23 GCDT: Σώμα Κειμένων της Γλώσσας των Εναγόμενων στο Ελληνικό Δικαστήριο

Αναστασία Κ. Κατρανίδου, Κατερίνα Θ. Φραντζή

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία παρουσιάζεται το GCDT (Greek Corpus of Defendants' Testimonies), το πρώτο σώμα κειμένων της γλώσσας εναγόμενων στο ελληνικό δικαστήριο, το οποίο δημιουργήθηκε από τις καταθέσεις τους μέσα στη δικαστική αίθουσα. Με σκοπό τη δημιουργία του γλωσσικού προφίλ των εναγόμενων, μελετώνται κάποια κειμενικά χαρακτηριστικά τα οποία θα μπορούσαν να δώσουν στοιχεία για την περιγραφή της γλώσσας τους. Παράλληλα, παρουσιάζεται το GCWT (Greek Corpus of Witnesses' Testimonies), ένα σώμα κειμένων το οποίο αποτελείται από καταθέσεις μαρτύρων μέσα στη δικαστική αίθουσα και δημιουργήθηκε με σκοπό να αποτελέσει ένα σώμα κειμένων αναφοράς για το GCDT.

GCDT: Corpus of Defendants' Testimonies in the Greek Court

Anastasia K. Katranidou, Katerina T. Frantzi

ABSTRACT

In this study we present GCDT, the Greek Corpus of Defendants' Testimonies. GCDT is the first Greek corpus of defendants' testimonies created of language material produced in a real Greek Court environment. We also present GCWT, a corpus that consists of witnesses' testimonies produced also in a real Greek Court environment. GCWT has been constructed as a reference corpus for GCDT. We study some stylometric features, aiming to the stylistic study of the speech and the stylometric profile of the defendants.

0 Εισαγωγή

Η εγκληματολογική ή δικανική γλωσσολογία αφορά στη μελέτη της γραπτής και προφορικής γλώσσας που σχετίζεται με το νομικό σύστημα και την εγκληματολογική έρευνα [9]. Προσφάτως, το ενδιαφέρον για την εγκληματολογική γλωσσολογία έχει αυξηθεί αισθητά καθώς η γλώσσα, όπως κάθε είδους εγκληματολογικό στοιχείο, μπορεί να χρησιμοποιηθεί κατά την αστυνομική έρευνα αλλά και στη δικαστική διαδικασία [2, 3, 4, 17, 21]. Σήμερα, η εφαρμογή νέων επιστημονικών μεθόδων στην ανάλυση καταθέσεων και η αξιολόγηση του γλωσσικού προφίλ κατηγορούμενων και μαρτύρων μπορεί να αποτελέσει καθοριστικό παράγοντα στη δικαστική διαδικασία.

Τα σώματα κειμένων είναι συλλογές από κείμενα φυσικά παραγόμενου λόγου και σχεδιάζονται και κατασκευάζονται έτσι ώστε να περιγράφουν όσο το δυνατόν πιο αντιπροσωπευτικά τη γλώσσα ή τη γλωσσική ποικιλία που αφορούν [16]. Οι μελέτες που βασίζονται σε σώματα κειμένων εγγυώνται την ακρίβεια και την πληρότητα της ποσοτικής επεξεργασίας [20]. Η ύπαρξη σωμάτων κειμένων σχετικών με νομικές διαδικασίες και συγκεκριμένα με καταθέσεις κατηγορούμενων είναι περιορισμένη και όσον αφορά στην ελληνική γλώσσα, δεν υπάρχει μέχρι στιγμής σχετικό σώμα κειμένων, καθώς η δημοσίευση δικαστικών πρακτικών είναι πολύ μικρή έως ανύπαρκτη.

Λόγω της έλλειψης σχετικών σωμάτων κειμένων, οι ερευνητές/τριες συχνά αναγκάζονται να κατασκευάσουν τα δικά τους «τεχνητά» σώματα κειμένων, ώστε να μελετήσουν την αποτελεσματικότητα των μεθόδων και εργαλείων τους [19, 26]. Είναι αυτονόητο ότι τέτοιου είδους σώματα κειμένων δεν έχουν την ίδια αξία με τα αυθεντικά. Οι Fitzpatrick & Bachenko [5] δίνουν οδηγίες σχετικές με την κατασκευή σωμάτων κειμένων προερχόμενα από δίκες. Ένα σώμα κειμένων κατασκευασμένο από πραγματικά γλωσσικά δεδομένα είναι αυτό του Fornaciari [6], το οποίο αφορά στην ιταλική γλώσσα και είναι το πρώτο σώμα δικαστικών κειμένων τα οποία έχουν συγκεντρωθεί σε πραγματικό περιβάλλον.

Μέχρι σήμερα δεν υπάρχει αντίστοιχη έρευνα που να βασίζεται σε ανάλυση λόγου ατόμων που εμπλέκονται σε δίκη στην Ελλάδα με σκοπό την εξαγωγή πληροφοριών σχετικά με το γλωσσικό τους προφίλ. Η παρούσα εργασία στοχεύει στην περιγραφή χαρακτηριστικών της γλώσσας που χρησιμοποιούν εναγόμενοι/ες που έχουν κατηγορηθεί για ανθρωποκτονία, κατά την απολογία τους μέσα στη δικαστική αίθουσα. Στη συγκεκριμένη μελέτη παρουσιάζεται το GCDT (Greek Corpus of Defendants' Testimonies), το σώμα κειμένων που δημιουργήθηκε από καταθέσεις κατηγορούμενων για ανθρωποκτονία μέσα στην δικαστική αίθουσα, καθώς και το GCWT, ένα σώμα κειμένων αναφοράς με σχετικά υφολογικά χαρακτηριστικά με το GCDT. Το GCWT αποτελείται από καταθέσεις μαρτύρων που έχουν συλλεχθεί μέσα στη δικαστική αίθουσα και σχετίζονται με υποθέσεις ανθρωποκτονίας. Στη συνέχεια παρουσιάζεται ανάλυση βασισμένη σε τυπικά χαρακτηριστικά της γλώσσας που χρησιμοποιούν οι κατηγορούμενοι/ες κατά την κατάθεση τους μέσα στη δικαστική αίθουσα.

1 GCDT και GCWT

Ο αρχικός σκοπός της έρευνας είναι η μελέτη της γλώσσας που χρησιμοποιείται από τους/τις εναγόμενους/ες κατά τη διάρκεια της δίκης στη δικαστική αίθουσα. Για το λόγο αυτόν, κατασκευάστηκε σώμα κειμένων που αποτελείται από μεταγραφές των προφορικών καταθέσεων των κατηγορούμενων κατά τη διάρκεια της ακροαματικής διαδικασίας μέσα στη δικαστική αίθουσα [14]. Οι εναγόμενοι/ες έχουν κατηγορηθεί για ανθρωποκτονία.

Η εύρεση και συλλογή του υλικού ήταν απαιτητική εργασία. Το υπάρχον γλωσσικό υλικό διατέθηκε από το Δικαστήριο της Θεσσαλονίκης αποκλειστικά για ερευνητικούς σκοπούς. Η εξουσιοδότηση για την πλήρη πρόσβαση στα δεδομένα που ενδιέφεραν στην έρευνα δόθηκε με την προϋπόθεση του σεβασμού των όποιων προσωπικών δεδομένων των εμπλεκόμενων ατόμων.

Το GCDT (Greek Corpus of Defendants' Testimonies), το πρώτο σώμα κειμένων της γλώσσας των εναγόμενων στο ελληνικό δικαστήριο, κατασκευάστηκε από μεταγραφές που περιλαμβάνουν τις ακριβείς καταθέσεις των κατηγορουμένων κατά την ακροαματική διαδικασία. Όλοι/ες οι εναγόμενοι/ες, των οποίων οι καταθέσεις αποτελούν μέρος του σώματος κειμένων έχουν κατηγορηθεί για ανθρωποκτονία. Το σώμα κειμένων είναι σε ηλεκτρονική μορφή και έχει δημιουργηθεί με φυσικό τρόπο, ως αποτέλεσμα της ακριβούς καταγραφής των καταθέσεων των κατηγορουμένων κατά την απολογία τους μέσα στη δικαστική αίθουσα, παραλείποντας άλλες συμμετοχές στην ακροαματική διαδικασία. Η επιλογή των γλωσσικών δεδομένων για την κατασκευή του GCDT αποδείχθηκε χρονοβόρα διαδικασία καθώς εμπλέκονταν αρκετά μη επιθυμητά γλωσσικά δεδομένα. Η αποθήκευση του υλικού έγινε σε μορφή απλού κειμένου με κωδικοποίηση ANSI και UTF-8, για τη δυνατότητα επεξεργασίας από τα περισσότερα λογισμικά επεξεργασίας γλωσσικού υλικού.

Συνολικά το σώμα κειμένων αποτελείται από 109.523 λέξεις, από 124 κατηγορούμενους/ες για ανθρωποκτονία σε 86 ακροαματικές διαδικασίες. Οι 110 από αυτούς/ές είναι άνδρες και οι 14 γυναίκες. Ως μητρική γλώσσα την ελληνική έχουν 91 από αυτούς/ές ενώ 33 κατέθεσαν με τη βοήθεια διερμηνέα[1]. Η μέση ηλικία των κατηγορούμενων τη στιγμή της ακρόασης είναι τα 38 χρόνια. Το μορφωτικό τους επίπεδο δεν είναι απολύτως βέβαιο, ενώ σχετικά με την εργασία τους, οι περισσότεροι των κατηγορουμένων είναι εργάτες, αγρότες, οικοδόμοι, ελεύθεροι επαγγελματίες, φοιτητές (δύο) ενώ είκοσι τέσσερις είναι άνεργοι. Σχεδόν σε όλες τις περιπτώσεις (91.7%) η απόφαση της εκδίκασης είναι καταδικαστική. Τις λίγες φορές για

[1] Υπάρχει μια αναπόφευκτη απώλεια στην ακρίβεια του λόγου του εναγόμενου κατά τη διάρκεια της μεταγραφής. Στην περίπτωση που χρησιμοποιείται διερμηνέας, η απώλεια ακρίβειας στην ομιλία του εναγόμενου είναι ακόμη μεγαλύτερη.

τις οποίες η απόφαση είναι αθωωτική είναι λόγω έλλειψης αποδεικτικών στοιχείων.

Ως σώμα κειμένων αναφοράς χρησιμοποιήθηκε στο παρελθόν ο Εθνικός Θησαυρός Ελληνικής Γλώσσας[2] (ΕΘΕΓ) [10], ο οποίος είναι ως σήμερα το μεγαλύτερο σώμα κειμένων γραπτού λόγου της νεοελληνικής γλώσσας και αποτελείται αποκλειστικά από γραπτό γλωσσικό υλικό. Ωστόσο, επειδή οι κατηγορούμενοι/ες έχουν συγκεκριμένο γλωσσικό ύφος κατά τη διάρκεια της ακροαματικής διαδικασίας και για την επίτευξη αξιόπιστων αποτελεσμάτων, κατασκευάστηκε σώμα κειμένων αναφοράς με παρόμοια υφολογικά χαρακτηριστικά με το υπό μελέτη σώμα κειμένων. Το νέο σώμα αναφοράς, το GCWT (Greek Corpus of Witnesses' Testimonies), αποτελείται από 395.925 λέξεις και προέρχεται από καταθέσεις μαρτύρων σχετικές με περιπτώσεις ανθρωποκτονίας. Τόσο το GCDT όσο και το GCWT έχουν κατασκευαστεί από μεταγραφές προφορικής γλώσσας κατά τη διάρκεια της δίκης. Το μέγεθος του σώματος αναφοράς είναι τέσσερις φορές μεγαλύτερο από το υπό μελέτη σώμα, πολύ κοντά στο ιδανικό μέγεθος ενός σώματος κειμένου αναφοράς [1, 15].

2 Θέματα ύφους

Αρχικά πραγματοποιήθηκαν τυπικές στατιστικές μετρήσεις: μελετήθηκαν οι συχνότητες των μερών του λόγου καθώς και οι συχνότητες των πιο συχνών λέξεων συμπεριλαμβανομένων των λειτουργικών λέξεων (άρθρα, προθέσεις, σύνδεσμοι κ.ά.). Για τη στατιστική επεξεργασία του σώματος κειμένων χρησιμοποιήθηκε το λογισμικό Wordsmith Tools v.5 [22]. Διαπιστώθηκε ότι τα ουσιαστικά που χρησιμοποιούν οι κατηγορούμενοι/ες σχετίζονται με την έννοια «έγκλημα», τα ρήματα βρίσκονται κυρίως στο παρελθόν και χρησιμοποιούνται αρκετά συχνά επιρρήματα αφού η γλώσσα τους τείνει να είναι περιγραφική. Όπως ήταν αναμενόμενο, στην κορυφή της λίστας των πιο συχνών λέξεων υπάρχουν λειτουργικές λέξεις, όπως «το», «με», «αυτό» κ.λπ., με τη λέξη «και» να καταλαμβάνει το 4% του συνολικού μεγέθους του σώματος. Οι 15 συχνότερα χρησιμοποιούμενες λέξεις στη λίστα καταλαμβάνουν περίπου το ένα τρίτο του σώματος κειμένων.

2.1 Σύγκριση με το Σώμα Αναφοράς

Η σύγκριση των συχνοτήτων του GCDT με τις αντίστοιχες ενός σώματος κειμένων αναφοράς δίνει πληροφορίες σχετικές με τα ιδιαίτερα χαρακτηριστικά της γλώσσας των καταθέσεων.

Το νέο σώμα κειμένων αναφοράς που κατασκευάστηκε περιέχει παρόμοια υφολογικά χαρακτηριστικά με το υπό μελέτη σώμα κειμένων GCDT. Στη συνέχεια πραγματοποιήθηκαν

[2] Εθνικός Θησαυρός Ελληνικής Γλώσσας, Ινστιτούτο Επεξεργασίας του Λόγου, ΑΘΗΝΑ Έρευνα & Καινοτομία Τεχνολογίες Πληροφορία, http://hnc.ilsp.gr

νέες μετρήσεις βασιζόμενες σε δείκτες που καθορίζουν το γλωσσολογικό προφίλ [2, 25].

Ο *λεξικός πλούτος* ενός κειμένου αναφέρεται στο πόσες διαφορετικές λέξεις εμφανίζονται στο κείμενο. Ο Πίνακας 1 δείχνει το ποσοστό των λέξεων με συχνότητα εμφάνισης στο σώμα κειμένου ένα ή δύο, δηλαδή των άπαξ λεγομένων και δις-λεγομένων, αντίστοιχα, και το λόγο των δις-λεγομένων προς τα άπαξ λεγόμενα, ο οποίος είναι ενδεικτικός του ύφους συγγραφέα [11]. Είναι εμφανές ότι τα άπαξ λεγόμενα καταλαμβάνουν σχεδόν το 50% όλων των λέξεων. Ο λόγος των μαρτύρων εμφανίζει λίγο μικρότερο ποσοστό άπαξ λεγομένων λόγω του ότι το λεξιλόγιο τους τείνει να επαναλαμβάνεται σε μεγαλύτερο βαθμό από ότι των κατηγορούμενων.

Πίνακας 1. Ο λεξικός πλούτος του GCDT και GCWT

	Άπαξ λεγόμενα %	**Δις-λεγόμενα %**	**Δις-/Άπαξ - λεγόμενα**
GCDT	49.61	15.40	0.31
GCWT	45.96	15.47	0.34

Σημαντικό συστατικό για το χαρακτηρισμό του ύφους αποτελούν οι συχνότητες μερών του λόγου [7, 29]. Για αυτόν το λόγο χρησιμοποιήθηκε ελληνικός επισημειωτής[3] και μετρήθηκαν οι σχετικές συχνότητες των λέξεων περιεχομένου (ουσιαστικά, ρήματα, επίθετα, επιρρήματα) και των λειτουργικών λέξεων (προθέσεις, άρθρα, αντωνυμίες κ.ά.). Η *λεξική πυκνότητα*, που αξιολογεί το ποσοστό των λέξεων περιεχομένου στο κείμενο, είναι ένα μέτρο της πληροφορίας που περιέχει το κείμενο [8]. Έτσι, τα κείμενα προφορικού λόγου τείνουν να έχουν χαμηλότερη λεξική πυκνότητα (σχεδόν 45%) από ότι τα κείμενα γραπτού λόγου (πάνω από 50%) [13, 27, 28]. Οι συχνότητες των λέξεων περιεχομένου, των λειτουργικών λέξεων και η λεξική πυκνότητα του GCDT και GCWT φαίνονται στον Πίνακα 2.

[3] Ομάδα Επεξεργασίας Φυσικής Γλώσσας, Τμήμα Πληροφορικής – Οικονομικό Πανεπιστήμιο Αθηνών, http://nlp.cs.aueb.gr/software.html

Πίνακας 2. Οι συχνότητες των λέξεων περιεχομένου, των λειτουργικών λέξεων και η λεξική πυκνότητα του GCDT και GCWT

	συχνότητα λέξεων περιεχομένου %	συχνότητα λειτουργικών λέξεων %	λεξική πυκνότητα %
GCDT	44.21	55.7	44.2
GCWT	45.83	54.1	45.8

Τόσο το GCDT όσο και το GCWT έχουν χαμηλή λεξική πυκνότητα σε σύγκριση με την τυπική λεξική πυκνότητα των γραπτών κειμένων, δεδομένου ότι προέκυψαν από μεταγραφές ομιλούμενης γλώσσας και είναι κατασκευασμένα από ειδικό γλωσσικό υλικό. Το σώμα αναφοράς έχει υψηλότερη λεξική πυκνότητα από το GCDT, λόγω του ότι το GCWT περιέχει μαρτυρίες από εξειδικευμένους μάρτυρες, όπως ιατροδικαστές και αστυνομικούς, οι οποίοι τείνουν να χρησιμοποιούν πιο περιγραφική γλώσσα και περισσότερες λέξεις περιεχομένου που περιέχουν πληροφορίες κατά την κατάθεση τους στο δικαστήριο.

Οι τυπικές αποκλίσεις τόσο του μήκους λέξεων όσο και του μήκους προτάσεων μπορούν επίσης να δώσουν πληροφορίες για τον τρόπο με τον οποίο οι κατηγορούμενοι/ες χρησιμοποιούν τη γλώσσα.

Πίνακας 3. Μήκος λέξεων, μήκος προτάσεων και τυπική απόκλιση των GCDT και GCWT

	Μέσος όρος μήκους λέξεων σε γράμματα	Τυπική απόκλιση μήκους λέξεων	Μέσος όρος μήκους προτάσεων σε λέξεις	Τυπική απόκλιση μήκους προτάσεων
GCDT	4.44	2.27	8.27	6.32
GCWT	4.64	2.54	8.76	6.46

Με τις μετρήσεις διαπιστώθηκε ότι υπάρχουν μικρές διαφορές μεταξύ του λόγου των εναγόμενων και των μαρτύρων, όπως φαίνεται στον Πίνακα 3. Υπάρχει μια μικρή διαφορά στο μήκος λέξεων μεταξύ των δύο σωμάτων κειμένων, καθώς οι μάρτυρες φαίνεται να χρησιμοποιούν περισσότερες και μεγαλύτερες λέξεις πιο συχνά από τους/τις κατηγορούμενους/ες. Ο μέσος όρος μήκους προτάσεων των κατηγορουμένων είναι μικρότερος από αυτόν των μαρτύρων, όπως και η τυπική απόκλιση (6,32 λέξεις) για τους/τις κατηγορούμενους/ες σε σύγκριση με τους μάρτυρες (6,46 λέξεις). Λαμβάνοντας υπόψη τη

φύση και των δύο σωμάτων κειμένων, οι χαμηλές τυπικές αποκλίσεις δεν προκαλούν έκπληξη. Και τα δύο σώματα κειμένων προέρχονται από καταθέσεις μέσα στο δικαστήριο και, εκτός από κάποια περιγραφικά κομμάτια ομιλίας, αποτελούνται από απαντήσεις. Συνήθως, οι κατηγορούμενοι/ες και οι μάρτυρες χρησιμοποιούν μονολεκτικές ή σύντομες απαντήσεις. Επιπλέον, το μορφωτικό επίπεδο των κατηγορουμένων είναι κατά μέσο όρο χαμηλότερο από αυτό των μαρτύρων και επομένως τείνουν να χρησιμοποιούν απλούστερες λέξεις και μικρότερες προτάσεις.

Η χρήση των λέξεων-κλειδιά δίνει μία άλλη πτυχή σύγκρισης της λίστας λέξεων που εξάγεται από το υπό μελέτη σώμα κειμένου, GCDT, με τη λίστα λέξεων που εξάγεται από ένα σώμα κειμένου αναφοράς [23]. Το αποτέλεσμα αυτής της σύγκρισης είναι η τιμή *keyness*, η οποία περιγράφει τη σημασία μιας λέξης μέσα σε ένα κείμενο. Όσο πιο μεγάλη είναι αυτή η τιμή μιας λέξης σε ένα κείμενο σε σύγκριση με ένα κείμενο αναφοράς τόσο περισσότερο χαρακτηριστική είναι η λέξη και θεωρείται λέξη κλειδί για το κείμενο που εμφανίζεται. Οι λέξεις-κλειδιά είναι «λέξεις ασυνήθιστης συχνότητας σε σύγκριση με ένα σώμα αναφοράς» [24]. Οι μετρήσεις έδειξαν ότι λέξεις-κλειδιά του GCDT είναι κυρίως ρήματα στο πρώτο πρόσωπο, ενικού αριθμού σε παρελθοντικό χρόνο, που χρησιμοποιούνται για να περιγράψουν μια ενέργεια ή ένα συναίσθημα. Επιπλέον, κάποιες λέξεις εμφανίζουν αρνητική τιμή *keyness*, δηλαδή εμφανίζονται αρκετά σπάνια σε σχέση με το σώμα αναφοράς. Για παράδειγμα, οι δύο λέξεις-κλειδιά «κατηγορούμενοι» και «θύματα» εμφανίζονται αρκετά συχνά στο σώμα αναφοράς σε σύγκριση με το GCDT, καθώς οι κατηγορούμενοι/ες σπάνια αναφέρονται σε αυτούς τους δύο όρους.

4 Συμπεράσματα και μελλοντική εργασία

Στην εργασία παρουσιάστηκε το GCDT, σώμα κειμένων της γλώσσας εναγόμενων στο ελληνικό δικαστήριο και το GCWT, σώμα κειμένων αναφοράς το οποίο αποτελείται από καταθέσεις μαρτύρων. Από τις στατιστικές αναλύσεις που πραγματοποιήθηκαν στο λόγο των κατηγορουμένων, διαπιστώθηκε ότι χρησιμοποιούν αρκετά συχνά λέξεις που χαρακτηρίζονται ως σπάνιες στην καθομιλουμένη. Τα ουσιαστικά που χρησιμοποιούν σχετίζονται με την έννοια «έγκλημα», τα ρήματα βρίσκονται κυρίως σε παρελθοντικό χρόνο και τα επιρρήματα είναι συχνά καθώς η γλώσσα των κατηγορουμένων τείνει να είναι περιγραφική. Όσον αφορά στις λέξεις με σχετικά υψηλές συχνότητες, παρατηρήθηκε ότι αφενός τόσο οι κατηγορούμενοι/ες όσο και οι μάρτυρες εμφανίζουν χαμηλή λεξική πυκνότητα σε σύγκριση με την τυπική λεξική πυκνότητα γραπτών κειμένων, αφ’ ετέρου το σώμα κειμένων αναφοράς είναι «πυκνότερο» λόγω της πλουσιότερης γλώσσας των καταθέσεων των ειδικών μαρτύρων. Παρά των φαινομενικά παρόμοιων υφολογικών

χαρακτηριστικών, ορισμένες λέξεις-κλειδιά του GCDT είναι ασυνήθιστα συχνές σε σύγκριση με τις εμφανίσεις τους στο GCWT.

Όσον αφορά στη μελλοντική εργασία, αυτή περιλαμβάνει καταρχάς την ενημέρωση του GCDT με καταθέσεις των κατηγορουμένων κατά την προανακριτική διαδικασία. Η ενημέρωση θα επιτρέψει τη σύγκριση μεταξύ της γλώσσας που χρησιμοποιούν οι κατηγορούμενοι/ες εντός και εκτός του δικαστηρίου. Η δεύτερη συνιστώσα της μελλοντικής εργασίας περιλαμβάνει τη διερεύνηση χαρακτηριστικών από το χώρο της Ανάκτησης Πληροφορίας και της Γλωσσικής Μοντελοποίησης [12, 18].

5 Βιβλιογραφία

[1] Berber-Sardinha, T. (2000). *Comparing corpora with WordSmith Tools: How large must the reference corpus be?* In Proceedings of the Workshop on Comparing Corpora, WCC 00, Hong Kong, October 7th, 2000, Association for Computational Linguistics 2000, 9, σσ. 7-13.

[2] Broussalis, G., Markopoulos, G., & Mikros, G. (2012). *Stylometric profiling of the Greek Legal Corpus*. Selected Papers of the 10th International Conference of Greek Linguistics, ICGL-10. Komotini: Democritus University of Thrace, σσ. 167-176.

[3] Cotterill, J. (2003). *Language and Power in Court: A Linguistic Analysis of the O.J. Simpson Trial*. Palgrave Macmillan.

[4] Coulthard, M. (2004). *Author identification, idiolect, and linguistic uniqueness*. Applied Linguistics, 25(4), σσ. 431-447.

[5] Fitzpatrick, E. & Bachenko, J. (2012) *Building a Data Collection for Deception Research*. Montclair State University

[6] Fornaciari, T. (2012). *Deception Detection in Italian Court testimonies*. University of Trento.

[7] Gamon, M. (2004). *Linguistic correlates of style: Authorship classification with deep linguistic analysis features*. In Proceedings of the 20th International Conference on Computational Linguistics, σσ. 611-617.

[8] García, A.M., & Martin, J.C., (2006). *Function words in authorship attribution studies*. Literary and Linguistic Computing 22(1), σσ. 49-66.

[9] Grant, T. & Perkins, R. (2013). *Forensic Linguistics*. In J. A. Siegel & P.J. Saukko (eds.) Encyclopedia of Forensic Sciences, 2nd edn, σσ.174-177.

[10] Hatzigeorgiu, N., Gavrilidou, M., Piperdis, S., Carayannis, G., Papakostopoulou, A., Athanasia, S., & Iason, D. (2000). *Design and implementation of the online ILSP Greek*

Corpus. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperdis & G. Stainhaouer (Eds.), Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), vol. III. Athens, Greece: ELRA, σσ. 1737-1742.

[11] Hoover, D. (2003). *Another perspective on vocabulary richness. Computers and the Humanities*, 37, σσ. 151-178.

[12] Houvardas, J., & Stamatatos, E. (2006). *N-Gram Feature Selection for Authorship 0Identification*. In: Euzenat, J., Domingue, J. (eds.). Artificial Intelligence: Methodology, Systems, and Applications, vol. 4183. Berlin/Heidelberg: Springer, σσ. 77-86.

[13] Johansson, V. (2008). *Lexical diversity and lexical density in speech and writing: a developmental perspective*, Working Papers 53, σσ. 61-79.

[14] Katranidou A., & Frantzi K. (2016). *The Greek Corpus of Defendants' Testimonies: frequent use of infrequent words*. European Journal of Humanities and Social Sciences 3, 2016, σσ. 25-29.

[15] Koppel, M., Argamon, S., & Shimoni, A.R. (2002). *Automatically categorizing written texts by author gender*. Literary and Linguistic Computing, 17(4), σσ. 401-412.

[16] McEnery T. & Wilson A. (2001). *Corpus linguistics: an introduction, Edinburgh textbooks in empirical linguistics*. Edinburgh: Edinburgh University Press.

[17] McMenamin, G. (2002). *Forensic Linguistics: advances in forensic stylistics*. Boca Raton: CRC Press LLC.

[18] Mikros, G. K. (2012). *Authorship Attribution and Gender Identification in Greek Blogs*. Methods and Applications of Quantitative Linguistics 21, σσ. 21-32.

[19] Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. (2003). *Lying Words: Predicting Deception From Linguistic Styles*. Personality and Social Psychology Bulletin, 29(5), σσ. 665–675.

[20] O'Keeffe, A. & McCarthy, M. (eds) (2010). *The Routledge Handbook of Corpus Linguistics*. London & New York: Routledge.

[21] Olsson, J. (2004). *Forensic Linguistics: An Introduction to Language, Crime and the Law*. London: Continuum.

[22] Scott, M. (1998). WordSmith Tools Version 3. Oxford: Oxford University Press.

[23] Scott, M. (2001). *Comparing corpora and identifying key words, collocations and*

frequency distributions through the WordSmith Tools suite of computer programs. In M. Ghadessy, A. Henry, & R. L. Roseberry, (Eds.), Small Corpus Studies and ELT. Amsterdam/Philadelphia: John Benjamins Publishing Co, σσ. 47-67.

[24] Scott, M., & Tribble, C. (2006). *Textual Patterns: keyword and corpus analysis in language education*. Amsterdam: Benjamins.

[25] Stamatatos, E. (2009). *A survey of modern authorship attribution methods.* Journal of the American Society for information Science and Technology, 60(3), σσ. 538-556.

[26] Strapparava, C. and Mihalcea, R. (2009). *The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language*. In Proceeding ACLShort '09 - Proceedings of the ACL-IJCNLP 2009 Conference Short Papers.

[27] To, V., Fan, S., & Thomas, D.P. (2013). *Lexical density and Readability: A case study of English Textbooks*. The International Journal of Language, Society and Culture, 37(7), σσ. 61-71.

[28] Ure, J. (1971). *Lexical density and register differentiation*. In G. Perren and J.L.M. Trim (Eds.), Applications of Linguistics. London: Cambridge University Press, σσ. 443-452.

[29] Zhao Y., & Zobel, J. (2005*). Effective and scalable authorship attribution using function words*. In Proceedings of the 2nd Asia Information Retrieval Symposium, AIRS 2005.

Αναστασία Κατρανίδου
Υποψήφια Διδάκτορας
Τμήμα Μεσογειακών Σπουδών
Πανεπιστήμιο Αιγαίου
Ηλ.ταχ.: katranid@gmail.com

Κατερίνα Φραντζή
Αναπληρώτρια Καθηγήτρια
Τμήμα Μεσογειακών Σπουδών
Πανεπιστήμιο Αιγαίου
Ηλ.ταχ.: frantzi@rhodes.aegean.gr