικτύου (ΤΝΔ) για την πρόβλεψη του φύλου του συγγραφέα βάσει των υφομετρικών χαρακτηριστικών του κειμένου.

Και τα τρία αυτά στάδια αποτελούν μια στενά εξαρτώμενη αλυσίδα παραδοχών που ορίζουν το θεωρητικό και μεθοδολογικό πλαίσιο της παρούσας έρευνας. Το πρώτο στάδιο εξετάζει τα βασικότερα ερευνητικά αποτελέσματα που έχουν προκύψει από την εξέταση των γλωσσικών διαφορών που παρουσιάζουν τα δύο φύλα στη γλωσσική τους χρήση, όπως αυτά εξετάζονται από τη νευροβιολογία και την κοινωνιογλωσσολογία. Και οι δύο επιστήμες έχουν παράγει σημαντικό ερευνητικό έργο στη έμφυλη διαφοροποίηση της γλωσσικής χρήσης το οποίο μόλις πρόσφατα άρχισε να συνεξετάζεται και να αξιολογείται διεπιστημονικά.

Η ερευνητική πιστοποίηση της έμφυλης γλωσσικής διαφοροποίησης συνεπάγεται ότι άνδρες και γυναίκες παράγουν γλωσσικές δομές οι οποίες διαφοροποιούνται τόσο σε δομικό όσο και σε λειτουργικό επίπεδο. Ο εντοπισμός αυτών των γλωσσικών χαρακτηριστικών, αλλά και η ποσοτική αξιολόγηση της επίδρασης τους στην έμφυλη διαφοροποίηση της γλωσσική χρήσης είναι το αντικείμενο της δεύτερης ερευνητικής φάσης. Τέλος, τα γλωσσικά χαρακτηριστικά που αξιολογήθηκαν στο δεύτερο στάδιο ως τα πιο σημαντικά όσον αφορά τη σχέση τους με το φύλο του συγγραφέα, εντάσσονται σε έναν ειδικό μηχανισμό μηχανικής μάθησης (Τεχνητό Νευρωνικό Δίκτυο - ΤΝΔ) το οποίο αφού εκπαιδευθεί με αυτά σε κείμενα όπου το φύλο του συγγραφέα είναι γνωστό, γενικεύει τους κανόνες μάθησης και ελέγχει την προβλεπτική ικανότητα των συγκεκριμένων χαρακτηριστικών σε κείμενα όπου το σύστημα δε γνωρίζει το φύλο του συγγραφέα.

[1] Η παρούσα έρευνα χρηματοδοτήθηκε από τον Ειδικό Λογαριασμό Κονδυλίων Έρευνας (ΕΛΚΕ) του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών μέσω του ερευνητικού προγράμματος ΚΑΠΟΔΙΣΤΡΙΑΣ (2009) με κωδικό αριθμό 70/4/8871.

2 Θεωρητικές προσεγγίσεις στις διαφορές των δύο φύλων σχετικά με τη γλωσσική ικανότητα και τη γλωσσική επιτέλεση

2.1 Βιολογικές διαφορές στον εγκέφαλο

2.1.1 Ανατομικές διαφορές

Για ένα μεγάλο χρονικό διάστημα η διαφοροποιημένη συμπεριφορά ανδρών και γυναικών αποδίδετο στην επίδραση της πολιτισμικής εκπαίδευσης και των πιέσεων που ασκούνται από τις ποικίλες διαδικασίες κοινωνικοποίησης που υφίσταται κάθε παιδί που μεγαλώνει σε ένα συγκεκριμένο κοινωνικό χώρο. Ωστόσο, τα τελευταία χρόνια μια σειρά από μελέτες που απέκλεισαν πειραματικά τον παράγοντα της πολιτισμικής εκμάθησης έδειξαν ότι μεγάλο μέρος της διαφοροποιημένης συμπεριφοράς των δύο φύλων έχει τη βάση του σε βιολογικές διαφορές με σημαντικότερη αυτή της ανατομίας του ίδιου του εγκεφάλου. Είναι άλλωστε χαρακτηριστικό πώς ήδη από τους πρώτους μήνες ζωής ενός μωρού το φύλο διαφοροποιεί σημαντικά τη συμπεριφορά του. Οι Anne Moir & David Jessel[2] αναφέρουν χαρακτηριστικά για τις έμφυλες διαφορές σε μωρά:

> *Αυτές οι διακριτές και μετρήσιμες διαφορές στη συμπεριφορά έχουν εγγραφεί πολύ πριν οι οποιεσδήποτε εξωτερικές επιδράσεις αποκτήσουν την ευκαιρία να λειτουργήσουν. [Οι διαφορές αυτές] αντανακλούν μια βασική διαφοροποίηση στον εγκέφαλο του νεογέννητου που ήδη γνωρίζουμε – την ανωτερότητα του ανδρικού εγκεφάλου στη χωρική επεξεργασία (spatial ability) και την αδιαμφισβήτητη ικανότητα του γυναικείου εγκεφάλου στη γλωσσική χρήση.*

Μια σημαντική ανατομική διαφοροποίηση μεταξύ ανδρικών και γυναικείων εγκεφάλων σχετίζεται με το μεσολόβιο (corpus callosum). Οι περισσότερες μελέτες συνδέουν αυτή τη διαφορά με αυξημένη ικανότητα των γυναικείων εγκεφάλων να συντονίζουν τα δύο ημισφαίρια τους. Η διαφορά αυτή έχει συσχετιστεί κατά καιρούς με τη γυναικεία «διαίσθηση»[3] και το μουσικό ταλέντο[4]. Πιο πρόσφατες έρευνες[5] έχουν επιβεβαιώσει την ανατομική διαφορά αλλά δεν καταλήγουν σε ακριβείς συσχετίσεις με γνωστικές ικανότητες.

Σημαντική διαφοροποίηση υφίσταται επίσης στο τμήμα του εγκεφάλου που ονομάζεται κατώτερος βρεγματικός λοβός (inferior-parietal lobule)[6]. Η περιοχή αυτή βρίσκεται λίγο πάνω από τα αυτιά στο ύψος των κροτάφων και εμφανίζεται σημαντικά μεγαλύτερη στους άνδρες σε σχέση με τις γυναίκες. Επίσης, στους άντρες ο δεξιός λοβός είναι μεγαλύτερος σε σχέση με τον αριστερό, ενώ στις γυναίκες αυτή η ασυμμετρία αντιστρέφεται. Ο δεξιός λοβός έχει συνδεθεί με την προσωρινή μνήμη που χρειάζεται ο εγκέφαλος για να καταλάβει και να χειριστεί χωρικές συσχετίσεις

---

[2] Moir, Anne, and David Jessel. *Brain Sex: The Real Difference between Men and Women*. New York: Delta, 1992.

[3] Gorman, Christine, and Madeleine Nash. "Sizing up the Sexes." *TIME* 20 January 1992: 36-43.

[4] Levitin, Daniel. *This Is Your Brain on Music: The Science of a Human Obsession*. New York: Dutton Adult, 2006.

[5] Clarke, Dave, et al. "Corpus Callosotomy: A Palliative Therapeutic Technique May Help Identify Resectable Epileptogenic Foci." *Seizure* 16.6 (2007): 545-53.

[6] Frederikse, Melissa E., et al. "Sex Differences in the Inferior Parietal Lobule." *Cereb. Cortex* 9.8 (1999): 896-901.

καθώς και την ικανότητα να αντιληφθεί τις σχέσεις που υφίστανται μεταξύ των διαφορετικών μερών του σώματος. Επίσης σχετίζεται με την αντίληψη των δικών μας συναισθημάτων. Αντίθετα, ο αριστερός λοβός εμπλέκεται περισσότερο στην αντίληψη του χρόνου και της ταχύτητας καθώς και στην ικανότητα της νοητικής περιστροφής τρισδιάστατων εικόνων.

2.1.2 Μελέτες Λειτουργικής Μαγνητικής Απεικόνισης (Functional Magnetic Resonance Imaging – fMRI)

Μια από τις σημαντικότερες εξελίξεις στο χώρο της διαγνωστικής απεικόνισης είναι η ανάπτυξη της Λειτουργικής Μαγνητικής Απεικόνισης - ΛΜΑ (Functional Magnetic Resonance Imaging – fMRI). Η συγκεκριμένη τεχνική όταν εφαρμόζεται στον εγκέφαλο μπορεί να δείξει σε πραγματικό χρόνο σε ποια τμήματα του εγκεφαλικού φλοιού παρουσιάζεται αυξημένη δραστηριότητα μετρώντας την ποσότητα του αίματος που κυκλοφορεί. Η χρήση ΛΜΑ σε πειραματικές συνθήκες με ελεγχόμενα ερεθίσματα μπορεί να αποκαλύψει ποιες περιοχές του εγκεφάλου σχετίζονται με συγκεκριμένες δεξιότητες.

Το βασικό απεικονιστικό εύρημα σε γλωσσικές δοκιμασίες είναι η λειτουργική πλευρίωση (functional lateralization) που παρατηρείται στις ΛΜΑ ανδρών[7], δηλαδή η χρήση μόνο του αριστερού λοβού για την επεξεργασία γλωσσικών δεδομένων. Αντίθετα οι γυναίκες φαίνεται να χρησιμοποιούν και τα δύο ημισφαίρια του εγκεφαλικού φλοιού όταν παράγουν, αλλά και όταν ακούν ανθρώπινη ομιλία. Η παράλληλη χρήση των δύο ημισφαιρίων στο γυναικείο εγκέφαλο έχει εξηγηθεί μερικώς και από το μεγαλύτερο μεσολόβιο που έχει ο γυναικείος εγκέφαλος (βλ. και παραπάνω στην υποενότητα 2.1.1). Η αμφίπλευρη λειτουργία των εγκεφαλικών ημισφαιρίων αποτελεί αυτή τη στιγμή τη βασικότερη βιολογική ερμηνεία της γυναικείας ανωτερότητας στη γλωσσική επεξεργασία. Η δυνατότητα κατανεμημένης γλωσσικής επεξεργασίας επιτρέπει ταχύτερη και ακριβέστερη επεξεργασία γλωσσικών δεδομένων[8]. Αντίθετα, ως αποτέλεσμα της πλευριωμένης γλωσσικής λειτουργίας οι άνδρες παρουσιάζουν διπλάσια ποσοστά στη δυσλεξία[9] και σημαντικά υψηλότερα ποσοστά αφασίας σε εγκεφαλικά[10]. Σχετικές έρευνες[11] δείχνουν ότι από τους ασθενείς που παθαίνουν κάποιου είδους βλάβη στο αριστερό ημισφαίριο του εγκεφάλου, περισσότεροι άνδρες (48,5%) από γυναίκες (30%) εμφανίζουν σημάδια αφασίας. Η σχέση περιοχής βλάβης και φύλου έχει προσδιοριστεί πλέον με μεγαλύτερη ακρίβεια και τώρα γνωρίζουμε ότι όταν πλήττεται το αριστερό πρόσθιο τμήμα του εγκεφαλικού φλοιού οι γυναίκες εμφανίζουν μεγαλύτερα ποσοστά αφασίας, ενώ όταν πλήττεται το πίσω αριστερό τμήμα του πρόσθιου εγκεφαλικού φλοιού εμφανίζονται περισσότεροι άνδρες με συμπτώματα αφασίας.

---

[7] Shaywitz, Bennet A. , et al. "Sex Differences in the Functional Organization of the Brain for Language." *Nature* 373 (1995): 607-09.

[8] Linn, Marcia C., and Anne C. Petersen. "Emergence and Characterization of Sex Differences in Spatial Ability: A Meta-Analysis." *Child Development* 56 (1985): 1479-98, Kimura, Doreen. *Sex and Cognition*. Cambridge, MA: MIT Press, 2000.

[9] Flannery, Kathleen A., et al. "Male Prevalence for Reading Disability Is Found in a Large Sample of Black and White Children Free from Ascertainment Bias." *Journal of the International Neuropsychological Society* 6.4 (2000): 433-42.

[10] McGlone, Jeanette "Sex Differences in Human Brain Organization: A Critical Survey." *Behavioral Brain Science* 3 (1980): 215-27.

[11] Kimura, Doreen. *Neuromotor Mechanisms in Human Communication*. Oxford: Oxford University Press, 1993. Kimura, Doreen, and Elizabeth Hampson. "Cognitive Pattern in Men and Women Is Influenced by Fluctuations in Sex Hormones." *Current Directions in Psychological Science* 3.2 (1994): 57-61.

Οι μελέτες ΛΜΑ που έχουν γίνει τα τελευταία χρόνια δεν έχουν παρουσιάσει απόλυτη ομοφωνία ως προς την πλευρίωση της γλωσσικής ικανότητας στους άνδρες. Η μεγαλύτερη μετα-ανάλυση δεδομένων ΛΜΑ[12] κατέληξε στο συμπέρασμα ότι με τα μέχρι τώρα δεδομένα δεν μπορεί να γίνει αποδεκτή με βεβαιότητα η θεωρία της λειτουργικής πλευρίωσης στο γενικό πληθυσμό. Ωστόσο συγκεκριμένες γλωσσικές δοκιμασίες διαφέρουν σημαντικά ως προς τον τόπο ενεργοποίησης μεταξύ ανδρών και γυναικών[13] δίνοντας βάση στην βιολογική διαφοροποίηση της γλωσσικής ικανότητας των δύο φύλων.

2.2 Κοινωνιογλωσσολογικές διαφορές

Η γλωσσική διαφοροποίηση μεταξύ ανδρών και γυναικών ήρθε έντονα στην επιφάνεια με την ανάπτυξη της κοινωνιογλωσσολογίας. Αν και τα ευρήματα στο συγκεκριμένο κλάδο είναι πολυάριθμα και καλύπτουν το σύνολο του θεωρητικού φάσματος που ενδύει τη μελέτη της γλώσσας στο κοινωνικό της περιβάλλον, εδώ θα επικεντρωθούμε μόνο στα ερευνητικά αποτελέσματα και τις θεωρητικές ερμηνείες που προκύπτουν μέσα από τον κλάδο της ποσοτικής κοινωνιογλωσσολογίας. Ο λόγος είναι ότι ο συγκεκριμένος κλάδος ασχολείται με την ποσοτική ανάλυση της γλωσσικής ποικιλίας και επομένως είναι απόλυτα συμβατός με το μεθοδολογικό πλαίσιο της παρούσας έρευνας.

Ένα από τα πιο ενδιαφέροντα χαρακτηριστικά της σχέσης της γλωσσικής χρήσης με το φύλο του ομιλητή είναι το ότι σε σχεδόν σταθερή βάση αποκαλύπτεται μεταξύ τους ένα είδος τυποποιημένης σχέσης, η μορφή της οποίας έχει επαναληφθεί και επιβεβαιωθεί σε διάφορες γλώσσες και πολιτισμούς. Πρόκειται για το φαινόμενο που ο Fasold[14] ονομάζει «κοινωνιογλωσσολογικό πρότυπο γένους» (sociolinguistic gender pattern). Είναι η κατάσταση στην οποία οι άνδρες ομιλητές χρησιμοποιούν συχνότερα από τις γυναίκες γλωσσικούς τύπους που είναι κοινωνικά στιγματισμένοι και αποκλίνοντες της γλωσσικής νόρμας. Αντίθετα, οι γυναίκες δείχνουν φανερή προτίμηση στους περισσότερο «σωστούς» γλωσσικούς τύπους και στην προφορά κύρους (prestige pronunciation).

Το παραπάνω πρότυπο συναντήθηκε ήδη από την πρώτη ποσοτική κοινωνιογλωσσολογική έρευνα που έγινε ευρύτερα γνωστή[15] και επιβεβαιώθηκε από πλήθος άλλων μελετών ανάλυσης της κοινωνιογλωσσολογικής ποικιλίας[16].

Ο Labov[17] συνοψίζοντας τα αποτελέσματα των κοινωνιογλωσσολογικών μελετών που έχουν γίνει μέχρι το 1982 σε μια ποικιλία εθνών κατέληξε για τις γλωσσικές διαφορές φύλου στο εξής:

> *Η γενική αρχή που αναδύθηκε από μελέτες στην Ευρώπη, Καναδά, ΗΠΑ και Λατινική Αμερική είναι ότι οι γυναίκες είναι περισσότερο συντηρητικές στις αντιδράσεις τους όσον αφορά την κοινωνικά αναγνωρισμένη ποικιλία.*

---

[12] Sommer, Iris, et al. "Do Women Really Have More Bilateral Language Representation Than Men? A Meta-Analysis of Functional Imaging Studies." *Brain* 127.8 (2004): 1845-52.

[13] Harrington, Greg S., and Sarah Tomaszewski Farias. "Sex Differences in Language Processing: Functional Mri Methodological Considerations." *Journal of Magnetic Resonance Imaging* 27 (2008): 1221-28.

[14] Fasold, Ralph W. *The Sociolinguistics of Language*. Oxford: Blackwell, 1990.

[15] Fischer, John L. "Social Influences on the Choice of a Linguistic Variant." *Word* 14 (1958): 47-56.

[16] Ενδ. Wolfram, Walt. *A Sociolinguistic Description of Detroit Negro Speech*. Washington, DC: Center for Applied Lingustics, 1969, Milroy, Lesley. *Language and Social Networks*. Oxford: Blackwell, 1980.

[17] Labov, William. "Building on Empirical Foundations." *Perspectives on Historical Linguistics*. Eds. Winfred P. Lehmann and Yakov Malkiel. Amsterdam & Philadelphia: John Benjamins, 1982. 79-92.

Η μοναδική περίπτωση αντιστροφής του κοινωνιογλωσσολογικού προτύπου γένους εντοπίστηκε στις αραβικές κοινωνίες, με τις γυναίκες να χρησιμοποιούν πιο σπάνια γλωσσικούς τύπους κύρους (κλασικά αραβικά). Αλλά και σε αυτήν την περίπτωση πιο σύγχρονες αναλύσεις[18] έδειξαν ότι όλες οι μελέτες που παρουσίαζαν τους άνδρες να προτιμούν τους γλωσσικούς τύπους κύρους δεν είχαν συνυπολογίσει την αραβική διγλωσσία.

Μια από τις πιο πολυσυζητημένες ερμηνείες ανήκει σε μια σειρά από ερευνητές οι οποίοι επιχείρησαν να ερμηνεύσουν το κοινωνιογλωσσολογικό πρότυπο γένους μέσα από τη υιοθέτηση στερεότυπων ρόλων που επιβάλλει η κοινωνία στα φύλα. Η Key[19] ισχυρίστηκε πως οι γυναίκες μέσα από την υιοθέτηση κοινωνικά καταξιωμένων γλωσσικών τύπων επιχειρούν να αποκτήσουν το κύρος που οι ανδροκρατικές κοινωνίες τους αφαιρούν στις υπόλοιπες εκφάνσεις της ζωής τους. Μέσα από τη γλώσσα ασυνείδητα διεκδικούν την ισότητα και την εξίσωση με το ανδρικό φύλο. Σε μια σειρά από παρόμοιες διαπιστώσεις οδηγείται και ο Trudgill[20] ο οποίος ερμηνεύει το κοινωνιογλωσσολογικό πρότυπο γένους μέσα από πολλές και διαφορετικές μεταξύ τους οπτικές.

Βασική θέση του Trudgill είναι ότι οι γυναίκες γενικά είναι περισσότερο ευαίσθητες στην αντίληψη του κύρους που ενέχουν διάφοροι κοινωνικοί θεσμοί και επομένως και η γλώσσα. Παράλληλα, η θέση των γυναικών μέχρι πρόσφατα στην κοινωνία ήταν χαμηλότερη των ανδρών. Η ανισότητα αυτή για τις γυναίκες αποτελεί μια μόνιμη πηγή ανασφάλειας η οποία αντιμετωπίζεται μέσα από πλάγιες μεθόδους, αφού τα κοινωνικά στερεότυπα αποτελούν φραγμό στις άμεσες ανατροπές των ανδροκρατικών ρόλων. Ένας από τους έμμεσους τρόπους ανάκτησης του περιορισμένου κύρους της γυναίκας είναι και η υιοθέτηση κοινωνικά «υψηλής» ομιλίας.

Συμπληρωματικά ο Trudgill ισχυρίζεται ότι το κοινωνιογλωσσολογικό πρότυπο γένους ενισχύεται λόγω του ότι οι γυναίκες στις περισσότερες κοινωνίες ασχολούνται με την ανατροφή των παιδιών και αναλαμβάνουν τη μετάδοση της πρωτογενούς γνωστικής και πολιτισμικής πληροφορίας σε αυτά. Μέσω αυτής της διαδικασίας αποκτούν μεγαλύτερη συνείδηση της κοινωνικά καταξιωμένης γλωσσικής ποικιλίας και την υιοθετούν για να μπορέσουν να τη μεταφέρουν στα παιδιά τους έτσι ώστε αυτά να έχουν ένα επιπλέον εφόδιο στην κοινωνική τους μετεξέλιξη.

Μια τρίτη ερμηνεία έρχεται μέσα από την εξέταση του κοινωνιογλωσσολογικού προτύπου γένους σε σχέση με την κοινωνική τάξη των ομιλητών. Το φύλο αλληλεπιδρά με την κοινωνική τάξη συχνά και σε πολλαπλά επίπεδα. της γλωσσικής παραγωγής. Όπως αναφέρει και ο Trudgill[21]:

> *. . . η γλώσσα της ΕΤ [εργατικής τάξης], όπως και άλλες πλευρές της πολιτισμικής ταυτότητας της ΕΤ, εμφανίζεται, τουλάχιστο σε ορισμένες δυτικές κοινωνίες, να παρουσιάζει συνδηλώσεις με την ανδροπρέπεια . . . κυρίως γιατί συνδέεται με την τραχύτητα και τη σκληρότητα που υποτίθεται ότι έχει η ζωή της ΕΤ . . . και τα οποία θεωρούνται επιθυμητά ανδρικά χαρακτηριστικά. Από την*

---

[18] Abd-el-Jawad, Hassan R. "Cross-Dialectal Variation in Arabic: Competing Prestigious Forms." *Language in Society* 16 (1987): 359-68, Haeri, Niloofar. "Male/Female Differences in Speech: An Alternative Interpretation." *Variation in Language: Nwav - Xv at Stanford. Proceedings of the Fifteenth Annual Conference on New Ways of Analyzing Variation*. Eds. Keith M. Denning, et al. Stanford: Department of Linguistics, Stanford University, 1987. 173-96.

[19] Key, Mary Ritchie. *Male/Female Language*. Metuchen, NJ: Scarecrow Press, 1975.

[20] Trudgill, Peter. *On Dialect: Social and Geographic Factors*. Oxford: Blackwell, 1978.

[21] Trudgill, Peter. "Sex, Covert Prestige and Linguistic Change in the Urban British Engliish of Norwich." *Language in Society* 1 (1972): 179-95.

άλλη, [αυτά] δεν θεωρούνται επιθυμητά γυναικεία χαρακτηριστικά.

Μέσα από μια τέτοια συνδηλωτική λειτουργία οι προφορές «χαμηλού» κύρους, που συναντώνται συχνά σε ομιλητές της εργατικής τάξης, συνδέονται με την ομιλία των αντρών, ενώ αντίθετα οι γυναίκες ταυτίζονται με την επιλογή γλωσσικών τύπων που χαίρουν ευρύτερης αποδοχής στις αστικές νόρμες. Στην πραγματικότητα η παραπάνω συνδήλωση αποτελεί το μόρφωμα κάποιων κρυμμένων αξιών που συνοδεύουν την ομιλία που δεν ακολουθεί τη νόρμα και ονομάστηκε «κεκαλυμμένο κύρος» (covert prestige). Ο βασικός μηχανισμός λειτουργίας του κεκαλυμμένου κύρους είναι ότι βοηθάει τις κοινωνικές ομάδες που ανήκουν σε κάποιου είδους μειονότητας ή κοινωνικού περιθωρίου να επιλέξουν χαρακτηριστικά της γλωσσικής τους συμπεριφοράς ως σύμβολα περιχαράκωσης της θέσης και των αξιών τους στην κοινωνία που βρίσκονται. Έτσι υιοθετούνται γλωσσικά στοιχεία που, αν και γίνονται αντιληπτά από την πλειονότητα των ομιλητών ως στιγματισμένοι τύποι, την ίδια στιγμή τα ίδια γλωσσικά στοιχεία μέσα στα πλαίσια της συγκεκριμένης γλωσσικής κοινότητας σηματοδοτούνται θετικά και η χρήση τους φέρνει καταξίωση και αποδοχή από τα μέλη του μειονοτικού συνόλου. Στο σημείο αυτό εδράζεται και η βασική διαφοροποίηση της ανδρικής και της γυναικείας γλωσσικής συμπεριφοράς. Οι άνδρες προτιμούν το κεκαλυμμένο κύρος και την αποδοχή που επιφέρει από τους ομότιμούς τους, ενώ οι γυναίκες στρέφονται στη γλωσσική νόρμα της «μεσαίας τάξης» η οποία τους δίνει τη δυνατότητα να γίνουν ευρύτερα αποδεκτές από το κοινωνικό περιβάλλον πέραν της συγκεκριμένης κοινωνικής τάξης στην οποία ανήκουν.

Σε ένα διαφορετικό θεωρητικό πρότυπο στηρίζει τις απόψεις της για τις διαφορές φύλου στη γλώσσα η Deuchar[22]. Χρησιμοποιεί τη θεωρία των Brown & Levinson[23] σχετικά με τη διαπροσωπική στρατηγική του κοινωνικού προσώπου (face) των συνομιλητών για να ερμηνεύσει το κοινωνιογλωσσολογικό πρότυπο γένους. Η παραπάνω θεωρία προβλέπει διατάραξη της ισορροπίας του θετικού και αρνητικού κοινωνικού προσώπου του ομιλητή όταν υπάρχει διαφορά δύναμης ή εξουσίας μεταξύ αυτού και του συνομιλητή του. Στην περίπτωση αυτή ο βρισκόμενος σε κατώτερη θέση θα υποχρεωθεί να ενισχύσει το θετικό πρόσωπο του συνομιλητή του υιοθετώντας υπερβολικά ευγενική γλώσσα που θα μπορούσε να φτάσει σε σημεία δουλοπρέπειας. Αντίθετα, ο ομιλητής που κατέχει θέση δύναμης θα αδιαφορήσει για το θετικό πρόσωπο του συνομιλητή του και θα χρησιμοποιήσει γλώσσα που θα το πλήξει. Υπό αυτό το θεωρητικό πρίσμα η Deuchar ισχυρίζεται ότι η προτίμηση στη χρήση των γλωσσικών τύπων κύρους από τις γυναίκες είναι ιδανική για να προστατευθεί το πρόσωπο τους χωρίς παράλληλα να επιχειρείται μια απευθείας «επίθεση» στο συνομιλητή τους. Με αυτή την επιλογή τους οι γυναίκες που βλέπουν την ισορροπία αρνητικού και θετικού κοινωνικού προσώπου να διαταράσσεται από την άνιση σχέση που υπάρχει με τους άντρες, προστατεύουν το πρόσωπό τους. Αυτό το κάνουν αξιοποιώντας τις θετικές συνδηλώσεις που προκαλεί η χρήση των γλωσσικών τύπων κύρους, διατηρώντας έτσι το πρόσωπο του άντρα συνομιλητή τους ανέπαφο.

Ωστόσο η παραπάνω θέση ερμηνεύει το κοινωνιογλωσσολογικό πρότυπο γένους αποκλειστικά μέσα από το μηχανισμό της εξουσίας και της πίεσης που ασκείται στις

---

[22] Deuchar, Margaret. "A Pragmatic Account of Women's Use of Standard Speech." *Women in Their Speech Communities: New Perspectives on Language and Sex*. Eds. Deborah Cameron and Jennifer Coates. London: Longman, 1988. 27-32.

[23] Brown, Penelope, and Steven Levinson. "Universals in Language Usage: Politeness Phenomena." *Questions and Politeness: Strategies in Social Interaction*. Ed. Esther N. Goody. Cambridge: Cambridge University Press, 1978. 56-311.

διαπροσωπικές σχέσεις και αγνοεί άλλες ψυχολογικές ή ακόμα και βιολογικές αιτίες που υφίστανται και έχει αποδειχθεί ότι συμμετέχουν στη διαμόρφωση του γλωσσικού παραγόμενου. Παράλληλα, μια τέτοια θεώρηση προβλέπει ότι η γλωσσική παραγωγή των γυναικών θα συμφωνεί με το κοινωνιογλωσσολογικό πρότυπο γένους εφόσον οι συνομιλητές τους θα είναι άντρες και όχι γυναίκες, γεγονός που έχει επιβεβαιωθεί από μερικές έρευνες[24], αλλά δεν αποτελεί μια καθολικά διαπιστωμένη πραγματικότητα.

3 Αυτόματη κατηγοριοποίηση κειμένων βάσει του φύλου του συγγραφέα

Μέχρι στις αρχές της δεκαετίας του ’90 ο μεγαλύτερος όγκος έρευνας στη διερεύνηση της επίδρασης του φύλου στις γλωσσικές επιλογές του ομιλητή / συγγραφέα γίνεται στο πλαίσιο της κοινωνιογλωσσολογίας με αποτέλεσμα να αναλύονται κυρίως ο προφορικός λόγος των ομιλητών με έμφαση στη φωνητική ποικιλία[25]. Τα τελευταία χρόνια όμως, παράλληλα με την αυξημένη διαθεσιμότητα των ΗΣΚ αναπτύχθηκε ένα πλήθος τεχνικών μηχανικής μάθησης οι οποίες ενσωματώθηκαν σε λογισμικά ανοιχτού κώδικα (open source) και έγιναν διαθέσιμα σε όλη την ερευνητική κοινότητα. Ο συνδυασμός γλωσσικών δεδομένων και ελεύθερου λογισμικού μηχανικής μάθησης δημιούργησε έναν ταχύτατα αναπτυσσόμενο υπο-κλάδο της Τεχνητής Νοημοσύνης που ονομάζεται Ανάκτηση Πληροφορίας (Information Retrieval). Ο συγκεκριμένος κλάδος εκτός της άμεσης ενασχόλησής του με θέματα που άπτονται των μηχανών αναζήτησης του διαδικτύου, ασχολείται ενεργά με ζητήματα κατηγοριοποίησης κειμένων βάσει συγκεκριμένων χαρακτηριστικών, όπως είναι το θέμα (text topic categorization), η απόδοση ενός κειμένου αγνώστης πατρότητας στο συγγραφέα του (authorship attribution) κ.ά. Στο πλαίσιο αυτής της προσέγγισης, τα τελευταία χρόνια έχουν γίνει κάποιες προσπάθειες αυτόματης κατηγοριοποίησης κειμένων βάσει του φύλου του συγγραφέα η οποίες ανάγονται στο γενικότερο ερευνητικό ερώτημα του καθορισμού των δημογραφικών και ψυχολογικών χαρακτηριστικών του συγγραφέα (author profiling).

Μια από τις πρώτες προσπάθειες πρόβλεψης του φύλου του συγγραφέα σε αγγλικά κείμενα έγινε από τους Koppel, Argamon & Shimoni[26]. Ως ΗΣΚ εκπαίδευσης χρησιμοποίησαν ένα υποσύνολο του British National Corpus – BNC (566 κείμενα ισόποσα μοιρασμένα σε άνδρες και γυναίκες) το οποίο ανέπτυξαν με τέτοιο τρόπο ώστε να ελέγχεται η επίδραση του κειμενικού γένους σε αυτό. Σε αυτό το ΗΣΚ μέτρησαν μια σειρά από υφομετρικούς δείκτες που δεν επηρεάζονται από το θέμα του κειμένου και περιλαμβάνουν τις 405 πιο συχνές λειτουργικές λέξεις και τα πιο συχνά ν-λεκτα των Μερών του Λόγου (n-grams Part of Speech Tags). Συνολικά χρησιμοποιήθηκαν 1081 χαρακτηριστικά τα οποία εκπαίδευσαν ένα αλγόριθμο γραμμικού διαχωρισμού των κατηγοριών. Η ακρίβεια πρόβλεψης του φύλου του συγγραφέα κυμάνθηκε από 79,5% στα κείμενα λογοτεχνίας έως 82,6% σε μη λογοτεχνικά κείμενα. Ένα από τα πιο ενδιαφέροντα ευρήματα ήταν ότι τα κείμενα λογοτεχνίας στηρίζονταν σε διαφορετικούς κανόνες και γλωσσικά χαρακτηριστικά για να δηλώσουν το φύλο του συγγραφέα σε σχέση με τα μη λογοτεχνικά κείμενα.

[24] Brouwer, Dédé, Marinel Gerritsen, and Dorian De Haan. "Speech Differences between Women and Men: On the Wrong Track?" *Language in Society* 8 (1979): 33-50.

[25] Juola, Patrick. "Authorship Attribution." *Foundations and Trends in Information Retrieval* 1.3 (2008): 233-334.

[26] Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. "Automatically Categorizing Written Texts by Author Gender." *Literary and Linguistic Computing* 17.4 (2002): 401-12.

Επίσης, μια σειρά από λέξεις που παλαιότερες έρευνες είχαν καταδείξει ως χαρακτηριστικές της γυναικείας επικοινωνίας (αντωνυμίες) και της ανδρικής (οριστικά άρθρα), φάνηκε ότι λειτουργούσαν και σε αυτή την έρευνα ως δείκτες του φύλου του συγγραφέα.

Η ίδια ερευνητική ομάδα εξέτασε ένα μεγάλο ΗΚΣ από ιστολόγια (blogs) (37,478 ιστολόγια με 300 εκ. λέξεις) και προσπάθησε να προβλέψει τόσο το φύλο όσο και την ηλικία των συγγραφέων[27]. Τα γλωσσικά χαρακτηριστικά που μετρήθηκαν ανήκουν στην κατηγορία των συχνών λειτουργικών λέξεων, των Μερών του Λόγου και των τυποτεχνικών συμβάσεων που χρησιμοποιούνται στα ιστολόγια, δηλ. σύνδεσμοι (hyperlinks), σύμβολα ψυχικών καταστάσεων (emoticons), λέξεις που χαρακτηρίζουν τη γλώσσα των ιστολογίων όπως π.χ. lol, haha κ.ά. Επίσης στη συγκεκριμένη μελέτη χρησιμοποιήθηκαν λίστες λέξεων περιεχομένου. Ο συνολικός αριθμός χαρακτηριστικών που χρησιμοποιήθηκε έφτασε τα 1502 και εκπαίδευσε τον αλγόριθμο μηχανικής μάθησης Multi-Class Real Winnow του οποίου η ακρίβεια στην πρόβλεψη φύλου έφτασε στο 80,1%. Ενδιαφέρον παρουσιάζει η διαπίστωση των συγγραφέων ότι παρά τη μεγαλή διαφοροποίηση που βρέθηκε μεταξύ στερεότυπων λέξεων περιεχομένου μεταξύ ανδρών και γυναικών, το μεγαλύτερο βάρος στη διάκριση του φύλου του συγγραφέα έπεσε στη χρήση των χαρακτηριστικών που είναι σημασιολογικά ουδέτερα (όπως οι συχνές λειτουργικές λέξεις και τα Μέρη του Λόγου).

Ο Corney[28] σε μια μελέτη κατηγοριοποίησης των μηνυμάτων ηλεκτρονικού ταχυδρομείου, εξέτασε την ακρίβεια πρόβλεψης του φύλου του αποστολέα ενός μηνύματος μετρώντας μια σειρά από υφομετρικά χαρακτηριστικά όπως, συχνότερες λειτουργικές λέξεις, μέγεθος λέξης και πρότασης κ.ά. Η μεγίστη ακρίβεια πρόβλεψης έφτασε στο 70,1% και οι σημαντικότερες μεταβλητές ήταν οι συχνότερες λειτουργικές λέξεις, το μέγεθος λέξης και η συχνότητα των γραμμάτων.

Τέλος, θα ήταν χρήσιμο να αναφερθούμε στην εργασία των Hota et al.[29] η οποία εξέτασε τη γλώσσα ανδρικών και γυναικείων χαρακτήρων σε 34 έργα του Σαίξπηρ. Το βασικό ερευνητικό ερώτημα που τέθηκε είναι κατά πόσο ένας άνδρας συγγραφέας όπως ο Σαίξπηρ καταφέρνει να προσεγγίσει τα χαρακτηριστικά της γυναικείας ομιλίας και να τα περάσει στο κείμενο όταν γράφει τους διαλόγους των γυναικών που μετέχουν στο έργο του. Χρησιμοποιήθηκαν τόσο σημασιολογικά ουδέτερα χαρακτηριστικά όπως οι συχνότερες λειτουργικές λέξεις, αριθμοί, προθέσεις και συντμήσεις καθώς και οι πιο συχνές λέξεις του ΗΣΚ (λέξεις με συχνότητα εμφάνισης μεγαλύτερη του 10). Η ακρίβεια πρόβλεψης του φύλου του συγγραφέα κυμάνθηκε μεταξύ 60% και 75% ανάλογα με τα χαρακτηριστικά που χρησιμοποιήθηκαν. Οι συγγραφείς ερμηνεύουν την κάπως μικρότερη ακρίβεια αναγνώρισης του φύλου, στο γεγονός ότι ο Σαίξπηρ αν και διαισθητικά πλησίασε τη γλώσσα των γυναικών χαρακτήρων του, δεν κατάφερε να την αποδώσει στην ολότητά της.

---

[27] Schler, Jonathan, et al. "Effects of Age and Gender on Blogging." *Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*. 2006.

[28] Corney, Malcolm Walter. "Analysing E-Mail Text Authorship for Forensic Purposes." Queensland University of Technology, 2003.

[29] Hota, Sobhan R., et al. "Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters." *Proceedings of Digital Humanities 2006*. 2006. 100-04.

4 ΗΣΚ και υφομετρικά χαρακτηριστικά

4.1 Περιγραφή του ΗΣΚ

Ένα σημαντικό πρόβλημα στις μελέτες αυτόματης κατηγοριοποίησης κειμένων είναι η έλλειψη ομοιογένειας των ΗΣΚ που χρησιμοποιούνται για εκπαίδευση (training corpus) και στα ΗΣΚ που αξιοποιούνται στην επικύρωση (validation corpus) του αλγόριθμου κατηγοριοποίησης. Ο Rudman[30] εξετάζοντας ευρύτερα τα διάφορα ΗΣΚ που χρησιμοποιούνται στις υφομετρικές μελέτες αναγνώρισης του συγγραφέα εντόπισε τις ακόλουθες αδυναμίες που σχετίζονται με την ποσοτική και ποιοτική σύστασή τους:

- Η ακατάλληλη επιλογή, η μη διαθεσιμότητα και η αποσπασματικότητα των κειμένων.
- Η κανονικοποίηση των ορθογραφικών συμβάσεων, της κωδικοποίησης και γενικότερα η επέμβαση του εκδότη ή του επιμελητή της έκδοσης στις τυποτεχνικές αλλά και στις γλωσσικές επιλογές του συγγραφέα .
- Τα κείμενα που χρησιμοποιούνται για διασταυρούμενη επικύρωση (cross-validation) θα πρέπει να ταιριάζουν με τα κείμενα εκπαίδευσης ως προς το κειμενικό γένος, το κειμενικό μέσο και τη χρονολογία.

Οι παραπάνω επισημάνσεις σχετίζονται με μια σημαντική ιδιότητα της γλωσσικής ποικιλίας, αυτήν της πολυπαραγοντικότητας. Μια συγκεκριμένη γλωσσική επιλογή σε ένα κείμενο μπορεί να είναι συνιστώσα αλληλεπιδράσεων που παρουσιάζει το κειμενικό θέμα με το κειμενικό γένος ή οποιουδήποτε άλλου συνδυασμού μεταδεδομένων. Ακόμα και οι πιο αφηρημένες και θεωρητικά «ουδέτερες» γλωσσικές μεταβλητές εμφανίζουν σημαντική συσχέτιση με εξωγλωσσικές μεταβλητές όπως το κειμενικό θέμα και το κειμενικό γένος. Σε μια πρόσφατη μελέτη[31], εξετάστηκαν μια σειρά από συχνά χρησιμοποιούμενες υφομετρικές μεταβλητές που χρησιμοποιούνται ευρέως σε μελέτες αναγνώρισης του αγνώστου συγγραφέα (author attribution studies) σε ένα ΗΣΚ το οποίο αναπτύχθηκε ειδικά ώστε να ελεγχθεί η επίδραση του κειμενικού γένους και του κειμενικού θέματος σε αυτές. Τα αποτελέσματα της ανάλυσης έδειξαν ότι πολλές από αυτές τις μεταβλητές, ενώ θεωρητικά σχετίζονται μόνο με το συγγραφικό ύφος, στην ουσία συσχετίζονται έμμεσα και με το θέμα και με το γένος του κειμένου. Ορισμένες μάλιστα από αυτές παρουσίαζαν συσχέτιση μόνο με το θέμα ή/και το κειμενικό γένος και δεν σχετίζονταν καθόλου με το συγγραφικό ύφος. Η χρήση τέτοιων μεταβλητών σε ένα ΗΣΚ που δεν έχει αναπτυχθεί λαμβάνοντας υπόψη την κατανομή των θεμάτων, κειμενικών γενών και μέσων σε αυτό μπορεί πολύ εύκολα να λειτουργήσει παραπλανητικά ως προς την ακρίβεια της κατηγοριοποίησης που θα επιχειρηθεί. Σε μια τέτοια υποθετική περίπτωση ο ερευνητής μπορεί να πετύχει υψηλά ποσοστά αναγνώρισης του συγγραφέα, αλλά στην ουσία να έχει πετύχει υψηλά ποσοστά αναγνώρισης του κειμενικού θέματος, αφού οι υφομετρικές μεταβλητές που χρησιμοποίησε συσχετίζονται και με το κειμενικό θέμα. Σε κάθε περίπτωση ΗΣΚ που δεν έχουν ελεγχθεί για την επίδραση των κειμενικών μεταδεδομένων στην

[30] Rudman, Joseph. "The State of Authorship Attribution Studies: Some Problems and Solutions." *Computers and the Humanities* 31.4 (1997): 351-65.

[31] Mikros, George K., and Eleni K. Argiri. "Investigating Topic Influence in Authorship Attribution." Benno Stein, Moshe Koppel and Efstathios Stamatatos eds. *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*. CEUR, 2007. 29-35.

επιχειρούμενη κατηγοριοποίηση επιδρούν αρνητικά στη συνολική αξιοπιστία των πειραματικών αποτελεσμάτων.

Για τις ανάγκες της παρούσας έρευνας αναπτύξαμε ένα ΗΣΚ στο οποίο ελέγχθηκε ταυτόχρονα το φύλο του συγγραφέα, το κειμενικό θέμα, το κειμενικό γένος και το κειμενικό μέσο. Ειδικότερα, οι αρχές σύστασης του ΗΣΚ που χρησιμοποιήσαμε είναι οι ακόλουθες:

- Όλα τα κείμενα θα πρέπει να έχουν δημοσιευθεί στην ίδια εφημερίδα (Ελευθεροτυπία) μέσα σε σύντομο χρονικό διάστημα (1 έτος).
- Ίσος αριθμός κειμένων (50) από κάθε συγγραφέα και ίσος αριθμός κειμένων ανδρών και γυναικών.
- Κάθε κείμενο από άνδρα συγγραφέα θα πρέπει να ταιριάζει πλήρως ως προς το θέμα και το γένος με κάθε κείμενο από γυναίκα συγγραφέα.
- Τα κείμενα θα πρέπει να ανήκουν σε πολλά και διαφορετικά μεταξύ του θέματα και γένη έτσι ώστε η γλωσσική χρήση που θα ερευνηθεί να καλύπτει ένα μεγάλο κοινωνιοπραγματολογικό εύρος.

Το ΗΣΚ που συλλέχθηκε περιλαμβάνει 700 κείμενα ισομερώς κατανεμημένα σε 7 άνδρες και 7 γυναίκες συγγραφείς. Αν και υπάρχουν μικρές ποσοτικές διαφορές μεταξύ συγκεκριμένων θεμάτων και γενών, το συγκεκριμένο ΗΣΚ θα πρέπει να θεωρείται ισορροπημένο ως προς τα κειμενικά μεταδεδομένα που ελέγχθηκαν. Το μέγεθος του ΗΣΚ σε αριθμό κειμένων και αριθμό λέξεων εμφανίζεται στον παρακάτω πίνακα (Πίνακας 1):

Πίνακας 1: Ποσοτική σύσταση του ΗΣΚ

	Θέμα / **Γένος**	**Επιστήμες**		**Κοινωνία**		**Οικονομία**		**Τέχνη**		**Σύνολο**	
		Κείμενα	Λέξεις	Κείμενα	Λέξεις	Κείμενα	Λέξεις	Κείμενα	Λέξεις	N	Λέξεις
Γυναίκες	Γνώμη	7	3169	84	60489	14	5748	17	14568	**122**	**83974**
	Είδηση	11	6308	111	77811	31	16982	15	10865	**168**	**111966**
	Συζήτηση	8	4999	33	23071	2	1646	17	21710	**60**	**51426**
	Υποσύνολο	*26*	*14476*	*228*	*161371*	*47*	*24376*	*49*	*47143*	***350***	***247366***
Άνδρες	Γνώμη	8	3847	88	60283	14	8453	17	11301	**127**	**83884**
	Είδηση	17	8353	117	68793	32	20471	16	11023	**182**	**108640**
	Συζήτηση			22	20595	2	1736	17	17218	**41**	**39549**
	Υποσύνολο	*25*	*12200*	*227*	*149671*	*48*	*30660*	*50*	*39542*	***350***	***232073***
	Σύνολο	**51**	**26676**	**455**	**311042**	**95**	**55036**	**99**	**86685**	**700**	**479439**

Το συγκεκριμένο ΗΚΣ έχει ποιοτικά και ποσοτικά χαρακτηριστικά που δυσκολεύουν σημαντικά τις τεχνικές υφομετρικής ανάλυσης. Είναι ομοιογενές ως προς τα κειμενικά μεταδεδομένα και επιπλέον περιέχει σημαντικό αριθμό κειμένων μικρού μεγέθους γεγονός που δυσκολεύει σημαντικά τις υφομετρικές τεχνικές. Πιο συγκεκριμένα το 84% των κειμένων είναι μικρότερα των 1000 λέξεων. Αυτό περιορίζει σημαντικά τη στατιστική κανονικότητα που εμφανίζουν οι υφομετρικές

μεταβλητές αφού οι περισσότερες επιδεικνύουν αξιοποιήσιμα υφομετρικά μοτίβα σε μεγαλύτερου μεγέθους κείμενα[32].
Τα κείμενα ανακτήθηκαν από το δικτυακό τόπο της εφημερίδας «Ελευθεροτυπία» με τη βοήθεια του εργαλείου διαδικτυακής συλλογής ΗΣΚ «Μινώταυρος»[33]. Η ανάλυση των κειμένων σε βασικές λεξικές μονάδες (tokenization) και ο μορφολογικός χαρακτηρισμός τους σε Μέρη του Λόγου (Part of Speech tagging) έγινε με την πλατφόρμα επεξεργασίας φυσικής γλώσσας «Έλλογον»[34] που αναπτύχθηκε από το Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών του Ερευνητικού Κέντρου «Δημόκριτος». Οι μετρήσεις σε διάφορες υφομετρικές μεταβλητές πραγματοποιήθηκαν με το ειδικό λογισμικό διαχείρισης και μετρήσεων ΗΣΚ «Corpus Manager»[35] καθώς και εξειδικευμένο λογισμικό που αναπτύχθηκε σε γλώσσα PERL.

4.2 Υφομετρικές μεταβλητές

Τα γλωσσικά χαρακτηριστικά που επιλέξαμε να μετρήσουμε ανήκουν στην ευρύτερη κατηγορία των υφομετρικών μεταβλητών και χρησιμοποιούνται ευρέως στην ανάλυση του συγγραφικού ύφους εδώ και 50 χρόνια. Η επιλογή των συγκεκριμένων μεταβλητών έγινε με γνώμονα τη σημασιολογική τους ουδετερότητα, τη δυνατότητα αυτοματοποιημένης και αξιόπιστης μέτρησής τους, καθώς και την υψηλή συχνότητα εμφάνισής τους έτσι ώστε η χρήση τους να μην είναι υποκείμενη σε συνειδητή τροποποίηση από μεριάς του συγγραφέα.
Ειδικότερα, μετρήσαμε έξι ευρύτερα σύνολα υφομετρικών χαρακτηριστικών που περιέχουν τόσο λεξιλογικές όσο και υπο-λεξικές μονάδες. Κάθε ένα σύνολο από αυτά συγκεντρώνει έναν αριθμό μεταβλητών που λειτουργούν συμπληρωματικά μεταξύ τους και ως σύνολο προσεγγίζει μια ευρύτερη αφηρημένη κειμενική ιδιότητα. Έτσι για παράδειγμα η κειμενική ιδιότητα *Γλωσσικός «πλούτος»* προσεγγίζεται από τις επιμέρους υφομετρικές μεταβλητές: Yule's K, Λεξιλογική Πυκνότητα, Ποσοστό των Άπαξ και των Δις-Λεγόμενα, Εντροπία, Ποσοστό λέξεων του κειμένου που δεν ανήκει στις 5000 και 10000 συχνότερες λέξεις της Νέας Ελληνικής. Κάθε μία από αυτές τις μεταβλητές επιχειρεί να ποσοτικοποιήσει μια συγκεκριμένη ιδιότητα του γλωσσικού «πλούτου» και λειτουργεί συμπληρωματικά ως προς τις υπόλοιπες. Με τον τρόπο αυτό επιχειρούμε μια πολυπαραγοντική προσέγγιση σε κειμενικές ιδιότητες οι οποίες είναι εξαιρετικά αφηρημένες και πολύπλευρες. Η λίστα που παρουσιάζεται εδώ δεν είναι εξαντλητική ωστόσο περιέχει την πλειονότητα των γλωσσικών μεταβλητών που κατά καιρούς έχουν χρησιμοποιηθεί στις περισσότερες υφομετρικές μελέτες. Τα υφομετρικά χαρακτηριστικά που μετρήθηκαν είναι τα ακόλουθα:

- *Γλωσσικός «Πλούτος»*
    - Yule's K: Υπολογίζεται ο δείκτης Characteristic K ο οποίος αποτελεί τον πιο αξιόπιστο δείκτη λεξιλογικού «πλούτου» αναφορικά με την

[32] Ledger, Gerard, and Thomas Merriam. "Shakespeare, Fletcher, and the Two Noble Kinsmen." *Literary and Linguistic Computing* 9 (1994): 235-48, Baillie, D. W. "Authorship Attribution in Jacobean Dramatic Texts." *Computers in the Humanities*. Ed. J. L. Mitchell. Edinburgh: Edinburgh University Press, 1974.
[33] Koutsis, Ilias, et al. "Minotavros: A Tool for the Semi-Automated Creation of Large Corpora from the Web." *Proceedings from the Corpus Linguistics Conference Series*. 2005.
[34] Petasis, Georgios, et al. "Ellogon: A New Text Engineering Platform." *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*. 2002. 72-78.
[35] Kouklakis, George, et al. "Corpus Manager: A Tool for Multilingual Corpus Analysis." Davies Matthew, et al. eds. *Proceedings of the Corpus Linguistics Conference (CL2007)*. 2007.

ευρωστία του (robustness) σε διαφορετικά κειμενικά μεγέθη. Ο υπολογισμός του ακολουθεί τον τύπο των Tweedie & Baayen[36]. Όσο πιο μικρός είναι ο δείκτης τόσο μεγαλύτερη λεξιλογική ποικιλία παρουσιάζει το κείμενο

- Λεξιλογική Πυκνότητα (Lexical Density): Ο λόγος του ποσοστού των λέξεων «περιεχομένου» (content words) προς το ποσοστό των λειτουργικών λέξεων (function words). Ο όρος «Λεξιλογική Πυκνότητα» χρησιμοποιείται στη θεωρία της λειτουργικής γραμματικής του Halliday με ελαφρά διαφορετικό τρόπο υπολογισμού. Εδώ ακολουθούμε την έκδοση που χρησιμοποιείται αρκετά συχνά σε μελέτες εντοπισμού αγνώστου συγγραφέα και απαντάται συχνά και με τον όρο «Λειτουργική Πυκνότητα» (Functional Density)[37]. Όσο μεγαλύτερος είναι ο δείκτης τόσο μεγαλύτερο τμήμα του κειμένου αποτελείται από λέξεις περιεχομένου.
- % Άπαξ και %Δις –λεγόμενα: Το ποσοστό των λέξεων του κειμένου που αποτελείται από άπαξ λεγόμενα (λέξεις με συχνότητα εμφάνισης 1 και 2 στο κείμενο).
- Δις / Άπαξ λεγόμενα: Ο λόγος των δις προς τα άπαξ λεγόμενα. Ο συγκεκριμένος λόγος έχει προταθεί ως ενδεικτικός του συγγραφικού ύφους[38].
- Εντροπία (Entropy): Υπολογίζεται η εντροπία του κάθε κειμένου, ο βαθμός δηλαδή της οργάνωσης και της προβλεψιμότητας των λεξικών συχνοτήτων. Ο υπολογισμός της ακολουθεί τον τύπο του Oakes[39]. Όσο μεγαλύτερη είναι η τιμή της εντροπίας τόσο λιγότερο αναμενόμενες, με τη στατιστική σημασία του όρου, είναι οι λέξεις που προκύπτουν στο κείμενο.
- Σχετική Εντροπία (Relative Entropy): Ο λόγος της θεωρητικά μέγιστης εντροπίας ενός κειμένου με την παρατηρηθείσα εντροπία. Μέγιστη εντροπία παρουσιάζει ένα κείμενο που κάθε λέξη που θα περιελάμβανε θα εμφανιζόταν 1 φορά και άρα στο κείμενο αυτό όλες οι λέξεις θα ήταν άπαξ λεγόμενα. Όσο μεγαλύτερη είναι η Σχετική Εντροπία, τόσο λιγότερο τυποποιημένο είναι το κείμενο, και άρα περισσότερο «πλούσιο» λεξιλογικά.
- Σπανιότητα Λεξιλογίου: Ποσοστό των λέξεων του κειμένου που δεν ανήκει στις 5000 και τις 10000 πιο συχνές λέξεις της Νέας Ελληνικής όπως αυτές έχουν υπολογιστεί στο ΗΣΚ «Εθνικός Θησαυρός Ελληνικής Γλώσσας»[40] του «Ινστιτούτου Επεξεργασίας του Λόγου».

- *Μήκος λέξης (Word Length)*
  - Μέσο Μήκος Λέξης (Average Word Length): Υπολογίζεται για κάθε κείμενο ο μέσος όρος του μήκους της λέξης μετρημένος σε χαρακτήρες.

---

[36] Tweedie, Fiona J., and Harald R. Baayen. "How Variable May a Constant Be? Measures of Lexical Richness in Perspective." *Computers and the Humanities* 32.5 (1998): 323-52.

[37] Miranda, García Antonio, and Martín Javier Calle. "Function Words in Authorship Attribution Studies." *Literary and Linguistic Computing* 22.1 (2007): 49-66.

[38] Hoover, David. "Another Perspective on Vocabulary Richness." *Computers and the Humanities* 37 (2003): 151-78.

[39] Oakes, Michael P. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1998.

[40] Hatzigeorgiu, Nick, et al. "Design and Implementation of the Online Ilsp Greek Corpus." Maria Gavrilidou, et al. eds. *Proceedings of the second International Conference on Language Resources and Evaluation (LREC 2000)*. ELRA, 2000. 1737-42.

- Φάσμα του Μήκους Λέξης (Word Length Spectrum): Η συχνότητα των λέξεων με 1, 2, 3 … 14 γράμματα κανονικοποιημένη σε δείγμα 1000 λέξεων.

- *Μήκος πρότασης (Sentence Length)*
  - Μέσο Μήκος Πρότασης (Average Sentence Length): Υπολογίζεται για κάθε κείμενο ο μέσος όρος του μήκους της πρότασης μετρημένος σε λέξεις.
  - % μεγάλων προτάσεων: Το ποσοστό των μεγάλων προτάσεων ανά κείμενο (>18 λέξεις).
- *Συχνότητες γραμμάτων (Character frequencies)*
  - Η συχνότητα κάθε γράμματος κανονικοποιημένη σε μέγεθος κειμένου 1000 λέξεων.
- *Συχνότητες των Μερών του Λόγου (Part of Speech frequencies)*
  - Η συχνότητα κάθε μέρους του λόγου ανά κείμενο, όπως αυτή υπολογίστηκε από το μορφολογικό αναλυτή, κανονικοποιημένη σε μέγεθος κειμένου 1000 λέξεων.
- *Συχνές Λειτουργικές Λέξεις (Frequent Function Words)*
  - Η συχνότητα εμφάνισης των 50 πιο συχνών λειτουργικών λέξεων ανά κείμενο κανονικοποιημένη σε μέγεθος κειμένου 1000 λέξεων.

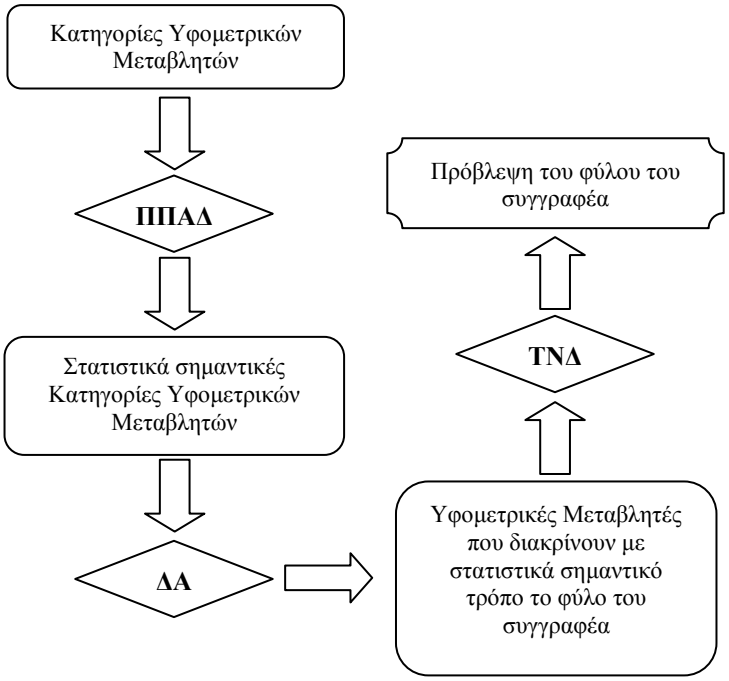
5 Στατιστική ανάλυση

Η ανάλυση της επίδρασης του φύλου του συγγραφέα στις υφομετρικές μεταβλητές που αναφέρθηκαν στην προηγούμενη ενότητα (4.2) απαιτεί πολυπαραγοντικές στατιστικές μεθόδους. Για να μπορέσουμε να αξιολογήσουμε ταυτόχρονα τη συμβολή όλων των υφομετρικών μεταβλητών θα χρησιμοποιήσουμε Πολυπαραγοντική Ανάλυση Διακύμανσης – ΠΠΑΔ (Multiple Analysis of Variance – MANOVA) με εξαρτημένες μεταβλητές τα υφομετρικά χαρακτηριστικά που μετρήθηκαν και ανεξάρτητη κατηγορική μεταβλητή το φύλο του συγγραφέα. Η ΠΠΑΔ είναι μαθηματική επέκταση της μονοπαραγοντικής Ανάλυσης της Διακύμανσης (Analysis of Variance – ANOVA) έτσι ώστε να μπορεί να αναλύσει την επίδραση δύο ή περισσότερων ανεξάρτητων κατηγορικών μεταβλητών σε δύο ή περισσότερες αριθμητικές εξαρτημένες μεταβλητές. Αν και το συγκεκριμένο πρόβλημα μπορεί να αντιμετωπιστεί με μια σειρά από μονοπαραγοντικά τεστ, το συνολικό λάθος α-επιπέδου (a-level error) θα αυξηθεί σημαντικά. Έτσι η πιθανότητα να διαπράξουμε λάθος Τύπου Ι, δηλαδή η πιθανότητα να απορρίψουμε εσφαλμένα τη μηδενική υπόθεση αυξάνεται σημαντικά. Η ΠΠΑΔ ελέγχει το λάθος Τύπου Ι και προσφέρει ένα συνολικό τεστ σημαντικότητας που λαμβάνει υπόψη του την επίδραση των ανεξάρτητων μεταβλητών σε όλες τις εξαρτημένες μεταβλητές ταυτόχρονα[41].
Η χρήση της ΠΠΑΔ ενδείκνυται[42] όταν οι εξαρτημένες μεταβλητές συσχετίζονται εννοιολογικά, όπως στην περίπτωση των υφομετρικών χαρακτηριστικών που χρησιμοποιούμε, και υπάρχει μέτρια συσχέτιση μεταξύ τους. Η ΠΠΑΔ αξιοποιεί την κοινή πληροφορία που μοιράζονται εννοιολογικά συσχετιζόμενες μεταβλητές και ελέγχει την επίδραση της ανεξάρτητης μεταβλητής (φύλο του συγγραφέα) με

[41] Weinfurth, Kevin P. "Multivariate Analysis of Variance." *Reading and Understanding Multivariate Statistics*. Eds. Laurence G. Grim and Paul R. Yarnold. Washington, DC: American Psychological Association, 1995. 245-76.
[42] Huberty, Carl J., and John D. Morris. "Multivariate Analysis Versus Multiple Univariate Analyses." *Psychological Bulletin* 105.2 (1989): 302-08.

πολυπαραγοντικό τρόπο (εξετάζοντας ταυτόχρονα όλες τις εξαρτημένες υφομετρικές μεταβλητές).
Η ΠΠΑΔ παράγει μια γραμμική εξίσωση (linear composite) των εξαρτημένων μεταβλητών που μεγιστοποιεί τη διακρίση των κατηγοριών της ανεξάρτητης μεταβλητής. Ένα μη στατιστικά σημαντικό αποτέλεσμα σημαίνει ότι η συγκεκριμένη ομάδα εξαρτημένων μεταβλητών δεν παρουσιάζει διαφοροποίηση των μέσων όρων μεταξύ των κατηγοριών της ανεξάρτητης μεταβλητής όταν αυτοί εξετάζονται ταυτόχρονα. Αντίθετα, μια στατιστικά σημαντική ΠΠΑΔ υποδεικνύει ότι τουλάχιστον μία από τις εξαρτημένες μεταβλητές διαφέρει σημαντικά μεταξύ των κατηγοριών της ανεξάρτητης μεταβλητής. Στη σχετική βιβλιογραφία[43] οι περισσότεροι ερευνητές μετά από μία στατιστικά σημαντική ΠΠΑΔ πραγματοποιούν πολλαπλά μονοπαραγοντικά τεστ (t τεστ στην περίπτωσή μας) προσαρμόζοντας το λάθος του α-επιπέδου στον αριθμό των τεστ (διόρθωση Bonferroni), έτσι ώστε να εντοπίσουν ποια/ες εξαρτημένη/ες μεταβλητή/ες διαφέρουν στατιστικά σημαντικά μεταξύ των κατηγοριών της ανεξάρτητης μεταβλητής. Αυτή η μεθοδολογία ωστόσο έχει δεχθεί κριτική[44] επειδή συγχέει τα μονοπαραγοντικά με τα εγγενώς πολυπαραγοντικά ερωτήματα, όπως είναι στην περίπτωση μας η διερεύνηση της επίδρασης του φύλου του συγγραφέα στο υφομετρικό προφίλ του κειμένου. Σε ένα τέτοιο ερευνητικό ερώτημα μας απασχολεί όχι ο μεμονωμένος εντοπισμός υφομετρικών χαρακτηριστικών που σχετίζονται με το φύλο του συγγραφέα, αλλά ο προσδιορισμός της βέλτιστης ακολουθίας των υφομετρικών μεταβλητών που αλληλεπιδρώντας μεταξύ τους μεγιστοποιούν τη διάκριση των κειμένων που γράφτηκαν από άνδρες και γυναίκες συγγραφείς. Το συγκεκριμένο ερευνητικό ζητούμενο ικανοποιείται από την επιλογή της Διακριτικής Ανάλυσης - ΔΑ (Discriminant Function Analysis - DFA) ως το επόμενο βήμα μιας στατιστικά σημαντικής ΠΠΑΔ. Η ΔΑ επιτρέπει στον ερευνητή να διερευνήσει σε βάθος τη γραμμική εξίσωση που προκύπτει από την ΠΠΑΔ και να αναλύσει τη δομή της καθώς και τους συντελεστές βαρύτητας (weights) κάθε εξαρτημένης μεταβλητής που συμμετέχει σε αυτήν[45]. Το σημαντικότερο βέβαια πλεονέκτημα είναι ότι διατηρείται ο πολυπαραγοντικός χαρακτήρας της ανάλυσης αφού η ΔΑ είναι στην ουσία ανεστραμμένη ΠΠΑΔ. Έτσι η εστίαση της ανάλυσης παραμένει στην αξιολόγηση της γραμμικής συνάρτησης όλων των εξαρτημένων μεταβλητών και όχι στη διερεύνηση της επίδρασης που έχει η κάθε μία από αυτές ξεχωριστά στη διάκριση των κατηγοριών της ανεξάρτητης μεταβλητής.
Η μεθοδολογία χρήσης των στατιστικών αναλύσεων που θα χρησιμοποιηθεί στην παρούσα εργασία απεικονίζεται στο παρακάτω διάγραμμα ροής (Διάγραμμα 1):

[43] Hair Jr, Joseph F., et al. *Multivariate Data Analysis*. 5th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1995, Stevens, James P. *Applied Multivariate Statistics for the Social Sciences*. 4rth ed. Hillsdale, NJ: Erlbaum, 2002.
[44] Bray, James H., and Scott E. Maxwell. "Analyzing and Interpreting Significant Manovas." *Review of Educational Research* 52.3 (1982): 340-67, Huberty, and Morris. "Multivariate Analysis Versus Multiple Univariate Analyses."
[45] Meyers, Lawrence S., Glenn Gamst, and A. J. Guarino. *Applied Multivariate Research. Design and Interpretation*. Thousand Oaks, CA: Sage, 2006.

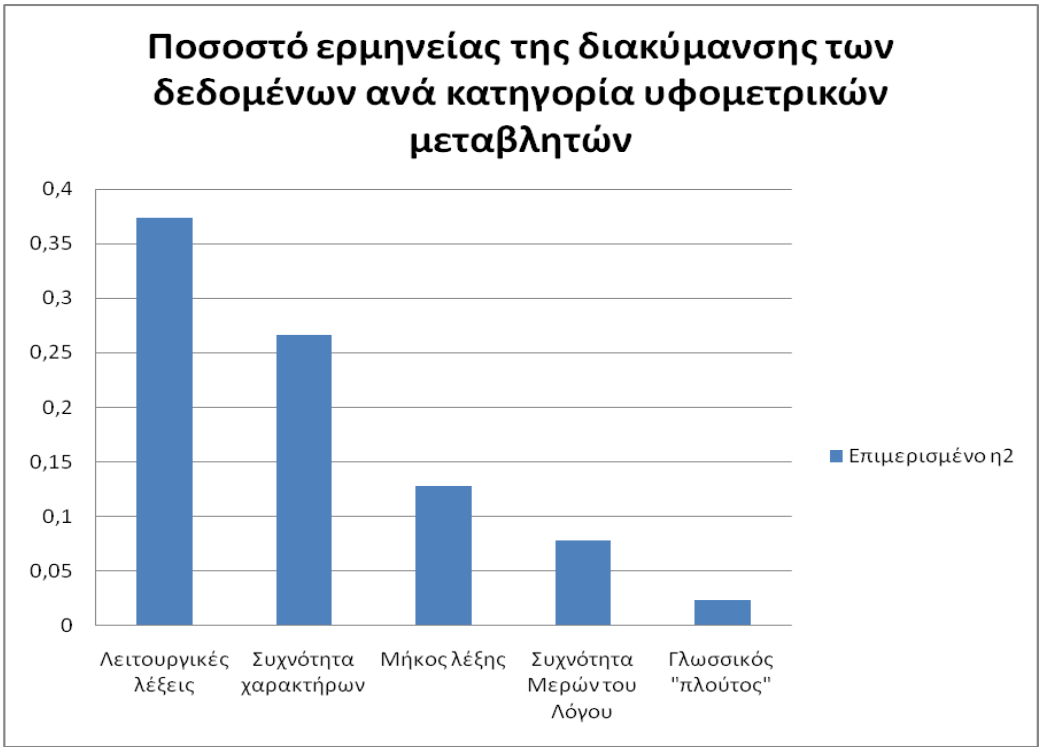


Διάγραμμα 1: Διάγραμμα ροής των στατιστικών αναλύσεων που θα χρησιμοποιηθούν στην παρούσα έρευνα

Οι υφομετρικές μεταβλητές που μετρήθηκαν θα οργανωθούν σε κατηγορίες και κάθε μία κατηγορία θα αξιολογηθεί μέσω ΠΠΑΔ για το αν διακρίνει το φύλο του συγγραφέα με στατιστικά σημαντικό τρόπο. Στη συνέχεια οι κατηγορίες μεταβλητών που εμφάνισαν στατιστική σημαντικότητα θα χρησιμοποιηθούν σε ΔΑ με εξαρτημένη μεταβλητή τα υφομετρικά χαρακτηριστικά της κάθε κατηγορίας και ανεξάρτητη μεταβλητή το φύλο του συγγραφέα. Σε κάθε ΔΑ όποιες υφομετρικές μεταβλητές συμβάλλουν με στατιστικά σημαντικό τρόπο στη διάκριση του φύλου του συγγραφέα θα αποτελούν τμήμα της ενοποιημένης ομάδας υφομετρικών χαρακτηριστικών που θα εκπαιδεύσουν το ΤΝΔ. Στην τελευταία φάση το ΤΝΔ εκπαιδεύεται με τις υφομετρικές μεταβλητές που κατά την ανάλυση με ΔΑ βγήκαν στατιστικά σημαντικοί ενδείκτες του φύλου του συγγραφέα.

6 Η επιλογή των υφομετρικών μεταβλητών που διακρίνουν το φύλο του συγγραφέα

Στα δεδομένα μας χρησιμοποιήσαμε ξεχωριστές ΠΠΑΔ για κάθε μία από τις έξι ομάδες υφομετρικών μεταβλητών με ανεξάρτητη μεταβλητή το φύλο του συγγραφέα. Υπολογίστηκε ο πολυπαραγοντικός δείκτης Hotelling T^2 ο οποίος αντιστοιχεί στο μονοπαραγοντικό δείκτη t. Επιπλέον, για κάθε μία ομάδα μεταβλητών υπολογίστηκε ο επιμερισμένος η^2 (partial η^2) ο οποίος αντιστοιχεί στο ποσοστό της διακύμανσης (variance) που ερμηνεύεται από το συγκεκριμένο συνδυασμό των υφομετρικών μεταβλητών. Το ακόλουθο διάγραμμα (Διάγραμμα 2) απεικονίζει οπτικά τη σχετική κατάταξη των ομάδων των μεταβλητών σχετικά με την ερμηνευτική τους δύναμη στη διάκριση του φύλου του συγγραφέα.



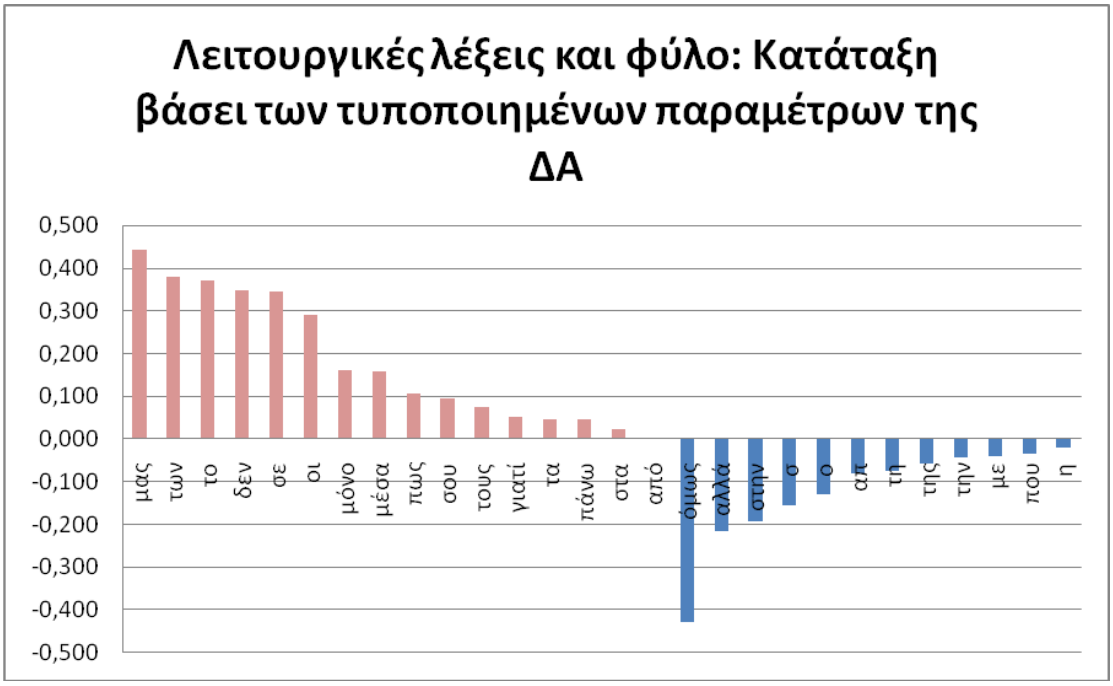
Διάγραμμα 2: Ιστόγραμμα της ερμηνευτικής δύναμης των διαφόρων υφομετρικών ομάδων σε σχέση με τη διάκριση του φύλου του συγγραφέα.

Στο παραπάνω διάγραμμα εμφανίζονται όλες οι υφομετρικές ομάδες που αναφέρθηκαν στην ενότητα 4.2 εκτός του *Μήκους Πρότασης* του οποίου η ανάλυση με ΠΠΑΔ ήταν μη στατιστικά σημαντική. Η ομάδα υφομετρικών μεταβλητών με τη μεγαλύτερη ερμηνευτική δύναμη όσον αφορά το φύλο του συγγραφέα είναι η *Συχνές Λειτουργικές Λέξεις* (37%) ακολουθούμενες από τις *Συχνότητητες Γραμμάτων* (27%), το *Μήκος Λέξης* (13%). Μικρή (<10%) αλλά στατιστικά σημαντική επίδραση εμφανίζουν οι κατηγορίες «*Συχνότητες των Μερών του Λόγου*» και *Γλωσσικός «Πλούτος»*.

Για κάθε μία από τις υφομετρικές ομάδες στις οποίος ο Hotelling T^2 βγήκε στατιστικά σημαντικός ακολούθησε μια ΔΑ με εξαρτημένη μεταβλητή το φύλο του συγγραφέα και ανεξάρτητες μεταβλητές τα υφομετρικά χαρακτηριστικά της συγκεκριμένης ομάδας. Με τον τρόπο αυτό μπορέσαμε να αναλύσουμε την διακριτική εξίσωση που παράγει η ΔΑ και να διερευνήσουμε την επίδραση κάθε μίας υφομετρικής μεταβλητής στη κατηγοριοποίηση του κειμένου βάσει του φύλου του συγγραφέα του.

6.1 Συχνότερες Λειτουργικές Λέξεις

Η ΔΑ με ανεξάρτητες μεταβλητές τις 50 *Συχνότερες Λειτουργικές Λέξεις* και εξαρτημένη μεταβλητή το φύλο του συγγραφέα προσδιόρισε 28 λειτουργικές λέξεις η συχνότητα χρήσης των οποίων διακρίνει με στατιστικά σημαντικό τρόπο το φύλο του συγγραφέα ενός κειμένου. Η διακριτική δύναμη της κάθε μεταβλητής υπολογίστηκε με βάση την τυποποιημένη παράμετρο (standardized coefficient) της ΔΑ. Στο παρακάτω διάγραμμα (Διάγραμμα 3) εμφανίζονται οι λειτουργικές λέξεις κατά φθίνουσα σειρά σπουδαιότητας όσον αφορά τη διακριτική τους λειτουργία ως προς το φύλο του συγγραφέα. Οι θετικές τιμές των τυποποιημένων παραμέτρων σχετίζονται με λειτουργικές λέξεις που προτιμούν οι γυναίκες ενώ οι αρνητικές τιμές των τυποποιημένων παραμέτρων σχετίζονται με λειτουργικές λέξεις που προτιμούν οι άνδρες συγγραφείς.



Διάγραμμα 3: Η επίδραση των λειτουργικών λέξεων στη διάκριση του φύλου του συγγραφέα.

Στο παραπάνω διάγραμμα είναι εμφανές ότι οι άνδρες συγγραφείς χρησιμοποιούν (με φθίνουσα σειρά σημαντικότητας) τις λέξεις *όμως, αλλά, στην, σ', ο, απ', τη, της, την, με, που, η*, ενώ οι γυναίκες συγγραφείς προτιμούν τις λέξεις *μας, των, το, δεν, σε (προσ. αντων.), οι, μόνο, μέσα, πως, σου, τους, γιατί, τα, πάνω, στα, από*.

Μεταξύ των λειτουργικών λέξεων που χαρακτηρίζουν τα κείμενα που παράγουν άνδρες μπορούμε να διακρίνουμε δύο επιμέρους κατηγορίες: α) αντιθετικοί σύνδεσμοι (αλλά, όμως) και β) συγκεκομμένοι τύποι προθέσεων (σ' απ'). Η πρώτη κατηγορία χαρακτηρίζει τη συντακτική δομή των κειμένων και παλαιότερες έρευνες την έχουν συσχετίσει με τη διάκριση του φύλου του συγγραφέα[46]. Η δεύτερη κατηγορία σχετίζεται με φαινόμενα κοινωνιογλωσσολογικής ποικιλίας και έχει επίσης συσχετιστεί με τον ανδρικό τρόπο γραφής.

Αντίθετα, στις λέξεις που χαρακτηρίζουν τις γυναίκες συγγραφείς μπορούμε να ξεχωρίσουμε την παρουσία των προσωπικών αντωνυμιών (μας, σε, σου). Η προτίμηση των προσωπικών αντωνυμιών στο γυναικείο λόγο έχει επιβεβαιωθεί και από άλλες έρευνες[47] και σχετίζεται με το γεγονός ότι ο γυναικείος λόγος χαρακτηρίζεται από διάθεση εμπλοκής του ομιλητή/συγγραφέα με τον ακροατή/αναγνώστη. Αυτή η εμπλοκή γλωσσικά κωδικοποιείται μέσα από τη χρήση προσωπικών αντωνυμιών αφού επιτρέπουν στο συγγραφέα να αναφερθεί άμεσα στον αναγνώστη και να τον κάνει κοινωνό των απόψεών του.

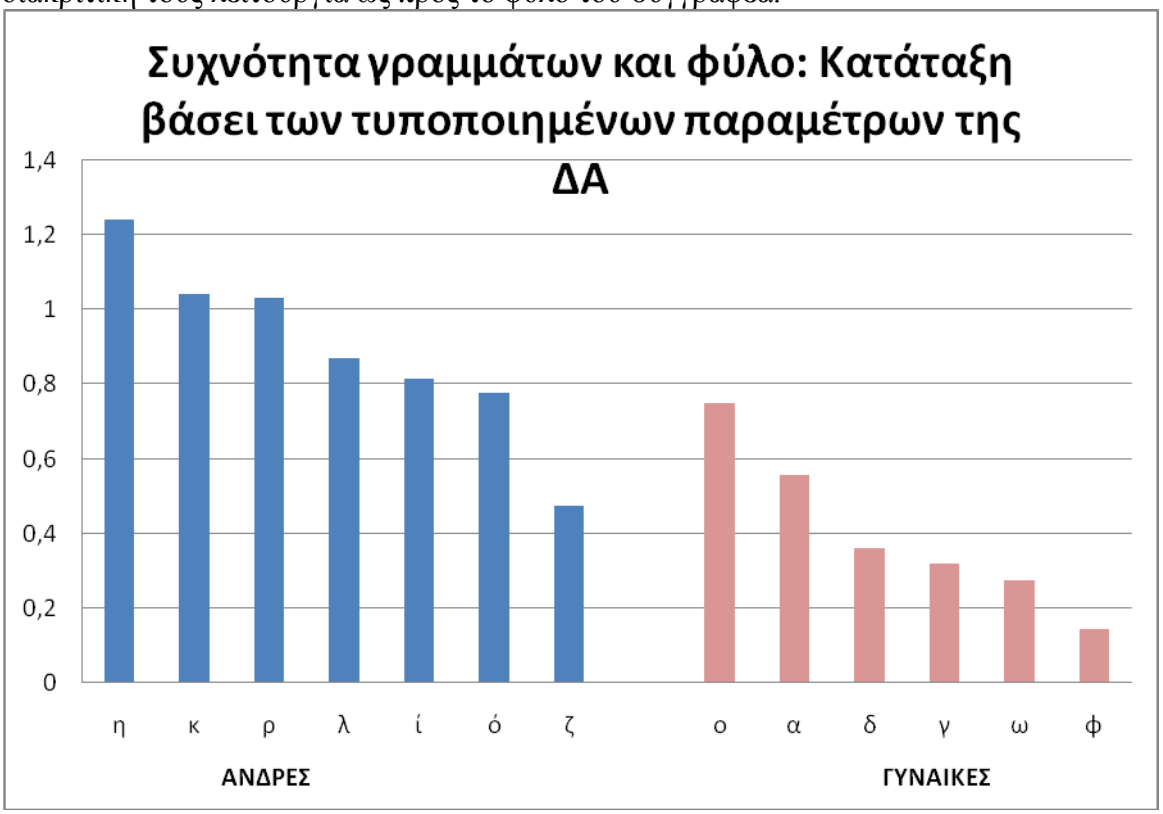
---

[46] Mulac, Anthony, James J. Bradac, and Susan Karol Mann. "Male/Female Language Differences and Attributional Consequences in Children's Television." *Human Communication Research* 11.4 (1985): 481-506, Mulac, Anthony, Lisa B. Studley, and Sheridan Blau. "The Gender-Linked Language Effect in Primary and Secondary Students' Impromptu Essays." *Sex roles* 23.9-10 (1990): 439-70.

[47] Holmes, Janet. "Hedges and Boosters in Women's and Men's Speech." *Language and Communication* 10.3 (1990): 185-205, Preisler, Bent. *Linguistic Sex Roles in Conversation: Social Variation in the Expression of Tentativeness in English*. Berlin: Mouton de Gruyter, 1986, Rayson, Paul, Geoffrey Leech, and Mary Hodges. "Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus." *International Journal of Corpus Linguistics* 2.1 (1997): 133-52, Argamon, Shlomo, et al. "Mining the Blogosphere: Age, Gender and the Varieties of Self–Expression." *First Monday* 12.9 (2007). 19/06/2009 <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2003/1878>.ττη

6.2 Συχνότητες γραμμάτων

Η ΔΑ με τις συχνότητες των γραμμάτων ως ανεξάρτητες μεταβλητές και το φύλο του συγγραφέα ως εξαρτημένη μεταβλητή ανέδειξε 12 γράμματα τα οποία διαφοροποιούν με στατιστικά σημαντικά τρόπο τους άνδρες από τις γυναίκες συγγραφείς. Η διακριτική δύναμη της κάθε μεταβλητής υπολογίστηκε με βάση την τυποποιημένη παράμετρο (standardized coefficient) της ΔΑ. Στο παρακάτω διάγραμμα (Διάγραμμα 4) εμφανίζονται τα γράμματα κατά φθίνουσα σειρά σπουδαιότητας όσον αφορά τη διακριτική τους λειτουργία ως προς το φύλο του συγγραφέα.



Διάγραμμα 4: Η επίδραση της συχνότητας των γραμμάτων στη διάκριση του φύλου του συγγραφέα.

Στο παραπάνω διάγραμμα διαπιστώνουμε ότι οι άνδρες συγγραφείς χρησιμοποιούν συχνότερα (με φθίνουσα σειρά σπουδαιότητας ως προς τη διάκριση του φύλου του συγγραφέα) τα γράμματα *η, κ, ρ, λ, ί, ό, ζ*. Αντίθετα, τα γράμματα που διακρίνουν τα κείμενα που έχουν γράψει γυναίκες συγγραφείς είναι τα *ο, α, δ, γ, ω, φ*.

6.3 Μήκος λέξης

Η ΔΑ με τις μεταβλητές του Μήκους λέξης ως ανεξάρτητες μεταβλητές και το φύλο του συγγραφέα ως εξαρτημένη μεταβλητή έδειξε ότι επτά μεταβλητές διαφοροποιούν με στατιστικά σημαντικά τρόπο τους άνδρες από τις γυναίκες συγγραφείς. Στο παρακάτω διάγραμμα (Διάγραμμα 5) εμφανίζονται οι μεταβλητές του Μήκους λέξης κατά φθίνουσα σειρά σπουδαιότητας όσον αφορά τη διακριτική τους λειτουργία ως προς το φύλο του συγγραφέα.

Διάγραμμα 5: Η επίδραση του Μήκους λέξης στη διάκριση του φύλου του συγγραφέα.

Στο παραπάνω διάγραμμα βλέπουμε ότι οι γυναίκες συγγραφείς χρησιμοποιούν με στατιστικά σημαντική διαφοροποίηση τις λέξεις των 2, 3, 13 και 14 γραμμάτων, ενώ οι άνδρες συγγραφείς εμφανίζουν υψηλότερη χρήση στις λέξεις των 4, 8, 9 και 10 γραμμάτων. Η εξέταση των τυποποιημένων παραμέτρων της ΔΑ δείχνει ότι οι πιο χαρακτηριστικοί δείκτες για τον εντοπισμό γυναικείας γραφής είναι το ποσοστό των λέξεων με 2 και 3 γράμματα ακολουθούμενα από τα ποσοστά των λέξεων με 13 και 14 γράμματα. Αντίστοιχα, οι πιο χρήσιμοι δείκτες για τον εντοπισμό των κειμένων που γράφονται από άνδρες είναι τα ποσοστά των λέξεων με 8 και 4 γράμματα ακολουθούμενα από τα ποσοστά των λέξεων που έχουν 9 και 10 γράμματα.

Αυτά τα αποτελέσματα μας επιτρέπουν τη διαμόρφωση μιας συγκεκριμένης εικόνας γύρω από τη σχέση του μήκους της λέξης με το φύλο του συγγραφέα. Οι γυναίκες χρησιμοποιούν το κατώτερο και το ανώτερο όριο του φάσματος στο οποίο εκτείνεται το γλωσσικό μήκος. Χρησιμοποιούν μικρές λέξεις (2-3 γραμμάτων) που στην πλειονότητά τους ανήκουν στην ομάδα των λειτουργικών λέξεων. Χρησιμοποιούν επίσης πολυγράμματες λέξεις (13 και 14 γραμμάτων) που σχετίζονται αντιστρόφως ανάλογα με τη χρήση των λειτουργικών λέξεων. Η επίδραση των πολυγράμματων λέξεων στο γυναικείο υφομετρικό προφίλ είναι μικρή σε σχέση με την επίδραση των μικρών λέξεων (2 και 3 γραμμάτων), γεγονός που μας κάνει να υποθέτουμε ότι δεν είναι άμεσοι δείκτες του γυναικείου τρόπου γραφής αλλά απλά αντανακλούν ανάστροφα τη έντονη προτίμηση των γυναικών στις μικρές λέξεις (2 και 3 γραμμάτων).

Αυτή η υπόθεση υποστηρίζεται περαιτέρω από τη διερεύνηση των συσχετίσεων των ποσοστών των λέξεων με 2, 3, 13 και 14 γραμμάτων με τη Λειτουργική Πυκνότητα στα γυναικεία κείμενα. Το ποσοστό των λέξεων με 2 και 3 γράμματα εμφανίζει στατιστικά σημαντική αρνητική συσχέτιση με τη Λεξιλογική Πυκνότητα ($r_{2\gamma\rho}$= -0,354, $r_{3\gamma\rho}$= -0,385) γεγονός που σημαίνει ότι η αύξηση στη Λεξιλογική Πυκνότητα (δηλ. περισσότερες λέξεις «περιεχομένου») σχετίζεται αντιστρόφως ανάλογα με τα ποσοστά των μικρών λέξεων (2 και 3 γραμμάτων). Από την άλλη μεριά, τα ποσοστά

των μεγάλων λέξεων (13 και 14 γραμμάτων) εμφανίζουν μικρότερη αλλά στατιστικά σημαντική θετική γραμμική συσχέτιση με τη Λεξιλογική Πυκνότητα ($r_{13γρ}$= 0,134, $r_{14γρ}$= 0,144) γεγονός που σημαίνει ότι καθώς η Λεξιλογική Πυκνότητα αυξάνεται, μια αντίστοιχη αύξηση παρατηρείται και στα ποσοστά των μεγάλων λέξεων με μικρότερο όμως ρυθμό.

6.4 Συχνότητες των Μερών του Λόγου

Η ΔΑ με τις συχνότητες των Μερών του Λόγου ως ανεξάρτητες μεταβλητές και το φύλο του συγγραφέα ως εξαρτημένη μεταβλητή ανέδειξε τη χρήση των επιρρημάτων (Wilk's λ= 0,987, p=0,003) και τη χρήση των επιθέτων (Wilk's λ= 0,995, p=0,049) ως στατιστικά σημαντικές μεταβλητές διάκρισης των ανδρικών και των γυναικείων κειμένων. Ειδικότερα, οι άνδρες συγγραφείς χρησιμοποιούν αυξημένα ποσοστά επιρρημάτων (M= 8,2, SD= 1,9) σε σχέση με τις γυναίκες (M= 7,8, SD= 1,7) και οι γυναίκες συγγραφείς χρησιμοποιούν αυξημένα ποσοστά επιθέτων (M= 8,2, SD= 2,2) συγκριτικά με τους άνδρες συγγραφείς (M= 7,9, SD= 1,9).

Συγκρίνοντας τη σχετική επίδραση των δύο συγκεκριμένων Μερών του Λόγου στη διάκριση του φύλου του συγγραφέα, παρατηρούμε ότι η χρήση των επιθέτων παρουσιάζει ισχυρή σχέση με τη διακριτική συνάρτηση (συντελεστής συσχέτισης: 0,256, τυποποιημένη παράμετρος: 0,464), ενώ η χρήση των επιρρημάτων παρουσιάζει ελαφρώς πιο αδύναμη σχέση (συντελεστής συσχέτισης: -0,403, τυποποιημένη παράμετρος: -0,435).

6.5 Γλωσσικός «Πλούτος»

Η ΔΑ με τις μεταβλητές του Γλωσσικού «Πλούτου» ως ανεξάρτητες μεταβλητές και το φύλο του συγγραφέα ως εξαρτημένη μεταβλητή ανέδειξε το ποσοστό των λέξεων του κειμένου που ανήκει στις 5000 συχνότερες λέξεις (Wilk's λ= 0,980, p=0,001) και τη χρήση της Σχετικής εντροπίας (Wilk's λ= 0,991, p=0,01) ως στατιστικά σημαντικές μεταβλητές διάκρισης των ανδρικών και των γυναικείων κειμένων. Ειδικότερα οι άνδρες συγγραφείς χρησιμοποιούν μικρότερο ποσοστό λέξεων που δεν ανήκουν στις 5000 συχνότερες λέξεις (M= 0,25, SD= 0,05) σε σχέση με τις γυναίκες (M= 0,26, SD= 0,04). Επιπλέον τα κείμενα που γράφουν οι γυναίκες συγγραφείς παρουσιάζουν μικρότερη σχετική εντροπία (M= 82,95, SD= 2,98) σε σχέση με αυτά των ανδρών (M= 83,54, SD= 3,13).

Η εξέταση της σχέσης των μεταβλητών του Γλωσσικού «Πλούτου» με το φύλο του συγγραφέα μας αποκάλυψε μια περίπλοκη και ετερογενή εικόνα που είναι ενδεικτική της πολυπλοκότητας που παρουσιάζει η σχέση του φύλου με το υφομετρικό προφίλ του κειμένου. Η Σχετική εντροπία και το ποσοστό των λέξεων του κειμένου που δεν ανήκει στις 5000 συχνότερες λέξεις αν και θεωρητικά μετρούν την ίδια αφηρημένη κειμενική ιδιότητα, το λεξιλογικό «πλούτο», ωστόσο σχετίζονται αντίστροφα με το φύλο του συγγραφέα. Οι γυναίκες γράφουν κείμενα με λεξιλόγιο που δεν είναι κοινόχρηστο και περιλαμβάνει μεγαλύτερο ποσοστό σπάνιων λέξεων, ενώ τα κείμενα των ανδρών παρουσιάζουν μικρότερη λεξιλογική επαναληπτικότητα, αποφυγή λεξιλογικών μοτίβων και γενικά λιγότερα χαρακτηριστικά λεξιλογικής τυποποίησης.

7 Εκπαίδευση Τεχνητού Νευρωνικού Δικτύου στην πρόβλεψη του φύλου του συγγραφέα

7.1 Τεχνητά Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα – ΤΝΔ (Artificial Neural Network – ANN) αποτελούν ένα από ισχυρότερα προβλεπτικά εργαλεία τεχνητής νοημοσύνης. Έχουν χρησιμοποιηθεί με μεγάλη επιτυχία σε ποικίλα προβλήματα κατηγοριοποίησης και τα τελευταία χρόνια χρησιμοποιούνται συστηματικά και σε προβλήματα υφομετρικής κατηγοριοποίησης[48].

Αναπτύχθηκαν για πρώτη φορά στη δεκαετία του ’60 έχοντας ως πρότυπο το βιολογικό μοντέλο του εγκεφαλικού νευρώνα. Όπως ο νευρώνας έτσι και το ΤΝΔ έχει εισόδους (ανεξάρτητες μεταβλητές) και εξόδους (εξαρτημένη μεταβλητή).

Το ΤΝΔ είναι μια μαθηματική κατασκευή η οποία μπορεί να εκπαιδευθεί μέσα από μια σειρά παραδειγμάτων αιτίας και αιτιατού. Μέσα από αυτή την εκπαίδευση το ΤΝΔ μπορεί να συνδέσει μαθηματικά το σύνολο των δεδομένων εκπαίδευσης (είσοδοι) με τα δεδομένα της κατηγορίας εξόδου. Οι είσοδοι (inputs) και οι έξοδοι (outputs) είναι αριθμοί και επομένως μπορούμε να αναπαραστήσουμε το σύνολο των εισόδων ή των εξόδων ως διάνυσμα: $x = [x_1, x_2, \ldots x_m]$. Για να εκπαιδεύσουμε το ΤΝΔ θα χρειαστούμε κάποιον αριθμό *n* διανυσμάτων εισόδου $\{x_k\}_{k=1}^{n}$ με γνωστά δεδομένα εξόδου $\{y_k\}_{k=1}^{n}$.

Ένα ΤΝΔ μπορεί να περιγραφεί ως ένας αριθμός νευρώνων που έχει οργανωθεί σε επίπεδα (layers), τα οποία συνδέονται μεταξύ τους με συναπτικά βάρη (weights) μέσω συναρτήσεων μετάβασης (transfer functions)[49]. Η οργάνωση των νευρώνων καθορίζει τον αριθμό των βαρών και την επίδρασή τους στο τελικό εξαγόμενο. Για ένα πρόβλημα κατηγοριοποίησης όπως αυτό που μελετάμε εδώ το ΤΝΔ θα έχει *m* νευρώνες στο επίπεδο εισόδου (input layer) (ένα για κάθε ένα από τα *m* στοιχεία του διανύσματος εισόδου x), κάποιο αριθμό νευρώνων, n_h, στο κρυφό επίπεδο (hidden layer) και έναν νευρώνα επιπέδου εξόδου (output layer) που στην ιδανική περίπτωση θα παίρνει δύο τιμές, Άνδρας ή Γυναίκα αντιπροσωπεύοντας την απόφαση ανάμεσα σε δύο πιθανότητες. Δεδομένου ότι έχουμε ένα συναπτικό βάρος για κάθε σύνδεση μεταξύ των νευρώνων, αυτό μας δίνει συνολικά $n_h(m+1)$ συναπτικά βάρη. Οι τιμές που παίρνουν τα συναπτικά βάρη καθορίζουν τον τρόπο με τον οποίο το ΤΝΔ θα μετασχηματίσει τα δεδομένα εισόδου στα δεδομένα εξόδου και αναλογεί με τη διαφορετική ισχύ των νευρικών συνδέσεων που υφίστανται στον εγκέφαλο. Μαθηματικά αυτό δεν είναι τίποτα περισσότερο από το να επιλέξεις μια πολύπλοκη συνάρτηση f(x) έτσι ώστε να συνδέσεις τα εισαγόμενα με τα εξαγόμενα. Αυτή η

---

[48] Singh, Sameer, and Fiona Tweedie. "Neural Networks and Disputed Authorship: New Challenges." *Artificial Neural Networks*. IEE, 1995. 24-28, Matthews, Robert, and Thomas Merriam. "Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher." *Literary and Linguistic Computing* 8.4 (1993): 203-09, Merriam, Thomas, and Robert Matthews. "Neural Computation in Stylometry Ii: An Application to the Works of Shakespeare and Marlowe." *Literary and Linguistic Computing* 9.1 (1994): 1-6, Lowe, D., and Robert Matthews. "Shakespeare Vs. Fletcher: A Stylometric Analysis by Radial Basis Functions." *Computers and the Humanities* 29 (1995): 449-61, Tweedie, Fiona, Sameer Singh, and David I. Holmes. "Neural Network Applications in Stylometry: The Federalist Papers." *Computers and the Humanities* 30 (1996): 1-10, Tearle, Matt, Kye Taylor, and Howard B. Demuth. "An Algorithm for Automated Authorship Attribution Using Neural Networks." *Literary and Linguistic Computing* 23.4 (2008): 425-42.

[49] Hagan, Martin T., Howard B. Deumuth, and Mark H. Beale. *Neural Network Design*. Boston: PWS Publishing Company, 1996.

συνάρτηση έχει έναν μεγάλο αριθμό παραμέτρων (συναπτικά βάρη) τα οποία μπορούν να προσαρμοστούν. Η επιλογή της αρχιτεκτονικής του ΤΝΔ (ο αριθμός και η οργάνωση των νευρώνων και ο τύπος των συναρτήσεων μετάβασης) καθορίζει τη μορφή αυτής της συνάρτησης. Η εκπαίδευση του ΤΝΔ συνίσταται στην προσαρμογή των συναπτικών βαρών έτσι ώστε αυτό να επιστρέφει ορθά τις κατηγορίες εξόδου που έχουν συνδεθεί με τα δεδομένα εισόδου, ακριβώς όπως ο ανθρώπινος εγκέφαλος ρυθμίζει την ισχύ μιας νευρικής σύναψης μετά από παρατεταμένη παρατήρηση. Μαθηματικά αυτή η λειτουργία οδηγεί σε ένα κλασικό πρόβλημα ελαχιστοποίησης όπου θα πρέπει να επιλεχθούν οι παράμετροι εκείνοι που μικραίνουν το σφάλμα της συνάρτησης, η διαφορά δηλαδή μεταξύ των προβλέψεων της κατηγορίας εξόδου μιας συνάρτησης για ένα συγκεκριμένο σύνολο δεδομένων εισόδου και τις πραγματικές κατηγορίες εξόδου που αντιστοιχούν σε αυτά τα δεδομένα εισόδου.
Υπάρχουν πολλοί διαφορετικοί αλγόριθμοι ΤΝΔ που έχουν αναπτυχθεί για να ελαχιστοποιήσουν το σφάλμα κατηγοριοποίησης. Στην παρούσα έρευνα θα χρησιμοποιήσουμε έναν από τους πρώτους και δημοφιλέστερους, γνωστός ως ανάστροφη μάθηση (backpropagation). Το σφάλμα μεταξύ των επιθυμητών και των πραγματικών κατηγοριών εξόδου χρησιμοποιείται σε μία συνάρτηση που τροποποιεί τα συναπτικά βάρη σε όλο το ΤΝΔ έτσι ώστε η κατηγοριοποίηση που πραγματοποιεί το ΤΝΔ να έρθει πιο κοντά στα πραγματικά δεδομένα εξόδου που συνδέονται με τα διανύσματα εισόδου.

7.2 Αποτελέσματα πρόβλεψης του ΤΝΔ

Για τις ανάγκες της παρούσας έρευνας εκπαιδεύσαμε ένα ΤΝΔ με μεταβλητές εισόδου τα υφομετρικά χαρακτηριστικά που προέκυψαν ως στατιστικά σημαντικές (συνολικά 54 μεταβλητές) μέσα από την προηγούμενη διαδικασία ελέγχου των έξι ομάδων μεταβλητών με ΔΑ (βλ. Ενότητα 6). Η αρχιτεκτονική του ΤΝΔ που επιλέχθηκε ήταν αυτή με ένα κρυμμένο επίπεδο που περιείχε 7 νευρώνες. Κάθε ένας από αυτούς τους νευρώνες ενεργοποιείται όταν η συνάρτηση του σταθμισμένου αθροίσματος (weighted sum) των μεταβλητών εισόδου φτάσει σε κάποια κρίσιμη τιμή. Η συνάρτηση αυτή ονομάζεται συνάρτηση ενεργοποίησης (activation function) και η επιλογή της καθορίζει σημαντικά τη συμπεριφορά του ΤΝΔ αφού καθορίζει τη μετάβαση από επίπεδο σε επίπεδο. Στο ΤΝΔ της παρούσας έρευνας θα πρέπει να καθοριστούν δύο συναρτήσεις ενεργοποίησης, μία για τη μετάβαση από τα δεδομένα εισόδου στο κρυφό επίπεδο και μία για τη μετάβαση από το κρυφό επίπεδο στο επίπεδο εξόδου. Για την πρώτη μετάβαση επιλέχθηκε η σιγμοειδής συνάρτηση, ενώ για τη μετάβαση από το κρυφό επίπεδο στο επίπεδο εξόδου επιλέχθηκε η συνάρτηση softmax έτσι ώστε η έξοδος να είναι υποχρεωτικά στο διάστημα 0-1, δηλαδή το αποτέλεσμα να είναι προσέγγιση της πιθανότητας ένα κείμενο να το έχει γράψει άνδρας ή γυναίκα[50].
Τα ΤΝΔ κατά την εκπαίδευσή τους μπορούν να δημιουργήσουν πολύ ισχυρές συνάψεις με τα δεδομένα εισόδου και στην ουσία να μάθουν να κατηγοριοποιούν τέλεια μόνο αυτά που τους δόθηκαν ως δεδομένα εκπαίδευσης. Το φαινόμενο αυτό ονομάζεται «υπερεκπαίδευση» (overtraining) και μπορεί να απαξιώσει την πρακτική χρησιμότητα των ΤΝΔ, αφού τα καθιστά ανίκανα να χειριστούν οποιαδήποτε άλλα δεδομένα εκτός αυτών στα οποία εκπαιδεύθηκαν. Για το λόγο αυτό θα πρέπει τα δεδομένα εκπαίδευσης θα πρέπει να χωρίζονται σε δύο τμήματα, τα αμιγώς δείγμα εκπαίδευσης (training sample) και το δείγμα δοκιμής (testing sample). Το ΤΝΔ καθώς

[50] Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern Classification*. 2nd ed. New York: John Wiley & Sons, 2000.

προσαρμόζει τα συναπτικά βάρη κατά την εκπαίδευσή του, ελέγχει ταυτόχρονα τις προβλέψεις που κάνει στο δείγμα δοκιμής. Όταν παρατηρηθεί σημαντική απόκλιση στην ακρίβεια κατηγοριοποίησης μεταξύ δείγματος εκπαίδευσης και δείγματος δοκιμής, ο αλγόριθμος τροποποιεί τα συναπτικά βάρη έτσι ώστε να μειωθεί η απόσταση μεταξύ τους. Στην παρούσα έρευνα για να μειώσουμε περαιτέρω την πιθανότητα υπερεκπαίδευσης και να εξασφαλίσουμε ότι το ΤΝΔ που αναπτύξαμε έχει ικανή γενικευτική δύναμη, εξαιρέσαμε ένα τμήμα των αρχικών δεδομένων από την διαδικασία της εκπαίδευσης και το χρησιμοποιήσαμε ως εξωτερικό δείγμα (holdout sample). Η ακρίβεια κατηγοριοποίησης στο εξωτερικό δείγμα προσεγγίζει ικανοποιητικά την απόδοση του ΤΝΔ σε ανεξάρτητα δεδομένα, αφού οι τιμές τους δεν έχουν ενσωματωθεί στην εκπαίδευση του. Το μέγεθος των δειγμάτων που χρησιμοποιήθηκαν για εκπαίδευση και επικύρωση είναι τα ακόλουθα: Δείγμα Εκπαίδευσης: 413 κείμενα (57% των συνολικών δεδομένων), Δείγμα Δοκιμής: 212 κείμενα (30,3% των συνολικών δεδομένων), Εξωτερικό Δείγμα: 75 κείμενα (10,7% των συνολικών δεδομένων)

Η ακρίβεια της κατηγοριοποίησης του ΤΝΔ που αναπτύχθηκε φαίνεται στον παρακάτω πίνακα:

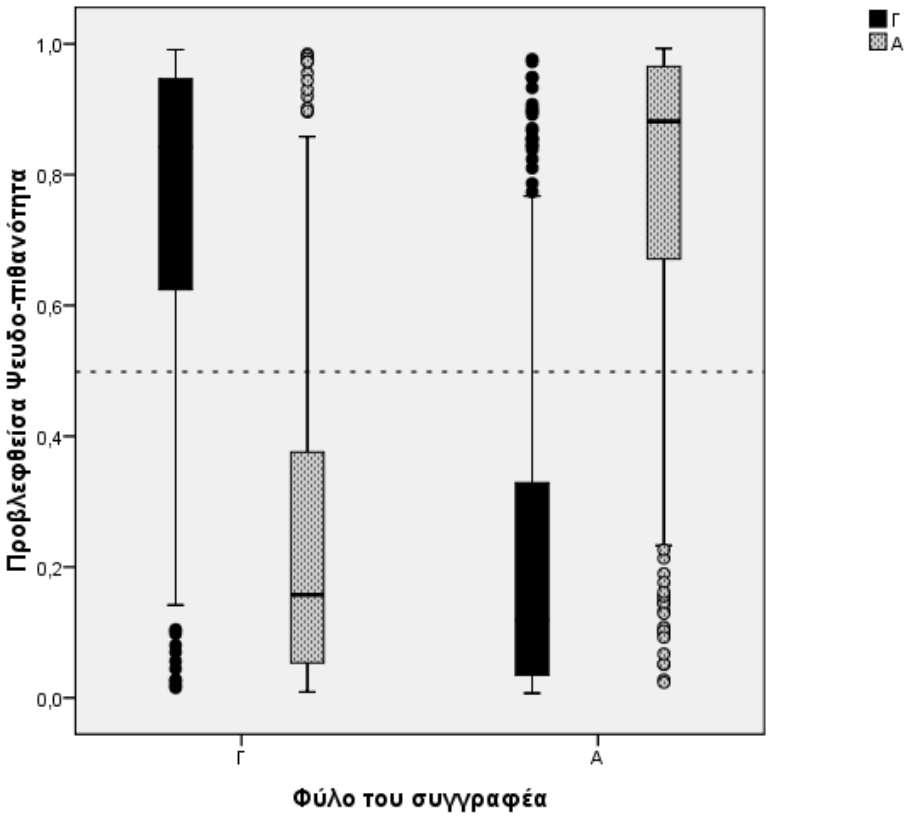
Πίνακας 2: Ακρίβεια της κατηγοριοποίησης του ΤΝΔ (Γ= Κείμενα Γυναικών, Α= Κείμενα Ανδρών)

Δείγμα	Πραγματικά δεδομένα	Δεδομένα πρόβλεψης Γ	Α	Ποσοστό σωστής πρόβλεψης
Εκπαίδευσης	Γ	178	26	87,3%
	Α	30	179	85,6%
	Συνολικό ποσοστό	50,4%	49,6%	**86,4%**
Δοκιμής	Γ	86	24	78,2%
	Α	18	84	82,4%
	Συνολικό ποσοστό	49,1%	50,9%	**80,2%**
Εξωτερικό	Γ	30	6	83,3%
	Α	9	30	76,9%
	Συνολικό ποσοστό	52,0%	48,0%	**80,0%**

Ο παραπάνω πίνακας συνοψίζει την απόδοση του ΤΝΔ στην πρόβλεψη του φύλου του συγγραφέα. Διαβάζεται από αριστερά προς τα δεξιά και δείχνει πόσα κείμενα που ανήκουν σε κάθε φύλο κατηγοριοποιήθηκαν ορθά από το ΤΝΔ ότι ανήκουν σε αυτό και σε πόσα το φύλο του συγγραφέα δεν προβλέφθηκε σωστά. Για παράδειγμα στο δείγμα εκπαίδευσης βλέπουμε ότι τα 178 κείμενα που είχαν γραφτεί από γυναίκες κατηγοριοποιήθηκαν ορθά ως γυναικεία κείμενα. Αντίθετα 26 κείμενα που γράφτηκαν από γυναίκες κατηγοριοποιήθηκαν εσφαλμένα ως κείμενα ανδρών (ακρίβεια πρόβλεψης των γυναικείων κειμένων στο δείγμα εκπαίδευσης 87,3%). Αντίστοιχα, εξετάζοντας την απόδοση του ΤΝΔ σε ανδρικά κείμενα βλέπουμε ότι 179 κείμενα που τα έχουν γράψει άνδρες κατηγοριοποιήθηκαν σωστά ότι ανήκουν σε άνδρες. Αντίθετα, 30 κείμενα που γράφτηκαν από άνδρες το ΤΝΔ εσφαλμένα τα κατηγοριοποίησε ως γυναικεία κείμενα (ακρίβεια πρόβλεψης των ανδρικών κειμένων στο δείγμα εκπαίδευσης 85,6%). Η συνολική ακρίβεια του ΤΝΔ στο δείγμα εκπαίδευσης είναι ο μέσος όρος της ακρίβειας που πέτυχε στα ανδρικά και στα γυναικεία κείμενα, δηλαδή 86,4% (87,3 + 85,6 / 2).

Η εξέταση των ποσοστών ακρίβειας δείχνει ότι το ΤΝΔ πέτυχε ακρίβεια πάνω από 80% σε όλα τα δείγματα, δηλαδή αναγνώρισε το φύλο του συγγραφέα με επιτυχία σε τουλάχιστον 8 από τα 10 κείμενα γεγονός που το τοποθετεί στα πιο ακριβή της σχετικής βιβλιογραφίας (βλ. και τις ακρίβειες που αναφέρονται στην Ενότητα 3). Τα ποσοστά ακρίβειας στο δείγμα δοκιμής και στο εξωτερικό δείγμα αν και ελαφρώς χαμηλότερα από την ακρίβεια του δείγματος εκπαίδευσης είναι αρκετά υψηλά γεγονός που υποδηλώνει ότι δεν υπήρξε υπερεκπαίδευση του ΤΝΔ στα δεδομένα εκπαίδευσης. Επομένως το συγκεκριμένο ΤΝΔ μπορεί να χρησιμοποιηθεί για να προβλέψει το φύλο του συγγραφέα σε κείμενα που δεν ανήκουν στο ΗΣΚ που χρησιμοποιήθηκε στην παρούσα έρευνα, εφόσον βέβαια αυτά ανήκουν στο ίδιο κειμενικό θέμα, γένος και μέσο.

Μια σημαντική παράμετρος που σχετίζεται με την απόδοση του ΤΝΔ στην πρόβλεψη του φύλου του συγγραφέα είναι η εξέταση της προβλεπτικής ακρίβειας ανά φύλο. Εκτός δηλαδή από τη γενική ακρίβεια πρόβλεψης του φύλου είναι σημαντικό να διερευνήσουμε κατά πόσο το ΤΝΔ έχει μάθει να διακρίνει το ίδιο καλά ανδρικά και γυναικεία κείμενα ή αντίθετα έχει εκπαιδευθεί ανισομερώς και επομένως διακρίνει καλύτερα μόνο μία κατηγορία. Για να απαντήσουμε στο συγκεκριμένο ερώτημα μπορούμε να διερευνήσουμε γραφικά την κατηγοριοποίηση ανά φύλο συγγραφέα μέσω του παρακάτω κιβωτιοδιαγράμματος (boxplot) (Διάγραμμα 6):



Διάγραμμα 6: Κιβωτοδιάγραμμα κατηγοριοποίησης του φύλου του συγγραφέα από το ΤΝΔ

Το παραπάνω διάγραμμα απεικονίζει την ακρίβεια κατηγοριοποίησης του ΤΝΔ ανά φύλο. Ο κάθετος άξονας (άξονας y) δείχνει την ψευδο-πιθανότητα ένα κείμενο να ανήκει στην κατηγορία (Γ= Γυναίκα συγγραφέας, Α= Άνδρας συγγραφέας) που ορίζεται στον παράλληλο άξονα (άξονας x). Όταν η ψευδο-πιθανότητα είναι πάνω από 0,5 τότε το κείμενο αυτό κατηγοριοποιείται στην αντίστοιχη κατηγορία του άξονα x. Αντίθετα, όταν ένα κείμενο εμφανίζει ψευδο-πιθανότητα κάτω από 0,5, τότε θεωρούμε ότι δεν ανήκει στην αντίστοιχη κατηγορία του άξονα x. Στο παραπάνω

διάγραμμα με μαύρο απεικονίζονται τα κείμενα που έχουν γράψει οι γυναίκες ενώ με γκρι συμβολίζονται τα κείμενα που έχουν γράψει οι άνδρες. Στην πρώτη από τα αριστερά μπάρα (μαύρη) εμφανίζονται τα κείμενα των γυναικών ο συντριπτικός όγκος των οποίων βρίσκεται πάνω από τη γραμμή αναφοράς (διακεκομμένη γραμμή) στο 0,5. Επομένως, πράγματι τα περισσότερα γυναικεία κείμενα ανάγονται ορθά στην κατηγορία Γ. Αυτά που δεν κατηγοριοποιήθηκαν σωστά αντίθετα εμφανίζεται κάτω από τη γραμμή αναφοράς ως μεμονωμένες κουκίδες. Αντίστοιχα, τα περισσότερα ανδρικά κείμενα της δεύτερης μπάρας (γκρι) εμφανίζονται κάτω από τη γραμμή αναφοράς με ψευδο-πιθανότητα να ανήκουν σε γυναίκες συγγραφείς να είναι κάτω του 0,5. Αντίστοιχα βλέπουμε ότι στην κατηγορία Α του άξονα x έχουν κατηγοριοποιηθεί σωστά τα περισσότερα ανδρικά κείμενα ενώ η πλειονότητα των γυναικείων κειμένων είναι κάτω από το 0,5 (τρίτη μπάρα). Συνολικά, βλέπουμε ότι η γραμμή αναφορά διαχωρίζει με ομοιογενές τρόπο τις επιμέρους κατηγοριοποιήσεις και δεν εμφανίζονται μπάρες όπου η γραμμή αναφοράς τις διχοτομεί. Αυτό δείχνει ότι το ΤΝΔ έχει μάθει εξίσου καλά να διακρίνει τόσο τους άνδρες όσο και τις γυναίκες συγγραφείς και επομένως δεν τίθεται θέμα μονομερούς εκπαίδευσης.

8 Συμπεράσματα

Η παρούσα εργασία επιχείρησε να διερευνήσει τη δυνατότητα πρόβλεψης του φύλου του συγγραφέα ενός κειμένου στηριζόμενη στις ακόλουθες προϋποθέσεις θεωρητικής και μεθοδολογικής υφής:

- Άνδρες και γυναίκες διαφοροποιούνται ως προς τη γλωσσική τους παραγωγή εξαιτίας ασυμμετριών στη βιολογική δομή και λειτουργία του εγκεφάλου, αλλά και λόγω διαφορετικών επιδράσεων από το κοινωνική δομή.
- Ένα μεγάλο κομμάτι της διαφοροποιημένης γλωσσικής συμπεριφοράς μεταξύ ανδρών και γυναικών στηρίζεται σε συνειδητές επιλογές που σχετίζονται άμεσα με τις επικοινωνιακές παραμέτρους της γλωσσικής χρήσης. Η συνειδητή και ενεργή διαφοροποίηση της γλωσσικών μέσων που χρησιμοποιούνται από άνδρες και γυναίκες δεν μπορεί να αξιοποιηθεί στην πρόβλεψη του φύλου του συγγραφέα αφού δεν μπορεί να αποδοθεί στα εγγενή χαρακτηριστικά του φύλου, αλλά απηχεί τις στρατηγικές επικοινωνίας του κάθε άτομου σε σχέση με την επικοινωνιακή περίσταση που βρίσκεται.
- Για να αποκλείσουμε την πιθανότητα η χρήση ενός γλωσσικού χαρακτηριστικού να αντιπροσωπεύει συνειδητές επιλογές εκ μέρους του ομιλητή / συγγραφέα, θα πρέπει να εξετάσουμε γλωσσικά χαρακτηριστικά τα οποία υπακούουν σε δύο βασικές αρχές: α) είναι πολύ συχνά και άρα ο ομιλητής / συγγραφέας δεν μπορεί να ελέγξει τη χρήση τους και β) δεν είναι φορείς λεξικής σημασίας και επομένως δε συσχετίζονται με την επικοινωνιακή περίσταση. Τα γλωσσικά χαρακτηριστικά που καλύπτουν τις παραπάνω προϋποθέσεις ανήκουν σε μια κατηγορία γλωσσικών μεταβλητών που ονομάζονται υφομετρικές και έχουν χρησιμοποιηθεί με επιτυχία κυρίως στην αναγνώριση του συγγραφέα σε κείμενα των οποίων η πατρότητα αμφισβητείται.
- Η χρήση υφομετρικών δεικτών προτείνεται παράλληλα με τη χρήση ενός προσεκτικά επιλεγμένου ΗΣΚ, που θα ελέγχει την αντιπροσώπευση των μετακειμενικών δεδομένων και θα ισορροπεί μεταξύ κειμενικής ποικιλίας και ισόποσης αντιπροσώπευσης συγκεκριμένων θεμάτων και γενών.

- Η μέτρηση των υφομετρικών χαρακτηριστικών στα κείμενα θα πρέπει να γίνεται με εργαλεία Επεξεργασίας Φυσικής Γλώσσας τα οποία αυτοματοποιούν τη διαδικασία της μέτρησης.
- Η στατιστική ανάλυση των δεδομένων που προκύπτουν από αυτές τις μετρήσεις είναι εγγενώς πολυπαραγοντική και θα πρέπει να λαμβάνει υπόψη της τη διαστασιμότητα του προβλήματος (dimensionality).
- Οι σχέσεις που αναπτύσσονται μεταξύ των υφομετρικών χαρακτηριστικών του κειμένου και του φύλου του συγγραφέα είναι περίπλοκες και πολλές φορές αντιφατικές. Για το λόγο αυτό τα Τεχνητά Νευρωνικά Δίκτυα είναι μία καλή επιλογή όσον αφορά τον αλγόριθμο κατηγοριοποίησης που μπορεί να χρησιμοποιηθεί, αφού είναι εγγενώς μη γραμμικά μοντέλα που παρουσιάζουν εύρωστα αποτελέσματα κατηγοριοποίησης ακόμα και με δεδομένα που δεν ακολουθούν την κανονική κατανομή, παρουσιάζουν ετεροσκεδασμό κ.ά.

Η παρούσα έρευνα κατέληξε σε ένα ΤΝΔ που προβλέπει σωστά το φύλο του συγγραφέα με ακρίβεια πάνω από 80%. Η ακρίβεια του συγκεκριμένου ΤΝΔ δικαιολογεί τις παραπάνω επιλογές και αποτελεί ισχυρή ένδειξη για την εγγενή διαφοροποίηση της γλωσσικής παραγωγής ανδρών και γυναικών.