UNIVERSITY OF THE AEGEAN

DOCTORAL THESIS

Text analysis and machine learning classification of defendants' testimonies in Greek Courtroom in order to create their linguistic profile

Author:

Anastasia Katranidou

Supervisor:

Prof. Katerina Frantzi

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Ph.D.) in the

Department of Mediterranean Studies, Faculty of Humanities

January, 2022

Advising Committee of this Doctoral Thesis

Katerina Frantzi, Supervisor
Professor, University of the Aegean,
Greece

Efstathios Stamatatos, Advisor
Professor, University of the Aegean,
Greece

Ioannis Stribis, Advisor
Assistant Professor, University of the Aegean,
Greece

Approved by the Examining Committee

Katerina Frantzi
Professor, University of the Aegean,
Greece

Efstathios Stamatatos
Professor, University of the Aegean,
Greece

Ioannis Stribis
Assistant Professor, University of the Aegean,
Greece

Konstantinos Magliveras
Professor, University of the Aegean,
Greece

Eleni Panaretou
Associate Professor, National and Kapodistrian University of Athens,
Greece

Christina Alexandris
Associate Professor, National and Kapodistrian University of Athens,
Greece

Georgios Fessakis
Professor, University of the Aegean,
Greece

Declaration of Authorship

I declare that this thesis, titled ***Text analysis and machine learning classification of defendants' testimonies in Greek Courtroom in order to create their linguistic profile,*** has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. All direct or indirect sources used are acknowledged as references and where I have consulted the published work of others, this is always clearly attributed.

Στην κόρη μου, Κυβέλη

(To my daughter, Kyveli)

"Words are free. It's how you use them that may cost you."

Rev J. Martin

ACKOWLEDGEMENTS

I would like to thank Prof. Katerina T. Frantzi for the supervision of my work and for the invaluable guidance, support, encouragement, and effort to offer every possible help to finish this thesis. It was a pleasant experience working with her and I am very glad finishing this thesis under her supervision.

During the first steps of my study, and afterwards, I received remarkable advice from Prof. Efstathios Stamatatos, which gave a boost in my research.

I would also like to thank Assistant Prof. Ioannis Stribis for serving as my committee member.

I also found full cooperation at the Court of Thessaloniki, where the data were collected, by authorizing the processing of the court files and the data collection and allowing the creation of my corpora. I had the fortune to know the court clerk Mrs. Niki Tsolaki who helped me with several bureaucratic and practical problems. Without her help this dissertation would not have even started.

It seems common to thank one's parents, however I am aware that without their continuous support in every phase of my life, I would not have managed to accomplish my goals. I am forever indebted to them for giving me the opportunities and experiences that have made me who I am.

This dissertation was completed a few months after my daughter was born, to whom I dedicate this research.

ABSTRACT

Forensic Linguistics attempts to analyze the language that relates to the law, either as evidence or as legal discourse. Language as legal discourse includes, among others, the discourse inside the court room. Crime profiling, or offender profiling, is one of the most important areas of research in Forensic Linguistics and should be its fundamental task, since by examining a criminal behavior one can evaluate or even predict future criminal actions. The identification of specific characteristics of an individual committing a crime is achieved by a thorough systematic observational process and an analysis of the crime scene, the victim, the forensic evidence, and the known facts of the crime. In this dissertation, using natural language analysis techniques from the field of author profiling, where one can extract information about the age, education, sex, etc., of the author of a given text, we attempt to define the linguistic profile of a criminals' category (that of the murderers) and, at a later stage, to develop a machine learning classifier which would predict whether a text belongs to that category, i.e., it has been written or said by a murderer or not.

First, we created three corpora from text data that we derived from real trial briefs of a Greek court. The first one concerned testimonies of defendants accused of murder, the second one was constructed from testimonies of witnesses and the last one consisted of testimonies of the defendants in their interrogation phase before their trial. It is obvious that the creation of this research would not have been possible without the possession of these trial briefs, which were difficult to get access to and required a time-consuming procedure. The latter is the main reason that no corresponding research has been done so far in Greece.

Having created these corpora, we quantified the way defendants of murder speak inside a Greek courtroom during their testimony, by studying several stylometric features of their language and comparing them with both the general language and the language of the witnesses. As a result, we have been able to extract some linguistic patterns used by murders in their testimonies. Moreover, some of these features proved to be more crucial, than others, in being able to describe the language profile of the speaker of a testimony.

The results we extracted of the quantitative analysis and knowing that the court proceedings and police investigations in Greece do not have appropriate and easy-to-use tools that can provide additional assistance in evaluating the statements of the accused, led us to the idea of constructing an automated text classifier using as training data the most useful stylometric features of the defendants' testimonies. Automated text classification has been considered as a vital method to manage a vast number of documents in digital form since its goal is the construction of a classification model (classifier) that is able to automatically assign labels to electronic texts by learning specific features of each category. In any case, statistics has been more concerned with testing hypotheses, whereas machine learning has been more concerned with formulating the process of generalization as a search through possible hypotheses.

Hence, we present a text classification machine learning model, the GDCT classifier, which was trained using the appropriate stylometric features, as demonstrated in our study. The experimental results of our corpora, covering the testimonies of 269 defendants and witnesses in total, verify the effectiveness of our method. Specifically, we prove that GCDT classifier can characterize a person who testifies, as guilty or not, with 93% accuracy. Our model does not seek to replace any judge or investigator but can offer to the trial procedure an additional tool in evaluating a murderer's testimony. This research is a pioneering method both in Greek Forensic Linguistics and in the Greek judicial process.

ΠΕΡΙΛΗΨΗ (ABSTRACT IN GREEK)

Η δικανική ή εγκληματολογική γλωσσολογία επιχειρεί να αναλύσει τη γλώσσα που σχετίζεται με το νόμο, είτε στην περίπτωση που η γλώσσα αποτελεί κάποιο αποδεικτικό στοιχείο μιας εγκληματικής πράξης είτε στην περίπτωση της γλώσσας ως νομικού λόγου. Η γλώσσα ως νομικός λόγος περιλαμβάνει, μεταξύ άλλων, το λόγο μέσα στη δικαστική αίθουσα. Το εγκληματικό προφίλ, το οποίο στη διεθνή βιβλιογραφία καταγράφεται και ως προφίλ του παραβάτη/δράστη, είναι ένας από τους σημαντικότερους τομείς έρευνας στην εγκληματολογική γλωσσολογία και θα έπρεπε να είναι και από τα κύρια καθήκοντά του, καθώς εξετάζοντας μια εγκληματική συμπεριφορά μπορεί κανείς να αξιολογήσει ή ακόμη και να προβλέψει μελλοντικές εγκληματικές ενέργειες. Για να προσδιοριστούν συγκεκριμένα χαρακτηριστικά του ατόμου που διαπράττει ένα έγκλημα, απαιτείται μια διεξοδική και συστηματική διαδικασία παρατήρησης και ανάλυσης της σκηνής του εγκλήματος, του θύματος, των αποδεικτικών στοιχείων και των γεγονότων του εγκλήματος.

Σε αυτήν τη διατριβή, χρησιμοποιώντας τεχνικές ανάλυσης φυσικής γλώσσας από το ερευνητικό πεδίο της δημιουργίας προφίλ του συγγραφέα (author profiling), όπου μπορεί κανείς να εξάγει πληροφορίες σχετικά με την ηλικία, την εκπαίδευση, το φύλο κ.α. του συγγραφέα ενός συγκεκριμένου κειμένου, επιχειρούμε να καθορίσουμε το γλωσσικό προφίλ μιας συγκεκριμένης κατηγορίας εγκληματιών, αυτής των ανθρωποκτόνων και, σε μεταγενέστερο στάδιο, να αναπτύξουμε ένα μοντέλο κατηγοριοποίησης ή ταξινόμησης (classifier) μηχανικής μάθησης που θα προβλέπει εάν ένα κείμενο ανήκει σε αυτήν την κατηγορία των εγκληματιών ή όχι, δηλαδή αν έχει γραφτεί ή ειπωθεί από έναν ανθρωποκτόνο ή όχι.

Αρχικά, δημιουργήσαμε τρία σώματα κειμένου (corpora) από κείμενα που προήλθαν εξ ολοκλήρου από καταθέσεις που έγιναν σε πραγματικές δίκες σε αίθουσες των ελληνικών δικαστηρίων από τις αντίστοιχες δικογραφίες. Το πρώτο σώμα κειμένου που κατασκευάσαμε αφορά σε απολογίες κατηγορουμένων που κατηγορούνταν για ανθρωποκτονία, το δεύτερο δημιουργήθηκε από καταθέσεις μαρτύρων που κατέθεταν στις ίδιες δικαστικές υποθέσεις των κατηγορουμένων, και το τελευταίο αποτελείται από καταθέσεις των κατηγορουμένων στον ανακριτή, κατά την προανακριτική διαδικασία, πριν ακόμα παραπεμφθούν σε δίκη. Είναι προφανές ότι η δημιουργία αυτής της διατριβής δε θα ήταν εφικτή χωρίς την κατοχή

αυτών των δικογραφιών, στις οποίες η πρόσβαση ήταν δύσκολη και η απόκτησή τους ήταν μια διαδικασία χρονοβόρα και απαιτητική. Το τελευταίο είναι ο κύριος λόγος που μέχρι στιγμής δεν έχει γίνει αντίστοιχη έρευνα στην Ελλάδα.

Έχοντας δημιουργήσει τα παραπάνω σώματα κειμένων από τις δικογραφίες, ποσοτικοποιήσαμε τον τρόπο με τον οποίο οι κατηγορούμενοι μιλούν μέσα σε μια ελληνική δικαστική αίθουσα κατά τη διάρκεια της απολογίας τους, μελετώντας τα κυριότερα υφολογικά χαρακτηριστικά της γλώσσας που χρησιμοποιούν και συγκρίνοντάς τα με την καθομιλουμένη γλώσσα και τη γλώσσα των μαρτύρων που καταθέτουν στις ίδιες δικαστικές υποθέσεις. Ως εκ τούτου, καταφέραμε να εξάγουμε ορισμένα γλωσσικά μοτίβα που χρησιμοποιούν οι ανθρωποκτόνοι στις καταθέσεις τους. Επιπλέον, μερικά από τα προαναφερθέντα υφολογικά χαρακτηριστικά αποδείχτηκαν πιο καθοριστικά, από κάποια άλλα, όσον αφορά στην ικανότητα τους να μπορούν να περιγράψουν το γλωσσικό προφίλ του ομιλητή μιας κατάθεσης.

Τα αποτελέσματα από την ποσοτική ανάλυση που εξήγαμε αναφορικά με το γλωσσικό προφίλ των ανθρωποκτόνων και γνωρίζοντας ότι οι δικαστικές διαδικασίες και οι αστυνομικές έρευνες στην Ελλάδα δε διαθέτουν κατάλληλα και εύχρηστα εργαλεία που να μπορούν να δώσουν μια επιπλέον βοήθεια στην αξιολόγηση των καταθέσεων των κατηγορουμένων, μας οδήγησαν στην ιδέα της κατασκευής ενός αυτοματοποιημένου μοντέλου ταξινόμησης κειμένων, χρησιμοποιώντας για δεδομένα εκπαίδευσης τα πιο χρήσιμα υφολογικά χαρακτηριστικά που εξήγαμε από τις καταθέσεις των κατηγορουμένων. Η αυτοματοποιημένη ταξινόμηση κειμένου έχει θεωρηθεί ως μια μέθοδος ζωτικής σημασίας για τη διαχείριση τεράστιου αριθμού εγγράφων που βρίσκονται σε ψηφιακή μορφή, καθώς στόχος της είναι η κατασκευή ενός μοντέλου ταξινόμησης που να είναι σε θέση να εκχωρεί αυτόματα ετικέτες σε ηλεκτρονικά κείμενα μαθαίνοντας από συγκεκριμένα χαρακτηριστικά της κάθε κατηγορίας. Σε κάθε περίπτωση, η στατιστική αφορά περισσότερο στη δοκιμή υποθέσεων, ενώ η μηχανική μάθηση έχοντας ως δεδομένο πιθανές υποθέσεις, προσπαθεί να διαμορφώσει μια διαδικασία γενίκευσης.

Επομένως, παρουσιάζουμε ένα μοντέλο μηχανικής μάθησης ταξινόμησης κειμένων, το μοντέλο ταξινόμησης GDCT, το οποίο εκπαιδεύτηκε χρησιμοποιώντας τα κατάλληλα, όπως αποδείχτηκαν από τη μελέτη μας, υφολογικά χαρακτηριστικά από τη γλώσσα που χρησιμοποιούν οι ανθρωποκτόνοι και οι μάρτυρες στις καταθέσεις τους. Τα αποτελέσματα

από τα πειράματα που έγιναν στα σώματα κειμένων μας, τα οποία αποτελούνται από τις μαρτυρίες 269 κατηγορουμένων και μαρτύρων συνολικά, επιβεβαιώνουν την αποτελεσματικότητα της μεθόδου μας. Συγκεκριμένα, αποδεικνύουμε ότι το μοντέλο ταξινόμησης GCDT μπορεί να χαρακτηρίσει ένα άτομο που καταθέτει, ως ένοχο ή όχι, με ακρίβεια 93%. Το μοντέλο μας δεν επιδιώκει σε καμία περίπτωση να αντικαταστήσει το ρόλο ενός δικαστή ή ανακριτή, αλλά μπορεί να προσφέρει στη δικαστική διαδικασία ένα επιπλέον εργαλείο για την αξιολόγηση της κατάθεσης ενός δολοφόνου. Η συγκεκριμένη έρευνα αποτελεί μια πρωτοπόρα μέθοδο τόσο για την ελληνική δικανική γλωσσολογία όσο και για την ελληνική δικαστική διαδικασία.

PUBLICATIONS

Parts of this thesis appeared in the following previous publications:

- Katranidou, A., & Frantzi, K. T. (2016). The Greek Corpus of Defendants' Testimonies: frequent use of infrequent words. *European Journal of Humanities and Social, 3*, 25-29. https://doi.org/10.20534/ejhss-16-3-25-29
- Frantzi, K. T., & Katranidou, A. K. (2017). A Corpus-Based Analysis of the Language Used by Defendants of Homicide in Court. *World Journal of Social Science Research, 4*(2). https://doi.org/10.22158/wjssr.v4n2p164

ABBREVIATIONS

(In alphabetical order)

AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under Curve
BOW	Bags of Words
CART	Classification and Regression Trees
CGT	Corpus of Greek Texts
CSG	Corpus of Spoken Greek
CW	Content Words
DRS	Direct Report Speech
ECHR	European Court of Human Rights
GCDT	Greek Corpus of Defendants' Testimonies
GCWT	Greek Corpus of Witnesses' Testimonies
GGLC	Greek General Language Corpus
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
FW	Function Words
HNC	Hellenic National Corpus
IR	Information Retrieval
KNN	k-Nearest Neighbors
LDA	Linear Discriminant Analysis

LR	Logistic Regression
NB	Naive Bayes
NLP	Natural Language Processing
OCR	Optical Character Recognition
POS	Part of Speech
PR	Precision-Recall
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency
TC	Text Classification (or Text Categorization)
TN	True Negatives
TNR	True Negative Rate
TP	True Positives
TPR	True Positive Rate
TTR	Types to Token Ratio

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1. Problem statement

One of the most important efforts of the investigating authorities has always been understanding criminal behavior, its motives, and its characteristics (Ainsworth, 2001) in order to solve, foresee or, ideally, prevent crimes. Despite many attempts to connect specific types of individuals with specific types of crime have been made, they were often unable to support these scientifically. In some cases, sociologists and psychologists were able to assist police authorities to analyze the recurring forms of specific crimes or advising police officers on what evidence they should collect from specific forms of crime in relation to a criminal's personality (Douglas et al., 1986). This has been the backbone of the criminal profile sketching method.

Forensic sciences, also known as criminalistics, is the application of scientific principles to provide physical evidence in criminal cases. For instance, forensic biology, which relies on DNA analyses, is one of the most revolutionary disciplines for the practice of crime scene investigations. Chemistry and physics support the inquiries with several methodologies aimed to accomplish tasks, such as revealing latent fingerprints, or identifying materials, etc. Forensic psychiatrists apply scientific and clinical expertise within a legal framework, by evaluating the competency of a defendant to stand trial, giving their opinion as expert witnesses, or giving mental state opinion of a defendant, etc.

During the last two decades, interest in Forensic Linguistics has greatly increased (Cotterill, 2003; Coulthard, 2004; McMenamin, 2002; Olsson, 2004) since language, as any other kind of evidence, can be used during police investigations and trial procedures. Forensic Linguistics concerns the study of written and spoken language mainly for legal purposes (Grant & Perkins, 2013). Thus, new scientific methods are applied to analyze testimonies given by suspects of criminal actions, since the evaluation of the witnesses' and defendants' profiles could be a determining factor in the trial procedure. Analyses are

often done for investigative purposes and when a specimen (e.g., a text, an email, an internet chat) is to be presented as evidence in court.

In Greece, research has been undertaken regarding the defendants from a sociological and psychological perspective (Gerolympoy, 1999; Kotsalis & Margaritis, 2007; Pitsogiannis, 1983; Rota, 2014), however there has been no research which is based on the natural language analysis of defendants and witnesses during a trial in order to extract information about their linguistic profile. Police interrogations and trial testimonies are recorded creating plentiful research material, however due to several legal, practical, and bureaucratic reasons, the acquisition of this material and its processing is almost impossible. Particularly, access to interrogation and trial records requires appropriate authorizations, which can be time consuming. Also, due to the sensitive data involved, the process can be delayed even further. Thus, until now, there has been no relevant research that analyzes the linguistic profile of defendants in Greek criminal courts. That means that so far there has been no collection of linguistic data, either compiled as written texts or as a transcription of recorded speech, of people involved in criminal proceedings in Greece. This leads to the fact that police investigations and court proceedings lack qualified and easily employable tools which can give them an additional help in testimony evaluation, as for example a text classifier would give based on the defendants' testimonies. Such a classifier should be seen as supplementary to judicial process and not a substitute for it.

1.2. Aims and research questions

The present doctoral thesis was born from the necessity of filling this gap, studying the linguistic profile of defendants accused of murder in Greek courtroom, taking into account the practical and procedural constraints of the Greek criminal legal system, and implementing a tool aimed to classify a defendant, guilty or not, based on his or her testimonies. Thus, the aim of this dissertation is, firstly, to construct a corpus of the spoken words of the defendants in front of the interrogator and inside the courtroom, secondly, to analyze the natural language of defendants through their testimonies in order to find linguistic patterns in their speech, and, finally, to develop a classification model from the available text data so as to decide whether an uncategorized testimony of a defendant has enough similarities to the linguistic patterns of perpetrators' testimonies,

so that the defendant can be classified as guilty or not. The latter will help us to achieve the development of a tool which can give support in testimony evaluation.

The main research questions in this study are grouped together in the following three paragraphs. Some of them existed from the beginning of our research and were the incentive for starting this thesis, while some others emerged during our research.

Our first concept was to construct a corpus composed of the words spoken from defendants accused of murder and testifying inside a courtroom, and the question was whether it was possible to construct a corpus from their testimonies. It was quite vague whether these data were accessible to us, and which ethical issues might arise regarding the processing of personal data. Another inquiry was what additional information we could extract from the transcripts which would enrich our research and thereby the research field of Forensic Linguistics in Greece. These procedures should be done without intervening in sensitive data, such as personal data, mainly names and addresses, of the people involved with the law.

Our second research query presupposes a positive answer to the first research question. In particular, the fulfillment of the corpus construction raises questions about how the defendants' speech inside the court can be quantified, in order to define their stylometric profile. In other words, we want to know if the defendants' speech follows some kind of linguistic patterns, which they are and how they may differ from general language. Having the corpus we mentioned above, it is interesting to enquire whether the speech of the defendants accused of murder differs from each other depending on their demographic and social characteristics. For instance, whether age, nationality or occupation could play a decisive role in the way they speak during their testimony in a court of justice. Another query is whether defendants use different linguistic patterns in front of a judge inside a courtroom and before their trial during their interrogation. Furthermore, we would like to know which are the linguistic differences between the defendants' testimonies inside the court with those of the witnesses.

Given that the two above research questions give us satisfactory answers, the following and more daring question is whether it is possible to develop a classifier, which can answer the question of whether a testimony belongs to a convicted murderer or not. This

thought emanates from our aspiration to offer an additional tool in evaluating a murderer's testimony before the criminal court decides whether the defendant is innocent or guilty. Surely, this question creates further queries such as how accurate the predictions of such a classifier could be, whether these predictions can be used to facilitate the investigative process and whether safe conclusions can be drawn.

At this point we should mention that only the crime of murder was chosen, due to the fact that there is enough material of murder cases in Greece to be processed, and because the verdict of a murder is either convicting or acquittal, which in computational language corresponds to a binary value, something that would help us in the development of a classifier.

1.3. Objective of the research

The focus of this dissertation is to analyze the linguistic patterns of the speech of defendants accused of murder inside a courtroom during their testimonies in order to support the judicial proceedings, since the use of computational techniques provides efficient, systematic, and precise information which is not possible by human judgement alone.

Discourse analysis can be applied in every field that deals with written, oral, or sign language and the area of law provides all three, containing written discourse and transcriptions of oral interactions that occur in a court of justice. Studies relevant to legal cases have been performed in the past (Fitzpatrick & Bachenko, 2009; Lidsky, 2000; Moens et al., 1999; Shuy, 2008; Vrij et al., 1997).

In Greece, this field of research suffers due to the difficulty of collecting data from judicial proceedings, or to develop automatic methods to identify stylometric linguistic characteristics. One of the first and most interesting research in this field in Greek is based on electronic textual resources of the proclamations of the terrorist organization '17th November' and on the apologies of its members in order to facilitate authorship identification (Frantzi, 2005, 2007, 2009). However, except for these few cases that occupied the public opinion and the media, the publication of trial proceedings is nonexistent. Thus, little research has been done which include testimonies collected in

natural environments, i.e., inside a courtroom in front of a judge, because until now there has been no relevant material of the Greek courts.

In this dissertation we addressed these limitations and we set ourselves the following objectives:

1. to collect a dataset in the context of criminal proceedings which would stem from the apologies of defendants accused of murder inside a courtroom in real-life conditions,
2. to construct a corpus which would consist of this dataset, i.e., real world text, suitable for performing language measurements and analyses,
3. to synthesize the linguistic profile of the defendants accused of murder,
4. to detect possible differences or similarities between the language profile of the defendants accused of murder and a reference corpus,
5. to predict the court's verdict of a defendant, i.e., guilty, or not guilty.

In order to accomplish the first objective, we applied for the trial briefs of murder cases in a Greek criminal court receiving all authorizations to have access to the data files and collect the data.

Our second objective, which was also the most time consuming, was achieved by creating a corpus of the defendants' testimonies inside the court, called GCDT (Greek Corpus of Defendants' Testimonies), a second corpus consisted of the witnesses' testimonies, called GCWT (Greek Corpus of Witnesses' Testimonies), and a third one which contained the testimonies of the defendants at the interrogation phase before their court summons, called pre-GCDT (pre-trial GCDT). Each time we were extracting the corresponding section of text that we were interested in, since the rest of the text was useless to us at that point of our study.

Our third objective was triggered by the notion that defendants might have common linguistic patterns during their testimonies inside a court. In order to ascertain if our speculation was grounded or not, we implemented quantitative analysis to the GCDT measuring several stylometric features that are widely used in authorship identification or author attribution research, which attempt to capture different shades of the personal style of the authors.

Our fourth objective was achieved by measuring the stylometric features of three reference corpora and comparing the results with the ones of GCDT in order to detect possible correlations. For the quantitative analysis we used three reference corpora successively, a general Greek language corpus, the GCWT that contained testimonies of witnesses from the respective cases of the defendants, and the pre-GCDT that contained testimonies of the same defendants before their trial in front of an interrogator.

Finally, we managed to fulfill our final objective by developing a machine learning classifier which would predict whether a testimony belonged to a murderer or not. This text classification was achieved by training our model with testimonies from GCDT in order to have training data from murderers and from GCWT in order to have training data from witnesses.

1.4. Thesis organization

The rest of this thesis[1] is organized as follows:

Chapter 2 is divided in three subsections. The first two concern the review of two distinct but correlated research areas in which our study resides, *Forensic Linguistics* and *Text mining* methods for discourse analysis. Firstly, we present a relevant history on the legal language, a review of the area being researched, i.e., discourse analysis in legal context, previous studies on crime profiling, and application domains of Forensic Linguistics. Moreover, we present current information surrounding corpus linguistics and, more specifically, we describe the extent to which previous studies have successfully investigated court language corpora, noting the gaps that our study attempts to address. In addition, stylometric approaches related to discourse analysis are described in which our stylometric analysis is based on. In the second part of this chapter, we provide the essential information regarding machine learning and text mining methods. In more detail, we describe a typical text classification approach using machine learning, helping

[1] This work is supported by APOLLONIS (http://apollonis-infrastructure.gr), the Greek Infrastructure for Humanities and Language Research and Innovation, and its ESFRI-related national research infrastructure CLARIN:EL (https://www.clarin.gr/en), the CLARIN-related Greek network for language resources, technologies and services.

the reader to understand the research problem and the significance of the results of our study. The third part of this chapter describes some related work from Greek and foreign researchers, regarding the compilation of Forensic Linguistics and text mining, and particularly regarding text mining and law language corpora, setting the context of our work.

Chapter 3 describes the procedure that we followed in order to collect the trial briefs and some preliminary investigations proceedings which were the source of our dataset. Among others we present the assumptions we had to make regarding our dataset. We explain the stages that a criminal case goes through until its trial inside a Greek courtroom. In addition, we describe what a trial brief contains, and finally we present the three corpora that we constructed from these trial briefs, namely the GCDT, the GCWT and the pre-GCDT.

In Chapter 4, we focus on quantifying the way defendants of murder speak, by studying several stylometric features of their language, either by comparing their speech depending on demographic data, or comparing their speech with reference corpora, such as the general Greek language and the language used by the witnesses. After the quantitative analysis of the characteristics of the defendants, we discuss the linguistic patterns of the defendants.

Chapter 5 introduces a text classification model that we built, which classifies the texts in two categories, as guilty or not guilty, purely based on verbal information contained in our corpora. We train our algorithm with the appropriate features derived from the stylometric study we made in the previous chapter, and we evaluate its results depending on several metrics of accuracy. We also present and discuss our algorithm's results.

In Chapter 6, we present the major findings of our thesis trying to explain their meaning. Moreover, we show the limitations of our findings, and we interpret any surprising or unexpected result.

Finally, Chapter 7 discusses the main conclusions and contribution drawn from this study and proposes possible future work directions.

2 BACKGROUND

2.1. Forensic Linguistics

Forensic Linguistics is the branch of linguistics which deals with forensic issues. It has spread its branches in several academic and research fields of study, starting its flourish in the nineties, with important articles on language and law (Gibbons, 1999; Rieber & Stewart, 1990), and books on the language of the courtroom (Solan, 1993; Stygall, 1994). Discourse analysis plays a significant role in the studies concerning legal language.

2.1.1. Legal language

Legal language is not a homogeneous discourse type but a set of related and overlapping discourse types. It has been referred as a specific field of Language of Specific Purposes since its content stems from a specific and specialized language and its objectives refer to a set of specialized needs (Trosborg, 1997). Gibbons (1999) writes: "Law is language. Laws are coded in language, and the processes of the law are mediated through language". In other sources, legal language is referred as Forensic Linguistics which is "the analysis of the language that relates to law, either as evidence or as legal discourse" (Olsson & Luchjenbroers, 2013). In an attempt to model the main structures and processes of law legal system and their associated discourses, legal language can be divided into a number of domains presumed to involve linguistic diversification (Maley, 1994).

Language of the law

The language of the law, that is, legal documents, can be divided in legislation and common law. The sources of legislation and the originating points of legal process are the legislative rules. They contain features of language and organization that are directly attributable to the pursuit of certainty. The linguistic forms of the legislative rule are selected so that they are explicit and precise (Maley, 1987). Explicitness means drafting a detailed and, if possible, exhaustive rule. The language of the rule refers to all the possible entities or actions to which the legislature intends the rule to apply. In terms of precision, legal drafting seeks "a degree of precision and internal coherence rarely met outside the language of formal logic or mathematics" (Dickerson, 1965). The language of

the law then, must be more precise than other styles of language. The language of legislation is written by the legislature and defends, among others, civil rights. Similar to the language of legislation is the language that is used in common law such as the regulations which are written by the authorities. Apart from the drafting of statutes the language of common law includes the contracting of agreements between individuals, such as wills, contracts, and deeds, the contracting of agreements between state and individuals, and the contracting of agreements between state authorities among them.

Pre-trial proceedings, trial proceedings and judicial judgements

Language as legal discourse includes, among others, the discourse inside the courtroom. The legal language can be divided in the professional language of law and the language of law encountered by the lay person (Dumas, 2007). In pre-trial procedure, legal language includes police and interrogator interview of the suspects. In this case, the individuals involved use legal language differently, depending on who the speaker is and on whom the speaker refers to. For instance, the police officer or the interrogator might use more structured and proficient language than the interviewee, and also the interrogator is likely to use more comprehensible and less formal language when addressing a suspect or a witness than if addressing a lawyer who is more familiar with the legal language.

The language of the courtroom varies according to the purpose of the communication. For example, in court proceedings' examination, the language that is used by a lawyer when he addresses a layman, either a client or a witness, is different from the language that is used when the lawyer addresses the judge. However, disparities in power are not limited to the police or interrogator examination. There is also a great disparity of power within the courtroom, between the legal professionals on the one hand, and the general public, particularly plaintiffs, defendants, and witnesses, on the other. This is a result of the use of the complex legal language. These disparities in power are both revealed and imposed through language (Gibbons, 1999). The language of the courtroom has its own features and rules in procedures, such as in re-examination, in intervention, in jury summation, in the final decision or when the judge declares the law.

After a trial, judicial judgments are written down and include the decision of the court and the trial proceedings. This form of legal language is quite heterogeneous since it contains transcripts of defendants' and witnesses' testimonies, judge declaration of the

law, judge and counsels' exchanges, counsels, and laymen (i.e., defendants, witnesses) exchanges, police reports, references to statutes, etc. This implies that judicial judgments cannot be analyzed linguistically as a single linguistic entity.

2.1.2. Legal discourse analysis

Discourse analysis deals with analyzing written, oral, or sign language. One of its main characteristics is that it can be applied in various contexts. Any continuous text, written or spoken, can be analyzed. The area of law, a highly written and verbal field, provides a fertile field for discourse analysis. It is generally regarded as a field containing written discourse, since all oral interactions that occur in court are recorded in printed form. Therefore, immense collections of both written text such as motions, counterclaims, and judges' opinions, and spoken words transcribed in writing, such as trial testimony, questioning, and argument, are preserved in written form (Shuy, 2008). Discourse analysis has been used in criminal cases yielding valuable knowledge in legal information extraction systems, i.e., locating information in texts by building a system that automatically abstracts Belgian criminal cases (Moens et al, 1999), for voice identification, defamation regarding the use of the name 'John Doe' in cyberspace (Lidsky, 2000), and mainly for outlining the profile of a criminal. Moreover, there is a large number of studies concerning the discourse analysis of law texts, such as Goodrich (1987) who examined the legal discipline and its concepts of language, text and sign, and constructed a theory of legal discourse as a linguistics of legal power, the book of Trosborg (1997) about the discourse analysis of statutes and contracts, showing that the discourse of English contract law selects patterns which are specific to the function of legal documents, namely regulation through legislation and common law, Bhatia et al. (2007) who studied the automatic analysis of lexicogrammar features, analysis of intertextuality and interdiscursivity in legal discourses, the research of Brousalis et al. (2012) who studied the application of discourse analysis to the language of Greek legislation and confirmed that factors such as the formulaic language, the preference to nouns and impersonal constructions, the use of technical vocabulary, and the length and complexity of sentences characterizing the Greek law texts, etc.

2.1.3. Crime profiling using text analysis

Criminal profiling, also referred to as offender or psychological profiling, designates a process of identifying specific characteristics of an individual committing a crime by a systematic observation and analysis of the crime scene, the victim, the forensic evidence, and the known facts of the crime (Chifflet, 2015). The profiling technique is used by behavioral scientists and criminologists to identify an unknown offender's significant personality and demographic characteristics through an analysis of their crimes, examine their criminal behavior and evaluate or even predict future criminal actions (Douglas, 1986; Davis, 1996). In other words, profiling is the process of drawing an offender's portrait from all available elements of the crime scene (Muller, 2000). Criminal profiling has raised immense popularity as both a topic of fascination for the general public as well as an academic field of study and scholarly attention has increased with various studies dealing with offender profiling (Dowden et al., 2007). However, some findings indicate no evidence for the assumption of a homology between crime scene actions and background characteristics and the homology assumption is too simplistic to provide a basis for offender profiling (Mokros, 2002). In case texts of a criminal are available, either written or transcribed from spoken, crime profiling can borrow techniques from other research fields such as author profiling, as described in the following subsection.

2.1.4. Application domains

Forensic linguists use large and structured set of spoken or written texts, namely corpora. These corpora include texts of suicide notes, mobile phone texts, police statements, police interview records and, in our case, defendants' testimonies and witnesses' statements. The following application domains can be implemented in the field of Forensic Linguistics with different degrees of reliability (Ariani, 2014).

Authorship analysis

Authorship analysis has two major approaches, i.e., author attribution and author characterization. Author attribution, also known as author identification, is the process of attempting to identify the likely authorship of a given document, given a collection of documents whose authorship is known (Stamatatos et al., 2000, 2016; Stammatatos et al., 2015). A set of documents with known authorship are used for training. The problem is

then to identify which of these authors wrote unattributed documents (Zhao, 2005). Therefore, author identification deals with classification problems and is directly related with the quantification of style of documents and more specifically the personal style of each author. The identification relies on analysis of author's idiolect, or patterns of language use such as vocabulary, collocations, pronunciation, spelling, grammar. The main attempts in authorship attribution research focused on defining features for quantifying writing style, research known as "stylometry" (Holmes, 1998). Hence, a great variety of measures including sentence length, word length, word frequencies, character frequencies, and vocabulary richness functions had been proposed. Authorship characterization attempts to formulate an author profile by making inferences about gender, education, and cultural backgrounds based on writing style. Authorship analysis is present in various applications (Stamatatos, 2009). The plethora of electronic texts, such as e-mails, blogs, online forum messages, source code, etc., has made the process of author recognition easy and fast, with a sharp increase in its application in various fields (Madigan et al, 2005; Rangel et al., 2018; Stammatatos et al., 2015) Some of them include matching messages or proclamations to known terrorists (Frantzi, 2009), applying authorship identification to extremist online messages (Abbasi & Chen, 2005), verifying the authenticity of suicide notes (Bennell, 2011; Shapero, 2011), identifying software plagiarism (El-Waned et al., 2007), recognizing copyright disputes (Adelsbach, 2003) and obtaining source code's author (Frantzeskou et al., 2006).

Our study has borrowed many of the measures that author identification proposes for quantifying the writing style, as we mentioned above, including sentence length, word length, word frequencies, character frequencies, and vocabulary richness, despite the fact that in our case we deal with spoken language.

Author profiling

Author profiling or characterization is the procedure of extracting information about the age, education, sex, etc., of the author of a given text (Burger et al., 2011; Koppel et al., 2002). Author profiling characterize authors by studying their sociolect aspect, i.e., how language is shared or how an author can be characterized from a psychological viewpoint (Rangel et al., 2018; Stammatatos et al., 2015). Author profiling is applied, among others, in forensics, security, and marketing. From a Forensic Linguistics' perspective, for

example, it is useful to learn about the linguistic profile of the author of a harassing text message and identify certain characteristics (Argamon et al., 2009). From a marketing viewpoint, companies may be interested to learn about the demographics of people who like or dislike their products, given blogs and online product reviews as analysis source (Abbasi et al., 2008).

The techniques of author profiling are applied in our study as we tried to learn about the linguistic profile of the speakers of our dataset (testimonies), i.e., the defendants accused of murder.

Forensic stylistics

Forensic stylistics is a subfield of Forensic Linguistics and it aims at applying stylistics to the context of author identification (Pavelec et al., 2007). Forensic stylistics is the study and interpretation of texts from a linguistic perspective (McMenamin, 1993). The basic claim of this approach is that every writer has his or her own linguistic patterns in unique combinations, and these patterns can be analyzed and described in aiming author identification. Stylistics can be classified into two different approaches, i.e., qualitative, and quantitative. Whereas the qualitative approach assesses errors and personal behavior of the authors, also known as idiosyncrasies, the second approach, which is very often referred as stylometry, is quantitative and computational, focusing on readily computable and countable language features, e.g., word length, phrase length, sentence length, vocabulary frequency, distribution of words of different lengths (Chaski & D, 2005; Tambouratzis et al., 2004). Apart from grammar, lexis, and semantics, stylistics is concerned with the examination of phonological properties and discursive devices as well (Simpson, 2004). There are some principles for individuality in stylistics features (Choudhary, 2018). Particularly, each matured writer has a handwriting which is personal and individual. Also, every writer has a unique style of using a language either in handwriting or verbal communication and every individual has his or her own distinctive characteristics which are unconsciously reflected in his or her handwriting.

Our study is based on the approach of stylistics that every writer, or speaker, has his or her own unique linguistic patterns and these patterns we would strive to analyze and describe. Thus, we would aim on the quantitative analysis of our dataset, focusing on readily computable and countable language features.

Linguistic dialectology

Linguistic dialectology refers to the scientific study of dialects (Chambers, 2004), and is a research area of sociolinguistics. It studies variations of language based mainly on their geographical distribution and related features. The study of dialects is also applied in court usually in the defendants and witnesses as they speak their own dialect and not the standard vocabulary that the legal representatives usually use. In case an interpreter is present between the interviewer and the interviewee, the study of the dialect becomes even more important.

In our case, where almost a quarter of the testimonies, which constitute our dataset, belongs to non-native speakers, dialectology could be applied, since we also had the recorded apologies of the non-native defendants in order to study variations of the language based mainly on geographical distribution. Unfortunately, the testimonies we have in our hands are already translated by an authorized interpreter.

Forensic phonetics

Forensic phonetics focuses on the analysis of spoken communication for the needs of criminal justice. It includes speaker identification, enhancing and decoding spoken messages, analysis of emotions in voice, authentication of recordings (Hollien, 2012). It deals with the production of accurate transcriptions of what was being said. The recent progress in acoustic engineering gave a boost in the study of forensic phonetics and established its presence in the forensic research area. Phoneticians can analyze the distinctive speech characteristics of a speaker relative to other candidate speakers in an inquiry. Forensic Phonetics examines aspects of recorded speech and offers opinions based on the observations arising from the analysis. Transcriptions can reveal information about a speaker's social and regional background (Olsson, 2004).

Although our dataset, stems from spoken language, we do not have the recordings of the testimonies in order to study the forensic phonetics. Thus, we approached our dataset as if it stemmed from written text.

Forensic transcription

Forensic transcription includes transcriptions of spoken words to written documents. This work is that of court stenographers, who take shorthand notes and transcribe them

into a written text, which becomes, after appropriate checking, the official version of the proceedings (Fraser, 2014). Text transcription should be accurate and reliable in order to become powerful evidence in criminal trials. Alongside, the introduction and rapid spread of audio-recording technology gave the opportunity to transcribe the speech captured in an audio or video recording in written form. Transcripts are frequently used for research purposes (Bucholtz, 2007; Heselwood, 2013), and specifically in forensic phonetics (French & Stevens, 2013; Shuy, 1993; Turell et al., 2008).

In our study, the testimonies that compose our dataset, are transcriptions of spoken words inside a courtroom to written documents. This procedure has been conducted from authorized secretaries, who are obliged to write down what they hear word by word. As a result, the transcripts are considered accurate, However, in subsection 3.3 we present some limitations and assumptions regarding the transcriptions' process.

Intra-author variation

Intra-author variation, i.e., the variation within one author's work is a field under study of Forensic Linguistics. Sometimes, the intra-author variation is higher than the variation of texts by two different authors, known as inter-author variation (Olsson, 2004). This perception raises many questions about author attribution. An assessment of the intra-author variation is difficult to obtain (Nini, 2013). Thus, a strong theoretical framework for authorship analysis should be introduced, in order to solve the problem of theoretical validity (Grant & Baker, 2001).

In our study, we could study possible intra-author variations since for most of the defendants we have both their testimonies in front of a judge and in front of an interrogator. However, in this study we focus on the characteristics of the speech of defendants as a genre.

2.1.5. Corpus linguistics

In language sciences a corpus is a collection of written texts or transcribed speech which can serve as a basis of linguistic analysis and description (Kennedy, 2014). Corpora are fundamental to corpus linguistics as an empirical endeavor. They form the basis of analysis and provide data for hypothesis-testing, language model construction, exemplification, and empirical grounding (Kirk, 1996). In corpus linguistics, the term

'corpus' covers a "large and principled collection of natural texts" (Precht et al., 1998). Corpora are built so that the representativeness of the language, sublanguage, special language they describe is achieved (McEnery & Wilson, 2003). Some definitions suggest that corpora necessarily consist of structured collections of text. Others indicate that corpora can consist of whole texts or collections of whole texts. There is a distinction between a corpus and a text archive or text database. A corpus is designed for linguistic analysis and normally is a systematic, planned, and structured compilation of text, whereas a text archive is an unstructured text repository (Leech, 1991). General text archives typically do not qualify as corpora but are seen as databases (Baker et al., 2006; Gries, 2009). A corpus can be analyzed and compared with other corpora to study variation.

In recent times the meaning and use of words has been extended using corpus-based techniques. Corpus linguistics is an area that focuses upon a set of procedures and methods for studying language (McEnery & Hardie, 2011). Corpus is described as a large body of linguistic evidence composed of attested language use (McEnery, 2019). Corpus can be both spoken and written. The choice of corpus depends on the research question and the chosen application. The set of texts or corpus dealt with is usually of large-scale size that requires the use of a machine-readable text.

The first machine-readable corpus, that rocketed corpus linguistics into the digital era, was W. Nelson Francis and Henry Kučera's Brown Corpus of written American English, which was completed in the early 1960s (Svartvik, 2007). Advances in computer technology have made possible the collection and storage of very large corpora from a variety of sources and computers have facilitated the analysis of these corpora (Precht et al., 1998). Corpus linguistics has evolved in tandem with computer technology and is linked to the computer which has introduced speed, accountability, accuracy, statistical reliability, and the ability to handle huge amounts of data. Computers not only allowed for storage and the processing of increasingly massive amounts of data, but they also enabled increasingly complex quantitative analysis, which is integral to the study of language use. Thus, common tasks of corpus-based analysis, like word frequencies, concordances, collocate and keywords, can be completed within a couple of minutes. While early corpus analysis consisted of word counting which required huge amounts of

processing by building-sized computers in nearly inaccessible computer labs in university basements, corpus linguists can now perform advanced statistical analyses on their laptops at home or in their offices, using platforms such as R (Gries, 2009), Python (Bird et al., 2009), or Perl (Hammond, 2003).

These technological advances have boosted corpus-based applications. For instance, Natural Language Processing (NLP) which includes a wide set of syntax, semantics, discourse, and speech tasks, uses corpus data as the raw data for several applications. Corpus linguistics makes it possible to identify the meaning of words by looking at their occurrences in natural contexts, common or uncommon words, patterns between words and non-linguistic factors, collocations and the use and distribution of synonyms. Corpus-based studies guarantee precision and completeness involving the processing of real language material (O'Keeffe, 2010).

2.1.6. Court language corpora

The existence of corpora for Forensic Linguistics' purposes, and mainly corpora from speech language containing defendants' or witnesses' testimonies is limited. This can be attributed to the difficulty of collecting such data due to issues of personal data protection and access to sensitive data. Due to the lack of forensic corpora, researchers are often forced to create their own 'laboratory corpora' in order to study the effectiveness of their methods and tools. Obviously, such corpora cannot have the potential of 'real language' corpora. Some researchers have described a set of guidelines for acquiring and developing corpora of court data (Fitzpatrick & Bachenko, 2012).

Regarding the Greek language, until now, there is no such corpus since the publication of trial proceedings is almost nonexistent. Interesting research in this field in Greek is in the notices of the terrorist Greek organization "November 17th" and their correlation with the testimonies of its members (Frantzi, 2007; 2009). Moreover, recent research of written and spoken courtroom discourse in military justice is published which attempts to identify, analyze, and address the main issues that affect them (Kapopoulos, 2021). Older publications concern the criminal proceedings of those accused as responsible for Regime of the Colonels, a far-right authoritarian military junta that ruled Greece from 1967 to 1974, have also been published as a book (Voultepsis, 1975). Moreover, the

proceedings of the Trial of the Six, which was the trial for treason, in late 1922, of the Anti-Venizelist officials held responsible for the Greek military defeat in Asia Minor, have been published as a book (Trial of the Six, 1976).

As for the English language, there are specific trial proceedings published on the Internet, regarding notorious trials such as these of O. J. Simpson (Cotterill, 2002, 2003; Fisher, 1997; Iegorova, 2018), of bomber Timothy McVeigh (Linder, 2011), and serial killer Harold Shipman (Smith, 1966). Harris (2001) examined the nature and structure of witness and defendant narrative accounts in the evidential portions of courtroom trials, using the trials of O.J. Simpson, Oklahoma Bombers, and Louise Woodward as a database, proposing a means of distinguishing narrative from non-narrative accounts, and using a model to analyze a series of representative example narratives taken from the trial data. Matoesian (2005) examined a questioning strategy in trial cross examination designed to control an evasive witness. The data segment that was used came from the William Kennedy Smith rape trial, a famous media trial that occurred in 1991, and concerns a defense attorney's cross examination of a witness. Galatolo (2005, 2006) studied the functions of Direct Report Speech (DRS) in legal testimonies, investigating the witnesses' answers to questions posed during direct and cross-examination. Her analysis focused on the evidential and moral function that DRS had, particularly, on lay witnesses. The data used in her study were taken from an Italian criminal trial, a murder case, that had attained a good deal of notoriety. A work focusing on real-life data is that of Fornaciari who had created a corpus in real life conditions which was the first corpus of deceptive Italian texts, not relying on material created in laboratory conditions but of language material collected in a natural environment, to create models for distinguishing true from false statements (Fornaciari & Poesio, 2013). Another related work presented a dataset consisting of truthful video clips, from real court trials, using the transcription of those videos to extract several linguistic features (Pérez-Rosas, 2015). Lee (2010) explored court interpreters' renditions of reported speech in Korean language contained in witnesses' evidence. The data of her study was formed by audio recordings of court proceedings.

2.1.7. Criminal's language stylometry

The personal style of the authors is based on frequent patterns found in their texts and the extraction of this stylistic information from documents is quantified based on a wide variety of measures that are used for stylistic purposes. This thesis attempts to represent the general properties of the criminal's language style by combining two objectives, i.e., the criminal and the author profiling, studying thoroughly the methods used by these two research areas. Recent approaches for authorship attribution and author profiling have been examined in a comprehensive survey (Stamatatos, 2009), in which characteristics for both text representation and text classification focusing on computational requirements are evaluated. Another relevant research presents the most distinctive stylometric characteristics, concluding that legal texts have a distinct and highly recognizable stylometric profile (Broussalis et al., 2012). The measures that are used commonly in author identification and forensic stylistics are described below (Abbasi et al., 2008; Vel, 2000; Zheng et al., 2006).

- Lexical: A text is considered as a sequence of tokens grouped in sentences, so these features are token-based. Lexical features can be further divided into word-based and character-based and features. Examples of word-based measures are sentence/word length counts, 10-word frequencies, stop word frequencies, n-grams of words, vocabulary richness measures, etc. Even though these features are easy to extract in most cases, are not suitable for some natural languages, such as Chinese, or for some text's domains consisting of multiple abbreviations or acronyms, such as e-mail messages and tweets. Character-based features include n-grams of characters, related alphabetic characters count, digit characters count, uppercase and lowercase characters count, letter frequencies, punctuation marks count, compression models, etc. Although the character-based information is easy to be extracted in any natural language, the dimensionality of this representation is considerably increased.
- Syntactic: this category can capture an author's writing style at the sentence level and includes function words, punctuation, and Part-of-Speech (POS) frequencies. The discriminating power of syntactic features is derived from people's different habits of organizing sentences. This type of information requires especially robust and accurate NLP tools (POS taggers, syntactic parsers, etc.) to analyze the documents. The

extraction of syntactic features is language-dependent relying on the availability of NLP tools for a specific natural language. However, the use of NLP tools increases the computational cost and POS tagging is still immature for some languages such as Chinese.

- Structural: these features represent the way an author organizes the layout of a piece of writing. This type of information refers to the logical structure of sentences and the relationships between different concepts. Examples of these features are total number of lines, total number of sentences, total number of paragraphs, number of sentences per paragraph, number of characters/words per paragraph, etc.
- Content-specific: Content-specific features are important keywords and phrases pertaining to certain topics. Content-specific keywords can be used to better capture the properties of an author's style within a particular text domain. For example, content-specific features on a discussion of crime may include the words 'police' and 'kill'. In case that all texts to be analyzed are on the same thematic area, content-based information may reveal some authorial choices. In more detail, given that the texts in question deal with certain topics and are of the same genre, one can define certain words frequently used within that topic or that genre.

The main types of features used in authorship attribution to capture the writing style of an author are the lexical ones which are easy to extract. When a deep linguistic analysis of texts is required, one should use more sophisticated and language dependent features. In this study, we used lexical, syntactic, and content-specific features (Table 2.1.1).

Type		**Stylometric features**
Lexical	character-based	# 2-grams
	word-based	# total words most frequent words average word length (in characters) average sentence length (in words) TTR (types to token ratio) hapax & dis legomena
Syntactic		frequency of function words frequency of content words lexical density

		functional density
Content-specific		frequency of keywords

Table 2.1.1 Adopted stylometric features in this study

2.2. Text Mining

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data (Witten & Frank, 2002). The idea is to find regularities or patterns from databases by seeking them automatically with the aid of computer programs. If strong patterns are found, they could be generalized to make accurate predictions for future data. Data mining methods can be applied to any kind of data in both structured and unstructured form (Aggarwal, 2015; Allahyari et al., 2017). Likewise, text mining, a subfield of data mining, is about looking for patterns in text. It is the process of analyzing text to extract information that is useful for particular purposes. Text is unstructured, amorphous, and difficult to deal with. Nevertheless, text is the most common medium for the information exchanges, thus the motivation for trying to extract information is compelling. Text mining focuses on the discovery and extraction of proper and non-trivial knowledge or patterns from a collection of text documents, such as emails, blogs, articles, HTML files, etc. (Hotho et al., 2005).

2.2.1. Machine learning

Machine learning provides the technical basis of data mining. It is used to extract information from the raw data in databases that is expressed in a comprehensible form and can be used for a variety of purposes. It is about techniques for finding and describing structural patterns in data. This research field is dedicated to the study and the understanding of the learning systems' function and it is seen as a subset of Artificial Intelligence (AI). Applying AI means building better and intelligent machines. Machine learning aims to build algorithms able to improve automatically with the experience they gain during their execution. It is a constantly evolving field interacting with applications and sciences such as statistics, engineering, computer science, cognitive science, etc. (Jordan & Mitchell, 2015; Sebastiani, 2002).

The basic function principle of a machine learning algorithm is the attempt to derive generalized rules, capable of dealing with the problem to be solved from a limited set of training data. The only available information is the limited training sample. Therefore, it is a process of Inductive Reasoning, in other words the derivation of general principles from specific observations (Copi, 2006). Therefore, the only thing that is guaranteed is that the target function is accurate for the training data, while for any other instance we can only assume. Thus, the fundamental assumption of inductive learning is that anyone involved in this field can benefit from examples using them as training data, capture characteristic functions and then predict various models with relative accuracy. Something extremely interesting, which has not been proven but works in all cases is the fact that any assumption which can approach the target function in a wide range of training data, will approach it in unknown instances as well (K. Sai Prasad, 2020; Witten et al., 2002).

Most of the data is unstructured, that is audios, videos, photos, documents, graphs, etc., and finding patterns in data is almost impossible for human brains. Also, data is already very massive and the time to compute it increases continuously. Machine learning can help people with significant data in minimum time.

2.2.2. Supervised learning

One type of machine learning is called supervised learning. Supervised learning algorithms are designed to learn by example (Russell & Norvig, 2010; Shams & Mercer, 2016). During training of a supervised learning algorithm, the training data consist of inputs paired with the correct outputs. If inputs are given with the corresponding correct outputs, then the learning is called supervised, in contrast to unsupervised learning, where inputs are unlabeled. In contrast to supervised learning, unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data. Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, machine is restricted to find the hidden structure in unlabeled data on its own.

The supervised learning algorithm will search for patterns in the data that correlate with the desired outputs. After training, the algorithm can accept new unseen inputs, determining which label the new inputs will be classified as, based on prior training data. The objective of a supervised learning model is to predict the correct label for newly presented input data. The function used to connect input features to a predicted output is created by the machine learning model during training.

Supervised learning is sometimes called Classification learning, since the learning data are presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples. The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The method operates under supervision by being provided with the actual outcome for each of the training examples. This outcome is called the class or the category of the example. Another type of supervised learning algorithm is called Regression, which is a predictive statistical process where the model attempts to find the important relationship between dependent and independent variables. The goal of a regression algorithm is to predict a continuous number, instead of a category, such as sales, income, and test scores. In this study, we are interested in classification learning algorithms, as we will analyze below.

2.2.3. Text classification

Classification is the type of supervised learning in which labelled data are used, and these data are used to make predictions in a non-continuous form, in contrary to regression which makes predictions in a continuous form. In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. During training, a classification algorithm is given data inputs with an assigned category or class (Wilson, 2019). The purpose of a classification algorithm is to take a new input value and assign it a class that it fits into, based on the training data provided. More specifically, text classification or text categorization, is the process of classifying the texts and assigning tags to natural language texts within a predetermined set of categories.

Typically, text categorization is the task of assigning a Boolean (true or false) value to each pair $(d_j, c_i) \in D \times C$, where D is a domain of documents and $C = \{c_1, \ldots, c_c\}$ is a set

of predefined categories. A true (T) value corresponding to a pair indicates a decision that the document d_j is assigned under the category c_i, while a false (F) value indicates a decision not to assign d_j under c_i (Sebastiani, 2002). Therefore, the goal is to approximate the unknown target function that describes how documents ought to be classified using a function called the classifier.

Instead of relying on manually crafted rules, text classification with machine learning learns to make classifications based on past observations. By using pre-labeled examples as training data, a machine learning algorithm can learn the different associations between pieces of text and a particular output, namely class or tag, is expected for a particular input, that is, a text. Text data is the simplest form of data which is unstructured in nature. Humans can clearly perceive and process unstructured text data, but it is difficult for machines to understand the same. This voluminous text data is an important source of knowledge and information. Therefore, to use the information extracted from text data effectively, methods and algorithms are needed.

Binary text classification

The case in which exactly one category is assigned to each document is often called single label. A special case of single label text classification is binary text classification, in which each document must be assigned either to category c_i or to its complement $\bar{c}_i$, i.e., a document is classified into one of two mutually exclusive categories or classes. Binary classification is the simplest and most widely studied case and can be extended for solving multi-class problems. It is noteworthy since this thesis is based on binary text classification techniques. The two mutually exclusive categories in our study are 'guilty' and 'not guilty', which we will analyze in detail in subsequent sections.

2.2.4. Data preprocessing

The process of text classification using machine learning techniques has been described thoroughly (Ikonomakis et al., 2005) and its simplified version, slightly modified, is depicted schematically in Figure 2.2.1.

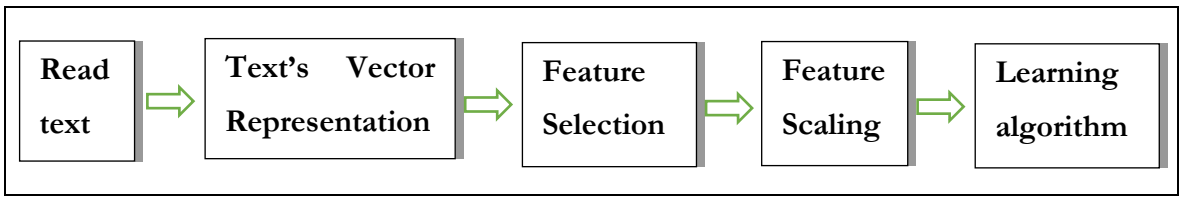


Figure 2.2.1 Text classification process

The data preprocessing phase is crucial in text classification since it brings the under-investigation text into a form that is predictable and analyzable from the machine learning algorithm. The basic stages are the vector representation of the text, the feature selection, and the feature scaling, i.e., normalization or standardization.

Vector representation of text

A text is a sequence of words, so each text is usually represented by an array of words. The set of all the words of a training set constitutes the feature set. In order to facilitate the processing of the feature set from a machine learning classifier, the words or phrases from a text are mapped to vectors of real numbers, through a set of language modeling and feature learning techniques, a procedure known as word embedding.

A simplifying representation used in NLP and Information Retrieval (IR) is called Bag-of-Words (BOW) model where a text is represented by its occurrence of words within the document (Deepu et al., 2016). Another approach represents a text by a binary vector, assigning the value 1 if the text contains the feature or 0 if the feature does not appear in the text. Another technique, the Term Frequency-Inverse Document Frequency (TF-IDF) model, contains information on the more important words and the less important ones as well (Zheng et al., 2019).

Not all the words presented in a document can be used to train the classifier (Madsen et al., 2004). There are several ways to reduce the size of the initial feature set. Thus, it is very common that some text preparation methods are used to remove information from documents and facilitate further processing. Typical pre-processing steps concern the removal of punctuation marks and special characters, the removal of stop words (e.g., auxiliary verbs, conjunctions, and articles), the removal of misspelled or words with the same stem, the use of the word's occurrence frequency instead of a Boolean indicator of

whether the word occurred in the text, etc. When topic of documents is concerned, information about punctuation marks and function word usage is not crucial. However, if style of documents is concerned, such as in our study, this kind of information can be very useful and for that reason, although we removed the punctuation marks, we kept every function word of our dataset.

Stemming is another common preprocessing step. To reduce the size of the initial feature set, misspelled or words with the same stem are removed keeping the stem or the most common of them as feature. For example, the words "kill", "killing", "killer" and "kills" can be replaced with "kill". However, there are some doubts on the actual importance of aggressive stemming (Sebastiani, 2002). In this study we did not proceed to stemming, since we were interested in the frequency of the POS separately, as well as the use of different tenses.

Great savings in training resources are made with the representation of the feature value (Leopold & Kindermann, 2002). Often a Boolean indicator of whether the word occurred in the document is sufficient. Other possibilities include the count of the number of times the word occurred in the document, a technique we adopted in our dataset, the frequency of its occurrence normalized by the length of the document, the count normalized by the inverse document frequency of the word. In situations where the document length varies widely, it may be important to normalize the counts.

Feature Selection

The first step towards training a classifier with machine learning is to extract features. Feature selection is the automatic or manual selection of those features which will contribute most to the prediction output in which we are interested in. Extracting the important features is a vital technique in dimensionality reduction (Beil et al., 2002; Khalid et al., 2014). The correct selection of features should target to overfitting reduction (see Subsection 5.1), accuracy improvement and training time reduction. There are several automatic methods for feature selection, depending on the variable type (numerical or categorical) of the input and output. Methods for feature selection use an evaluation function that is applied to a single word (Soucy & Mineau, 2003).

The so-called Best Individual Features methods use scoring of individual words, which can be performed using some of the measures, for instance, document frequency, term frequency, mutual information, information gain, odds ratio, chi-square, and term strength (Forman, 2003). On the contrary, Sequential Forward Selection methods firstly select the best single word evaluated by given criterion (Montañés et al., 2003) and then they add one word at a time until the number of selected words reaches desired k words. The widely adopted approach in text classification is the filtering approach based on scoring the features, sorting them according to this score and selecting a predefined number of the best ones. In our case, the feature selection was made using the filtering approach manually, scoring the features by taking into consideration which of the stylometric features contributed the most i.e., were indicative of a linguistic profile.

Feature Scaling

Feature scaling is a crucial part of the data preprocessing stage. In case the dataset features have different scales or units, there is a chance that higher weightage is given to features with higher magnitude. This will impact the performance of the machine learning algorithm and obviously, the algorithm will be biased towards one feature. Therefore, the feature scaling, before employing a machine learning algorithm, contributes to the objectivity of the result (Juszczak et al., 2002).

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling, since it is the most common technique of normalization. The formula for Min-Max scaling is:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

X_{max} and X_{min} are the maximum and the minimum values of the feature, respectively. When the value of X is the minimum value in the column, the numerator will be 0, hence X' is 0. On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator, thus the value of X' is 1. If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1 as well.

2.2.5. Text classification algorithms

After the features' extraction, the documents can be easily represented in a form that can be used by machine learning algorithms. The machine learning algorithm is fed with training data that consists of pairs of feature sets (vectors for each text input) and tags or classes, to produce a classification model (Figure 2.2.2)[2].

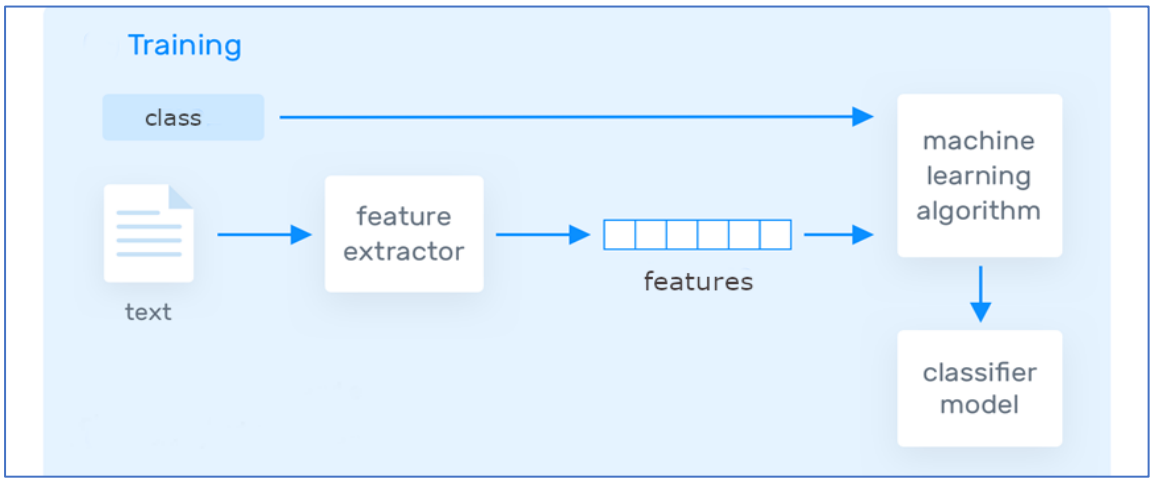


Figure 2.2.2 Classification training phase

Once the machine learning model is trained with enough training samples, it can begin to make accurate predictions. The same feature extractor is used to transform unseen text to feature sets which can be fed into the classification model to get predictions on tags. For example, in our case a tag is either guilty or not guilty. Figure 2.2.3[3] represents schematically the classification prediction phase.

[2] Retrieved from: https://monkeylearn.com/text-classification/

[3] Retrieved from: https://monkeylearn.com/text-classification/

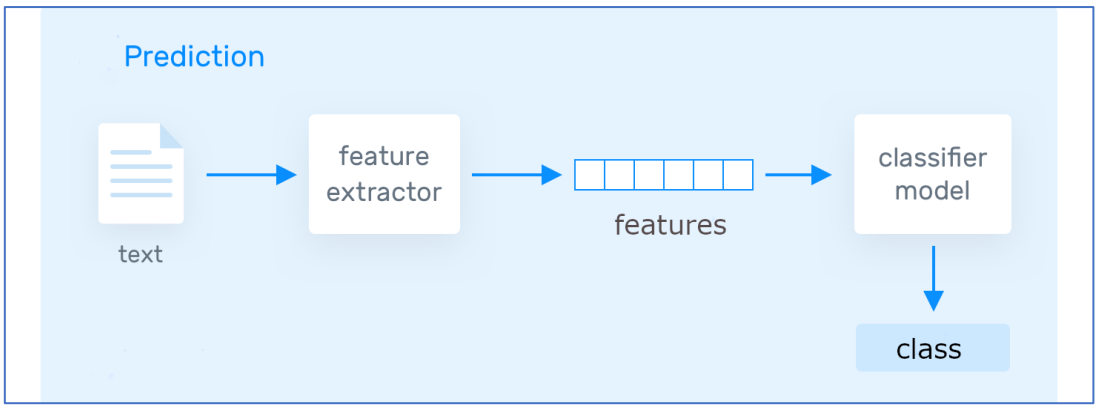


Figure 2.2.3 Classification prediction phase

Apart from manual, automated classification APIs (Application Program Interface) are also used to categorize the key texts in a document to utilize the important words.

In our case the purpose of classification algorithm is to determine if a testimony belongs to a murderer or not. Considering that a murderer is guilty, and a witness is not guilty, the algorithm can predict if a person is guilty or not by analyzing his or her testimony. This problem is called a binary classification problem, since the algorithm has two classes to choose from (guilty, or not guilty). The algorithm is given training data with testimonies that belong both to murderers and to witnesses. The model will find the features within the data that correlate to each class and create the function $Y=f(x)$, where Y is the predicted output that is determined by a mapping function that assigns a class to an input value x. Then, when provided with a new testimony, the model will use this function to determine whether it belongs to a murderer or not.

Some examples of classification problems are speech recognition, handwriting recognition, bio metric identification, document classification, etc. Classification problems can be solved with a numerous number of algorithms (Caruana & Niculescu-Mizil, 2006). Whichever algorithm is chosen depends on the data and the situation (Kotsiantis et al., 2007).

In order to check which classification algorithm performs better on our problem or what configurations to use, we tested 6 different algorithms: Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Classification and Regression Trees, Gaussian Naive Bayes, Support Vector Machines.

Logistic Regression

Logistic Regression (LR) is used when the dependent variable (target) is categorical and is intended for binary classification problems. For example, to predict whether an email is spam (1) or not (0), or in our case whether a text belongs to a guilty person (1) or not (0). It is a predictive analysis method and is used to describe data and to explain the relationship between one dependent binary variable (target) and one or more nominal, ordinal, interval or ratio-level independent variables (input). LR assumes no error in the output variable, thus outliers and possibly misclassified instances should be removed from the training data. It is a linear algorithm with a non-linear transform on output and it assumes a linear relationship between the input variables with the output. Like linear regression, the model can overfit if there are multiple highly correlated inputs (Tabachnick et al., 2007).

However, LR has limitations that suggest at the need for alternate linear classification algorithms. For instance, LR is intended for two-class or binary classification problems. It can be extended for multi-class classification but is rarely used for this purpose. Also, LR can become unstable when the classes are well separated or when there are few examples from which to estimate the parameters.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) does address each of the LR's limitations and is the suitable linear method for multi-class classification problems. LDA is the preferred linear classification technique if there are more than two classes, but it works even with binary-classification problems. LDA can work as a dimensionality reduction technique and as a classifier algorithm. The characteristics of the dataset will guide a researcher about the decision of applying LDA as a classifier or a dimensionality reduction algorithm to perform a classification task.

The main of LDA is basically separate example of classes linearly moving them to a different feature space, therefore if a dataset is linear separable, LDA can be applied as a classifier. However, if the dataset is not linear separable the LDA will try to organize the dataset in another space as the maximum linearly separability as possible, but it still be examples overlapping between classes because of non-linearly characteristic of data. In this case, the use of another classification model should be applied to deal with nonlinear

data such as neural network with multiple hidden layers, neural network with radial basis function or SVM with nonlinear Kernels. LDA assumes that each input variable has the same variance (Balakrishnama et al., 1999).

In this study we used LDA algorithm for performing text classification in our dataset, since it performed the best results comparing to other classification algorithms.

K-Nearest Neighbors

The k-nearest neighbors (k-NN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The k-NN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. This algorithm focuses on the detection of the most similar documents with the one in question. Then, the test document is assigned to the category most of its k most similar training documents belong to. The inputs consist of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors, where k is a positive integer. If k equals to 1, then the object is simply assigned to the class of that single nearest neighbor (Beyer et al., 1998). It is crucial to calculate the similarity among documents with an appropriate measure to make k-NN robust with noisy data (Larose, 2005). In addition, k-NN requires a high cost in the application phase (Hand et al., 2001). Lim proposed a method which improves performance of k-NN based text classification by using well estimated parameters (Lim, 2004).

Classification and Regression Trees

Classification and Regression Trees (CART) is a term used to describe decision tree algorithms that are used for classification and regression learning tasks. The representation for the CART model is a binary tree. Each root node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric). The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. Given a new input, the tree is traversed by evaluating the specific input started at the root node of the tree.

The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Classification trees are designed for dependent variables that take a finite number of unordered values and regression trees are for dependent variables that take continuous or ordered discrete values (Loh, 2011).

Gaussian Naive Bayes

The Naive Bayes (NB) classifier is a family of simple probabilistic classifiers based on a common assumption that all features are independent of each other, given the category variable, and it is often used as the baseline in text classification. The assumptions on distribution of features are called event models of the NB classifier (Xu, 2018). When dealing with continuous features, a typical assumption is Gaussian distribution. It remains a popular method for text categorization, i.e., judging documents as belonging to one category or the other, with word frequencies as the features. It is often used in text because of its simplicity and effectiveness (Kim et al., 2002). Its performance is often degraded because it does not model text well. However, some of its problems can be solved by some simple corrections (Schneider, 2005). With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines.

Support Vector Machines

A Support Vector Machine (SVM) is also a supervised machine learning model that uses classification algorithms for two-group classification problems (Cortes & Vapnik, 1995). After giving an SVM model sets of labeled training data for each category, it can categorize a new text. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space, where N is the number of features, that distinctly classifies the data points (Joachims, 1998). To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The goal is to find a plane that has the maximum margin, i.e., the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these

support vectors, the margin of the classifier is maximized. SVM algorithms provide excellent precision, but poor recall.

2.2.6. Hyperparameters tuning

Machine learning algorithms involve a number of hyperparameters that have to be set before running them. In contrast to model parameters, which are determined during training, these tuning parameters have to be carefully optimized to achieve maximal performance (Probst et al., 2019).

An approach is to objectively search different values for model hyperparameters and choose a subset that results in a model that achieves the best performance on a given dataset. This is called hyperparameter optimization or hyperparameter tuning.

In order to select an appropriate hyperparameter configuration for a specific dataset one can resort to default values of hyperparameters that are specified in implementing software packages or manually configure them, for example, based on recommendations from the literature, experience or trial-and-error.

Alternatively, one can use hyperparameter tuning strategies, which are data-dependent, second-level optimization procedures (Guyon et al., 2010), which try to minimize the expected generalization error of the inducing algorithm over a hyperparameter search space of considered candidate configurations, usually by evaluating predictions on an independent test set, or by running a resampling scheme such as cross-validation (see Subsection 2.2.8). However, the tuning strategies range from simple grid or random search (Bergstra & Bengio, 2012) to more complex, iterative procedures such as Bayesian optimization (Bischl et al., 2017) or iterated F-racing (Lang et al., 2017).

In this study we use the grid search technique due to its simplicity in the implementation and parallelization, and its reliability in low dimensional spaces. Grid search exhaustively enumerates all combinations of hyperparameters and evaluates each combination. Depending on the available computational resources, the nature of the learning algorithm and size of the problem, each evaluation may take considerable time. Thus, the overall optimization process is time consuming. In our case, the grid search was the best choice considering the simplicity versus the time of the evaluation.

2.2.7. Evaluation

The evaluation metrics commonly used in text classification have their origin in Information Extraction which preluded the use of machine learning in automated text processing and understanding (Boyer & Lapalme, 1985). Several studies have analyzed systematically the performance measures used in the complete spectrum of machine learning classification tasks (M & M.N, 2015; Sokolova & Lapalme, 2009).

Performance measures for classification

The correctness of a classification can be evaluated by computing the following metrics:

- True Positive (TP): when a case was positive (1), and it was predicted positive (1) and is equal to the number of correctly recognized class cases.
- True Negative (TN): when a case was negative (0), and it was predicted negative (0) and is equal to the number of correctly recognized cases that do not belong to the class.
- False Positive (FP): when a case was negative (0), but it was predicted positive (1) and is equal to the number of cases that were incorrectly assigned to the class.
- False Negative (FN): when a case was positive (1), but it was predicted negative (0) and is equal to the number of cases that were not recognized as class examples.

In case of binary classification these four metrics constitute a confusion matrix (Table 2.2.1) which shows the distribution of correct and wrong prediction over the two classes. It is a way of tabulating the number of misclassifications, i.e., the number of predicted classes which ended up in a wrong classification based on the actual classes. On y-axis confusion matrix has the true values, and on the x-axis the values given by the predictor.

		Predicted class	
		0 (false)	**1 (true)**
Actual class	**0 (false)**	TN	FP
	1 (true)	FN	TP

Table 2.2.1 Confusion matrix for binary classification

The most often used metrics for binary classification based on the values of the confusion matrix are accuracy, precision, recall, F1-score, specificity and the area under the curve (AUC), which are depicted in Table 2.2.2.

Metric	**Formula**
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall (Sensitivity or True Positive Rate)	$\frac{TP}{TP + FN}$
F1-score	$2 * \frac{Precision * Recall}{Precision + Recall}$
Specificity	$\frac{TN}{TN + FP}$
False Positive Rate	$\frac{FP}{FP + TN}$
AUC	$\frac{Recall + Specificity}{2}$

Table 2.2.2 Metrics for binary classification

Accuracy

Accuracy shows the overall effectiveness of a classifier. It is the most intuitive performance measure, and it is simply a ratio of correctly predicted observations to the total observations. Accuracy is a great measure but only when we have symmetric datasets where values of false positive and false negatives are similar. Therefore, we must look at other parameters to evaluate the performance of our model.

Precision

Precision, also called positive predictive value, answers to the question of what percent of the classifier's predictions are correct. For each class precision is defined as the ratio of true positives to the sum of true and false positives.

Recall

Recall, also known as sensitivity or true positive rate, answers to the question of what percent of the positive cases can the classifier catch. Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

F1-score

F1-score answers to the question of what percent of positive predictions are correct. It gives the harmonic mean of precision and recall.

Specificity

Specificity, also referred to as the True Negative Rate (TNR), answers to the question of how effectively a classifier identifies negative labels. It is the proportion of samples that test negative using the test in question that are genuinely negative.

Area Under the Curve and Receiver Operating Characteristic

In order to explain the Area Under the Curve (AUC) metric, we firstly describe what a Receiver Operating Characteristic (ROC) curve is (Figure 2.2.4). The performance of any binary classifier can be depicted in the space defined by True Positive Rate (TPR) and False Positive Rate (FPR), called the ROC. A ROC curve is typically used in binary classification to study the output of a classifier. It is a graphical plot that illustrates the performance of one classification model at all decision thresholds. It can be used to evaluate the strength of a model. The diagonal line of Figure 2.2.4[4] serves as a reference line since it is the ROC curve of a diagnostic test that randomly classifies the condition. It is called No-skill model does not have any ability to distinguish between the two classes and therefore, TPR = FPR at any decision threshold. The top left corner of the plot is the ideal point because the FPR equals to zero, and the TPR equals to one.

[4] Retrieved from https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

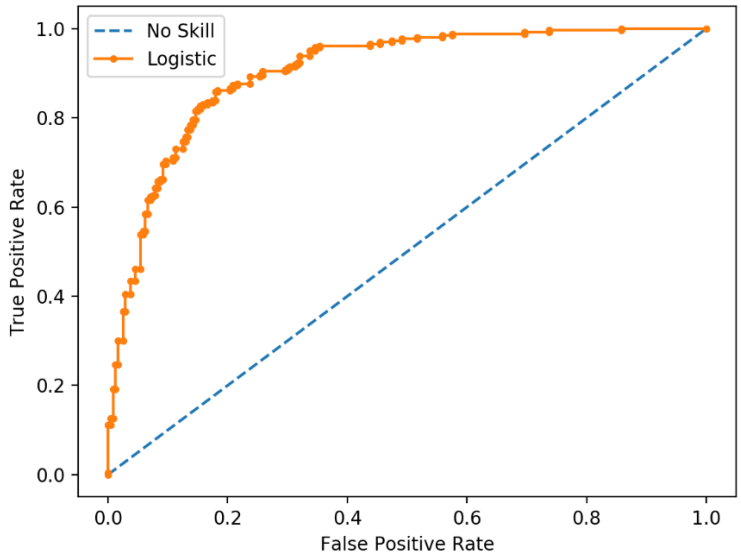


Figure 2.2.4 ROC curve of a logistic regression model and a no skill classifier

Smaller values on the x-axis of the plot indicate lower FP and higher TN. Moreover, larger values on the y-axis of the plot indicate higher TP and lower FN. Several points in ROC space are important to note. The lower left point (0,0) represents the strategy of never issuing a positive classification. Such a classifier commits no FP errors but also gains no TP. The opposite strategy, of unconditionally issuing positive classifications, is represented by the upper right point (1,1). Informally, one point in ROC space is better than another if it is to the northwest of the first. Classifiers appearing on the left-hand side of a ROC graph, near the X axis, may be thought of as 'conservative', because they make positive classifications only with strong evidence, so they make few false positive errors, but they often have low true positive rates as well. Classifiers on the upper right-hand side of a ROC graph may be thought of as 'liberal' because they make positive classifications with weak evidence, so they classify nearly all positives correctly, but they often have high false positive rates (Fawcett, 2006).

ROC curves can also be used to compare two or more models. Typically, a ROC curve illustrates TPR, or Sensitivity, on the Y axis, and FPR, or 1-Specificity, on the X axis. That is, each point on the ROC curve represents a different decision threshold (cutoff value). The points are connected to form the curve. Cutoff values that result in low FPR tend to result low TPR as well. As the TPR increases, the FPR increases. The better the

diagnostic test, the more quickly the TPR reaches 1 (or 100%). Figure 2.2.5[5] depicts a ROC curve of a random and of a perfect classifier.

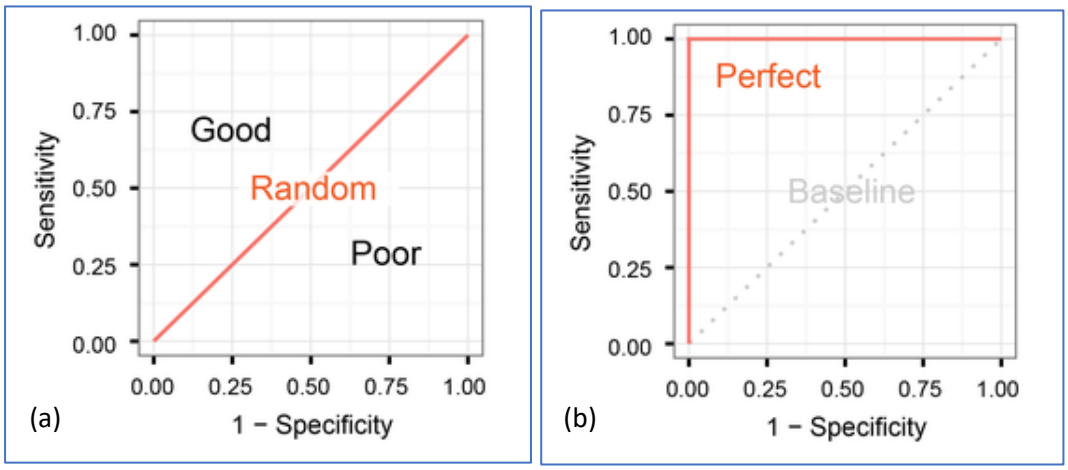


Figure 2.2.5 A ROC curve of a (a) random classifier, (b) perfect classifier

A near-perfect diagnostic test would have an ROC curve that is almost vertical from (0,0) to (0,1) and then horizontal to (1,1). A model with perfect skill is represented by a line that travels from the bottom left of the plot to the top left and then across the top to the top right. The point (0,1) represents perfect classification, i.e., the closer the ROC curve of a classifier is to the upper left corner of the ROC space (FPR=0, TPR=1), the more effective the classifier is.

The area under the curve (AUC) can be used as a summary of the model skill. It measures the entire two-dimensional area underneath the entire ROC curve from point (0,0) to point (1,1). A model with no skill is represented by a diagonal line from the bottom left of the plot to the top right and has an AUC of 0.5. If a classifier has greater AUC than another one, then it has better average performance, too. For instance, in Figure 2.2.6[6] shows four AUC scores. The score is 1.0 for the classifier with the perfect performance level (P) and 0.5 for the classifier with the random performance level (R).

[5] Retrieved from https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot/

[6] Retrieved from https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot/

ROC curves clearly show classifier A outperforms classifier B, which is also supported by their AUC scores (0.88 and 0.72, respectively).

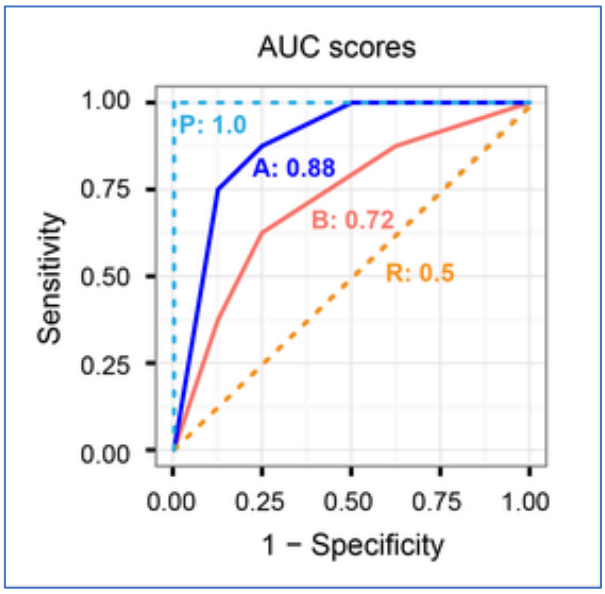


Figure 2.2.6 The AUC scores under ROC curves

Area Under the Curve and Precision-Recall curve

Apart from the above metrics, we refer to Precision-Recall curve (PR Curve) which is also a diagnostic tool that helps in the interpretation of probabilistic forecast for binary classification predictive modeling problems, and it is used in cases where there is an imbalance in the observations between the two classes (Branco et al., 2015). This metric is used for our classifier evaluation, since one class of the dataset observations is 42% greater than the other one.

A PR curve is a plot of the precision (y-axis) and the recall (x-axis) for different probability thresholds.

Figure 2.2.7[7] shows a PR curve of a random classifier which is depicted as a straight line equal to P / (P + N), where P the positives and N the negatives. A random classifier (no-skill) line changes depending on the distribution of the positive to negative classes. For

[7] Retrieved from https://classeval.wordpress.com/introduction/introduction-to-the-precision-recall-plot

instance, the line is y = 0.5 when the dataset is balanced and thus the ratio of positives and negatives is 1:1, whereas 0.25 when the ratio is 1:3.

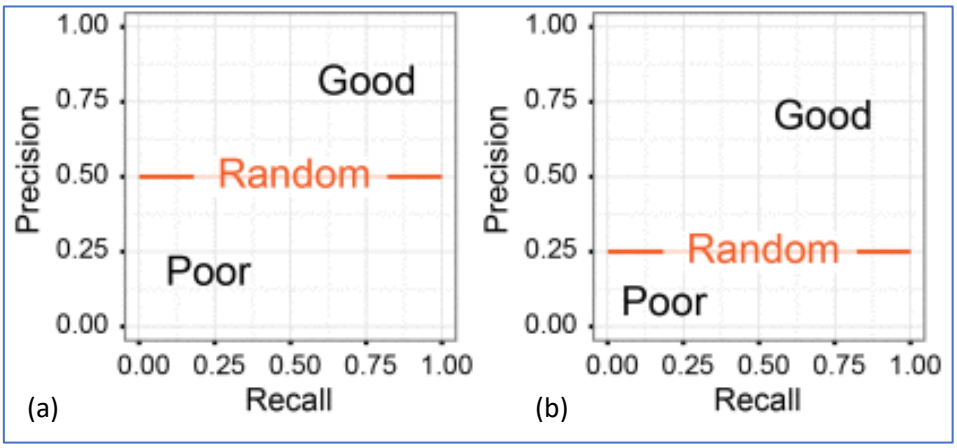


Figure 2.2.7 PR curve of a random classifier when the ratio of positives and negatives is (a) 1:1 and (b) 1:3.

This line separates the PR space into two areas. The separated area above the line is the area of good performance levels. The other area below the line is the area of poor performance.

Respectively, Figure 2.2.8[8] shows a PR curve of a perfect classifier which is depicted as combination of two straight lines, i.e., from the top left corner (0.0, 1.0) to the top right corner (1.0, 1.0) and further down to the end point (1.0, P / (P + N)). The end point depends on the ratio of positives and negatives. For instance, the end point is (1.0, 0.5) when the ratio of positives and negatives is 1:1, whereas (1.0, 0.25) when the ratio is 1:3.

[8] Retrieved from https://classeval.wordpress.com/introduction/introduction-to-the-precision-recall-plot/

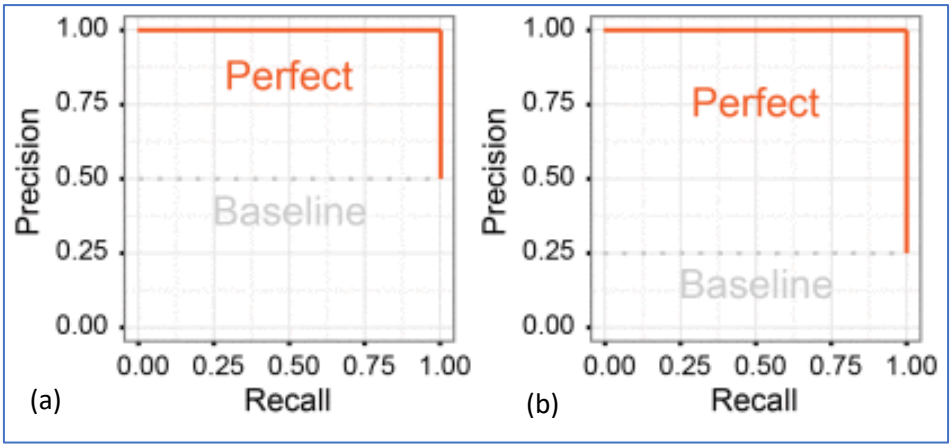


Figure 2.2.8 PR curve of a perfect classifier when the ratio of positives and negatives is (a) 1:1 and (b) 1:3.

A skillful model is represented by a curve that bows towards a coordinate of (1,1) (Figure 2.2.9[9]). A no-skill classifier is one that cannot discriminate between the classes and would predict a random class or a constant class in all cases. It is easy to compare several classifiers in the PR plot. Curves close to the perfect PR curve have a better performance level than the ones close to the baseline. In other words, a curve above the other curve has a better performance level. Similar to ROC curves, the AUC (the area under the precision-recall curve) score can be used as a single performance measure for PR curves. As the name indicates, it is an area under the curve calculated in the PR space. It summarizes the curve with a range of threshold values as a single score. The score can then be used as a point of comparison between different models on a binary classification problem. Although the theoretical range of AUC score is between 0 and 1, the actual scores of meaningful classifiers are greater than P / (P + N), which is the AUC score of a random classifier, with a score of 1.0 represents a model with perfect skill.

[9] Retrieved from https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

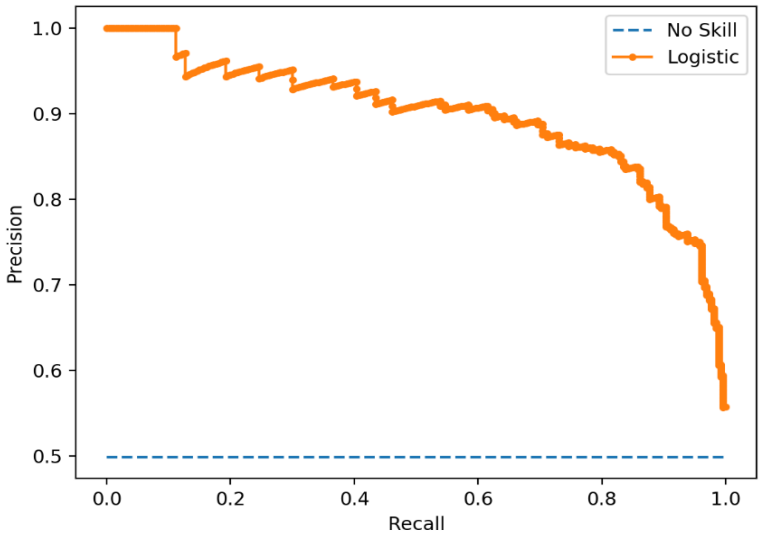


Figure 2.2.9 PR curve for a logistic regression model and a no skill classifier

Generally, ROC curves should be used when there are roughly equal numbers of observations for each class and PR curves should be used when there is a moderate to large class imbalance (Saito & Rehmsmeier, 2015). A ROC curve and a PR curve should indicate the same performance level for a classifier. Nevertheless, they usually appear to be different, and even interpretation can be different. In addition, the AUC scores are different between ROC and PR for the same classifier (Davis & Goadrich, 2006).

2.2.8. Validation

Validation techniques in machine learning are used to get the error rate of the machine learning model. If the training data volume is large enough to be representative of the statistical population, we may not need the validation techniques. However, as we work with samples of training data validation techniques seems to be mandatory. Model validation is the process of evaluating a trained model on test data set. This provides the generalization ability of a trained model. Using proper validation techniques can help us to understand our model. The main validation techniques that are used in machine learning and we used in our model are described below.

Hold-out

In case all the data is used for training the model and the error rate is evaluated based on the outcome versus the actual value from the same training data set, this error is called

the re-substitution error and this technique is called the re-substitution validation technique. To avoid the re-substitution error, it is used the holdout validation technique which is the most common method (Yadav & Shukla, 2016).

In this method, the given data is split into two different datasets labeled as a training and a testing dataset. This can usually be an 80/20 (i.e., 80% of the data is used as training dataset and the rest 20% of the data is held back and it is used as the testing dataset) or 70/30 or 60/40 split. The classifier fits a function using the training set only. Then the output values are predicted for the data in the testing set. The errors it makes are accumulated to give the mean absolute test set error, which is used to evaluate the model. The advantage of this method is that it is usually fast to compute. However, its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the test set, and thus the evaluation may be significantly different depending on how the division is made. In this case, there is a likelihood that uneven distribution of different classes of data is found in training and test dataset. To fix this, the training and test dataset is created with equal distribution of different classes of data. This process is called stratification.

K-fold cross-validation

A technique known as cross-validation is to perform multiple evaluations on different test sets and then to combine the scores from those evaluations (Yadav & Shukla, 2016; Zhang & Yang, 2015). An advantage of using cross-validation is that it allows us to examine how widely the performance varies across different training sets. If we get very similar scores for all N training sets, then we can be confident that the score is accurate. On the other hand, if scores vary widely across the N training sets, then we should probably be skeptical about the accuracy of the evaluation score.

More specifically, k-fold cross validation is one way to improve over the holdout method. Generally, k-fold cross-validation is conducted to verify that the model is not over-fitted. The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other 'k-1' subsets are put together to form a training set. Then the average error across

all 'k' trials is computed (Figure 2.2.10)[10]. It is necessary to automatically find a trade–off between the percentage of data used to train the classifier, and the tightness of the estimated error (Anguita et al., 2012). Every data point gets to be in a test set exactly once and gets to be in a training set 'k-1' times. The variance of the resulting estimate is reduced as 'k' is increased. The error rate of the model is average of the error rate of each iteration. The disadvantage of this method is that the training algorithm must be re-run from scratch k times, which means it takes k times as much computation to make an evaluation. This technique can also be called as a repeated hold-out method. The error rate could be improved by using stratification technique.

Figure 2.2.10 K-fold cross-validation scheme

A variant of this method is to randomly divide the data into a test and training set k different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over. Furthermore, even though the individual folds might be too small to give accurate evaluation scores on their own, the combined evaluation score is based on a large amount of data and is therefore quite reliable.

[10] Image from http://karlrosaen.com/ml/learning-log/2016-06-20/

Stratified k-fold approach is a variation of k-fold cross-validation that returns stratified folds, i.e., each set containing approximately the same ratio of target labels as the complete data. Thus, the splitting of data into folds ensures that each fold has the same proportion of observations with a given categorical value, such as the class outcome value.

2.3. Text Mining and Law Language Corpora

In this section, we refer to some key prior work in text mining and law corpora, outlining the range of its approaches to set the context of our work. The following studies deal with the process of legal texts with the aid of text mining techniques in order to automate a specific procedure for the benefit of law professionals. The main objective of these studies was the categorization of trial decisions to automatically identify legal arguments, properties, and relationships among them. Thus, legal professionals can identify relevant cases and material in the corpus of trial decisions. These studies have many similarities with our study regarding the text mining algorithms, although our corpora derived from the real-time testimonies of defendants and witnesses inside a court, whereas the studies which are presented below used corpora derived from the legal language of court decisions which present a more structured speech.

2.3.1. Worldwide related work

Text classification methods have been successfully applied to several NLP tasks and applications. The legal domain of law professionals would greatly benefit from the possibility of automation provided by machine learning. Regarding to text mining of arguments, Wyner et al. (2010) described relevant approaches using text-mining to automatically profile and extract arguments from legal cases. Another indicative study which investigated the extent to which one can automatically identify argumentative sentences in legal text, their argumentative function and structure, is that of Mochales-Palau and Moens (2009) who used a corpus containing legal texts extracted from the European Court of Human Rights (ECHR), classifying argumentative vs. non-argumentative sentences with an accuracy of 80%. Another recent study proposed a computational method to predict decisions of the ECRH, where textual information is

represented using contiguous word sequences, i.e., N-grams, and topics, with an accuracy of 79% on average (Aletras et al., 2016).

Another model was developed to extract cases which are relevant to the current case in terms of the comments on the quality of the case, e.g., whether it has been appealed, affirmed, overturned, overruled, explained, or distinguished, focusing on automated extraction of citation relations, not on argumentation or case factors (Jackson et al., 2003). Similarly, Boella et al. (2011) proposed a classification approach which identified the relevant domain to which a specific legal text belongs, using TF-IDF weighting and Information Gain for feature selection and SVM for classification, attained F1-score of 76% for the identification of the domains related to a legal text and 97.5% for the correct classification of a text into a specific domain.

An automatic summarization of court rulings was presented, based in Canadian court rulings, where the introduction, context, reasoning, and conclusion were found to be independent of the ruling itself (Farzindar & Lapalme, 2004). A system of classifying sentences for the task of summarizing court rulings was proposed, using SVM and Naive Bayes applied to BOW, TF-IDF, and dense features (e.g., position of sentence in document), obtaining 65% F1-score (Hachey & Grover, 2006). Similarly, the study of Gonçalves and Quaresma (2005) used BOW, POS tags, and TF-IDF to classify legal text in 3000 categories, based on taxonomy of legal concepts, and reported 64% and 79% F1-score. Sulea et al. (2017) presented a study in order to predict the accuracy of the ruling of the French Supreme Court and the law area to which a case belongs to, applying machine learning techniques and reporting results of 98% F1-score in predicting a case ruling, 96% F1-score for predicting the law area of a case, and 87.07% F1-score on estimating the date of a ruling.

Wyner and Milward (2008) developed text mining tools to automatically search for elements that are found in commercial case law search engines, such as indices for citation index, judges, jurisdiction, and so on. The second objective of their study was to develop searches for features of the case beyond those found in such search engines, such as case features or the identification of violation of some norm.

Finally, another study collected a new open domain deception dataset which contained demographic data such as gender and age in order to explore deception, gender and age detection in short texts using a machine learning approach. The feature sets' extraction included n-grams, shallow and deep syntactic features, semantic features, syntactic complexity, and readability metrics. After building classifiers that aim to predict deception, gender, and age, the study showed that deception detection can be performed in short texts even in the absence of a predetermined domain, but gender and age prediction in deceptive texts was a challenging task (Perez-Rosas & Mihalcea, 2015).

2.3.2. Greek related work

The goal of the research in Greek related work was to identify studies relevant to our work. Although there are several studies which associate text mining with law corpora, these corpora consist mainly of legal terms derived from online Greek legal databases, in contrast to our work that is based, exclusively, on language spoken from laymen inside a courtroom.

Research of Koutsogoula (2014) analyzed court decisions in order to extract the factors that are used as input data for the training of an AI model, so that it could predict the litigation outcome of public-works claims. The analysis was made in 34 court decisions, which were collected from the Greek online platform NOMOS, by extracting the factors that define their litigation outcome. The factors were chosen in order to have only two possible values, true or false. The conclusions of this study showed that it was difficult to construct an AI model, with little complexity which could foresee the courts decisions on all types of claims from the execution of technical works and will be characterized by high prediction accuracy.

Another study which processed a specific number of Greek legal documents aiming to create from scratch a synonym dictionary with terms of legal interest, extracted the data from legal information database "Νομοτέλεια". Target of the specific study was the clustering of documents in order to assess the dictionary's effectiveness, using text mining techniques (Niforas, 2016).

The subject of another study based on Greek legal documents was the management of information in sources of legal decisions. In particular, the study involved the extraction

of header summarization, labeling of legal references and categorization of legal decisions drawn from Areios Pagos, Supreme Court of Greece. The author created an automated process that detected reports from each category and produced a vector indicating the frequency of each category. The vectors were then used as inputs for various neural network models (Katsampos, 2015).

3 THE DATASET

In this chapter we present the corpora that we built from scratch using real trial data. In total we created three corpora, i.e., a corpus with testimonies from defendants accused of murder, a corpus with witnesses' testimonies and a corpus with testimonies from the defendants during the pre-investigation procedure. Before presenting our corpora, we describe the content and origin of our dataset, its characteristics, several obstacles we had to overpass, and assumptions we had to make.

3.1. Context Description

In order to determine the context that our data belongs and to clarify the procedure followed for their collection, we describe the process that a trial of a felony goes through in the Greek court, the content of the trial briefs, and the genre of the testimonies.

3.1.1. Felony hearings in Greek court

There are specific stages of investigation in Greece before a criminal case goes to court: the preliminary examination, the preliminary criminal investigation, and the main criminal investigation. During the preliminary examination, the prosecutor determines whether a complaint is well-founded and whether an offense has been committed. Preliminary criminal investigation is carried out if the suspect has been arrested red-handed or if there is an immediate danger due to a delay. Main criminal investigation is carried out only in case of serious offenses, such as felonies. Main criminal investigation takes place with a view to establishing, collecting, and preserving evidence as well as obtaining evidence of the crime. The criminal investigation or interview is conducted solely by the Investigating Magistrate and is written down by an authorized secretary. Following this procedure, it is decided whether the defendant should stand trial or be released. The three-member Criminal Court of Appeal or the Mixed Grand Jury are competent to deal with cases of felonies while the final decision is made by judges and the jury at the Mixed Grand Juries.

Hearings in criminal courts of justice follow a predefined procedure which depends on rules determined by the Code of Criminal Procedure. This means that the procedure of

every hearing follows specific regulations, providing an opportunity to collect data in relatively homogeneous conditions, even when the cases differ. The protagonist of each hearing is the defendant who gives the testimony. He or she answers the questions posed by the judge, the public prosecutor, the jury, and the defendant's lawyer. The defendant cannot be absent from any hearing and so the testimonies have the form of a dialogue. It is possible that other individuals intervene, for example more than one public prosecutor, or more than one defendant lawyer, defense witnesses and prosecution witnesses.

Apart from the defendant, another person that gives information about a crime is the witness either a prosecution or a defense one. A witness gives his or her own testimony in front of the judge, the public prosecutor, the jury, and the defendant's lawyer, and answers the questions posed by any of them. It is possible that a hearing has no witnesses at all.

The defendant, the defense witnesses and the prosecution witnesses are interviewed by the judge, the jury, the prosecutor, and the defense lawyer. In the Greek Court of law, all the exact words of the testimonies are written down word by word by an authorized trial secretary. An implementation of a project, referred to as the Integrated Court Transcripts System (ICTS)[11], has now been developed which aims to the digital recording, archiving and distribution of court hearing transcripts and the digital recording, archiving and distribution of the hearing transcripts of the courts of appeal (civil and criminal), the courts of first instance, and the courts of peace throughout the country, hereinafter. The project includes both the procurement and installation of all necessary equipment as well as the provision of services for the transcription of the recorded court hearings.

3.1.2. Trial Briefs

A trial brief is written by a trial secretary at the responsibility of the judge conducting the hearing. A trial brief shall state the place, time of the hearing and its breaks, as well as the time set for each repetition, the names of the judges, of the prosecutor and of the secretary; the name and anything else contributes to the identification of the litigants,

[11] https://www.ospd.gr/

their representatives and lawyers, the names of witnesses, interpreters, experts, and technical advisers, and the swearing in of witnesses, interpreters, and experts.

A trial brief must contain briefly the testimonies of the witnesses and the additions or differences of the statements made at the hearing from those made at the interrogation, as well as the conclusions of the experts and technical advisers, the testimonies and statements of the defendants and technical advisers, the proposals and requests of the prosecutor and the litigants, the decisions of the court and the provisions of the person conducting the hearing and generally any significant event during the hearing. Whoever leads the hearing makes sure that those parts of the testimonies or statements that he deems essential for the purposes of the evidence are recorded verbatim. In felony trials the observance of the trial briefs by voice recording is mandatory[12].

3.1.3. Testimonies and direct speech

As it mentioned above, in the Greek Court of law all the exact words of the testimonies are written down word by word by an authorized trial secretary. For instance, if a defendant says, 'I am not sure', the secretary is expected to write this down as 'I am not sure' without adding a reporting clause such as 'the defendant says'. A rendition such as 'The defendant is not sure' would be easily perceived as information diffusion. Since evidence law, which is concerned with the reliability and fairness of evidence presented, requires precision and exactness in witnesses' testimony, the trial secretary is expected to strive to convey the original utterances as faithfully as possible.

In interpreted courtroom examination, what the defendant or witness presents in reported speech, namely what was allegedly uttered in the 'reported event', is interpreted by the interpreter into the language the court understands, and the interpreted evidence is the evidence the court hears, and the trial secretary writes down, with a layer of complexity added to the already hetero-linguistic discourse in the courtroom (Lee, 2010). Thus, the requirement of a verbatim rendering is a legal stipulation designed to minimize

[12] http://www.opengov.gr/ministryofjustice

any interference by the interpreter, whose institutional role can influence and determine the outcome of the case.

We denote that the testimonies which comprise our dataset use DRS which retains the form of the original utterance. Defendants and witnesses give evidence of a conversation in the first person, namely in direct speech. In particular, the bulk of testimonies stems from lay defendants and witnesses, thus they use free direct speech, meaning that their speech includes also slang words, which have been kept intact, maintaining the exactness, and meaning of their words.

3.2. Data Collection

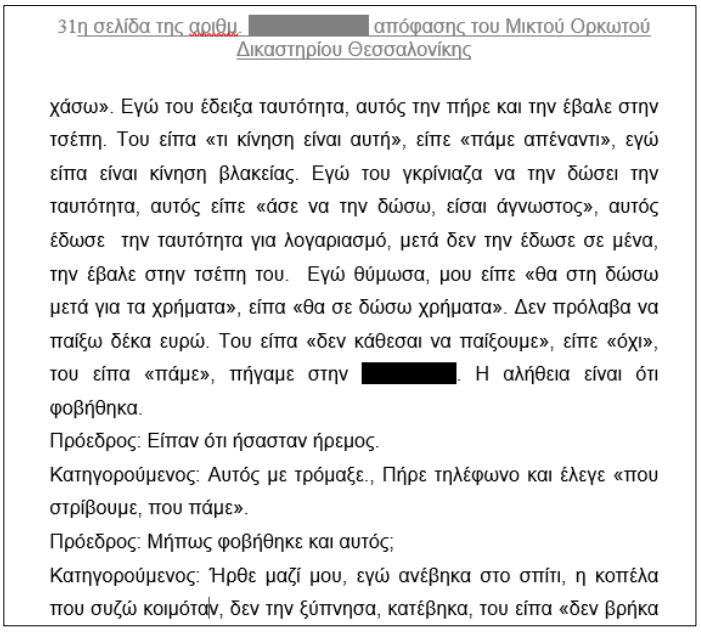
In order to study the testimony language in the Greek Court of law, we built corpora coming from testimonies collected in real conditions. The aim was to find testimonies that involved several subjects in the criminal case with adequate length of speech. Our corpora concern testimonies regarding murders.

The collection of such language material was a challenging task. To collect this kind of data, contacts have been made with a Court of Law in the Greek city of Thessaloniki. The aim was to examine all the relevant documents to extract the texts for our scientific purposes. All authorizations have been received in order to have access to the data files and collect the data. From our side, there was the assurance of using the receiving information in anonymous form, out of respect and a legal obligation for the privacy of the subjects involved.

The trials were held between 2008 and 2015 in Thessaloniki, a Greek city. Before 2008 all the trials briefs in the Court of Law of Thessaloniki were kept in analog form, i.e., manuscripts written on a typewriter or, in the worst case, by hand, thus their processing would be difficult or almost impossible.

The bulk of the data is in digital form and consists of the relevant trial briefs, i.e., all the documents related to the trial of a case. In Figure 3.2.1 it is depicted a sample of a trial brief in digital form, where there are questions of the presiding judge ('Πρόεδρος') and answers by the defendant ('Κατηγορούμενος'). We erased every sensitive information from the image, i.e., the number of the trial brief and a person's name.

The rest of the data that we retrieved were in analog form (i.e., manuscripts) and concern the defendants' testimonies during their criminal investigation which takes place after the defendants' arrest and before their trial. Some of the manuscripts had been typed clearly enough and printed so they could be scanned and converted in a digital form (text files) with an OCR (Optical Character Recognition) application. We used ABBYY FineReader[13] program which allows the conversion of image documents into editable electronic formats, i.e., text files. However, some manuscripts were handwritten, so they had to be typed in digital form manually.



31η σελίδα της αριθμ. [redacted] απόφασης του Μικτού Ορκωτού Δικαστηρίου Θεσσαλονίκης

χάσω». Εγώ του έδειξα ταυτότητα, αυτός την πήρε και την έβαλε στην τσέπη. Του είπα «τι κίνηση είναι αυτή», είπε «πάμε απέναντι», εγώ είπα είναι κίνηση βλακείας. Εγώ του γκρίνιαζα να την δώσει την ταυτότητα, αυτός είπε «άσε να την δώσω, είσαι άγνωστος», αυτός έδωσε την ταυτότητα για λογαριασμό, μετά δεν την έδωσε σε μένα, την έβαλε στην τσέπη του. Εγώ θύμωσα, μου είπε «θα στη δώσω μετά για τα χρήματα», είπα «θα σε δώσω χρήματα». Δεν πρόλαβα να παίξω δέκα ευρώ. Του είπα «δεν κάθεσαι να παίξουμε», είπε «όχι», του είπα «πάμε», πήγαμε στην [redacted]. Η αλήθεια είναι ότι φοβήθηκα.

Πρόεδρος: Είπαν ότι ήσασταν ήρεμος.

Κατηγορούμενος: Αυτός με τρόμαξε., Πήρε τηλέφωνο και έλεγε «που στρίβουμε, που πάμε».

Πρόεδρος: Μήπως φοβήθηκε και αυτός;

Κατηγορούμενος: Ήρθε μαζί μου, εγώ ανέβηκα στο σπίτι, η κοπέλα που συζώ κοιμόταν, δεν την ξύπνησα, κατέβηκα, του είπα «δεν βρήκα

Figure 3.2.1 A sample of a trial brief in digital form

Figure 3.2.2 is a photo of a pre-trial testimony sample in handwritten form where every sensitive information had been erased.

In this pre-trial sample one can see two questions of the interrogator ('ΕΡΩΤΗΣΗ') and the answer of the defendant ('ΑΠΟΚΡΙΣΗ'). The questions and answers are filled out,

[13] https://pdf.abbyy.com/

written by hand, in real time as the interrogator cross-examines the accused person and he or she testifies.

Hearings in Court are events strongly formalized with rules determined by the Code of Criminal Procedure. The main character of each hearing is the subject who gives the testimony.

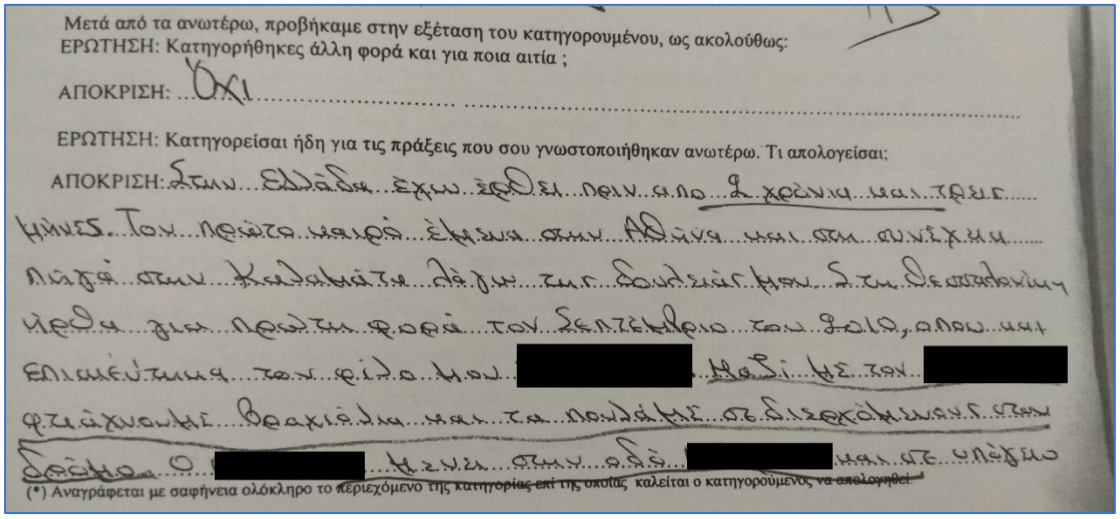

Μετά από τα ανωτέρω, προβήκαμε στην εξέταση του κατηγορουμένου, ως ακολούθως:
ΕΡΩΤΗΣΗ: Κατηγορήθηκες άλλη φορά και για ποια αιτία ;
ΑΠΟΚΡΙΣΗ: Όχι
ΕΡΩΤΗΣΗ: Κατηγορείσαι ήδη για τις πράξεις που σου γνωστοποιήθηκαν ανωτέρω. Τι απολογείσαι;
ΑΠΟΚΡΙΣΗ:
(*) Αναγράφεται με σαφήνεια ολόκληρο το περιεχόμενο της κατηγορίας επί της οποίας καλείται ο κατηγορούμενος να απολογηθεί

Figure 3.2.2 A sample of a pre-trial testimony in handwritten form

In our case the subject is either a defendant or a witness. Due the type of the specific cases (felonies), apart from the aforementioned persons, there are four jurors who can set their questions as well. Therefore, the considered testimonies have the form of a dialogue, in which at least eight individuals are present. It is possible that other persons intervene, for example more than one public prosecutor, or more than one defendant lawyer, or a lawyer for the victim of the crime, or a police officer, etc.

The procedure of creating our corpora was time consuming due to two major reasons. Firstly, as we mentioned in a previous section, a trial brief may be voluminous and include plenty of data, therefore the extraction of the information we needed was time intensive. Secondly, each manuscript had to be converted to a machine-readable form, either by typing it manually or by scanning it and converting it in text file using the appropriate applications.

3.3. Limitations and Assumptions

Our access to authentic proceedings was the first obstacle that we managed to overpass. However, another problem arose, that of ethical considerations, i.e., whether the testimonies can be discussed and analyzed, as the analysis may reveal some personal data or issues, which may become detrimental to the participants of the proceedings. Especially, regarding the cases which tend to go further to courts of appeal. The solution was using the receiving information in anonymous form, keeping the privacy of the subjects involved. Thus, every person's name involved in the trial briefs or other clues that would reveal personal information, such as addresses or workplaces, were erased from our dataset.

The problem of the researcher's presence in the proceedings, which might influence the behavior of the participants to interactions under research (Heffer, 2005, p. 52-53), or otherwise create emotionally based bias in the researcher's mind, is eliminated since we used pre-existing trial proceedings.

A potential disadvantage is that by having access only to transcripts and not to the original interaction, a considerable amount of contextual information as well as paralinguistic features, such as prosodic features, are lost (Linelu et al., 1988). However, our study focused on linguistic features and not on the meaningfulness of the testimonies.

The process of testimonies' transcription, i.e., the transformation of speech into writing by court reporters, results in adjusting spoken language to written style (Heffer, 2005, p. 54). Therefore, their complete accuracy is hardly achievable (Fraser, 2003; Tkačuková, 2010; Walker, 1986). As Gibbons (Dumas, 2007, p. 31) notes, "the process of transforming speech into a readable form can involve radical change". As a result, the transcripts, even being accurate, often miss such features of naturally occurring interaction as overlap, false starts. etc., moreover, they may contain corrections in grammar and syntax. The latter is a limitation of our dataset since the court decisions we received consisted of transcriptions without audio recordings which would theoretically confirm their validity.

Another assumption we made was the unavoidable loss of precision on the defendant's speech during the transcription procedure in case where an interpreter was interfered. The trial briefs we used contained testimonies of foreigners, who testified with the aid of an interpreter. Inevitably, an interpreter translates the original words by possibly correcting the syntax or even the vocabulary of the speaker. A recent study notes that errors in interpretation from the court-appointed interpreter may impinge key constitutional rights and suggests ways courts can better defend the constitutional rights of limited English proficiency defendants and defendants relying on the testimony of limited English proficiency witnesses (Santaniello, 2018). Hovland (1993) examines appellate cases in which non-English speaking criminal defendants and bilingual criminal defendants raised the issue of inaccurate interpretation and concludes that the present procedure for appellate review is inadequate. Moreover, in Greek criminal courts the interpretation is simultaneous with the testimony of the defendant or the witness. Relevant research studies the merits of consecutive interpretation versus simultaneous interpretation in the courtroom and concludes that a much higher degree of accuracy can be attained with consecutive interpretation (Tse, 1998).

Even though our dataset stemmed from transcriptions that may have been 'edited' in this way, they can still be considered a reliable source of material for this study. That is explained considering that our study aims at producing findings by seeking to reveal generic features of defendants' and witnesses' narrative testimonies, which are units of a higher hierarchical linguistic level than phono-stylistics or paralinguistics, thus the possible deviations of the transcripts from the real interaction which transpired in the court can be regarded as insignificant.

Furthermore, to the possible question of whether the testimonies are written by lawyers, the answer is that, in case this was perceived by us (we encountered eight such cases), they were not included in our dataset. The times this was done, it was clearly understood that the speech was intelligibly written by a law professional, as the testimony was written in an additional printed form, different of the trial proceedings. Moreover, most of the times the defendants testified in front of the judge, thus our dataset consists mainly of direct report speech.

3.4. Corpora

The dataset which we were provided, was used to build three corpora. All three corpora are stored in digital form, i.e., text files ('.txt' files) in order to be processed at a subsequent time with the appropriate word processing software.

3.4.1. Greek Corpus of Defendants' Testimonies

Our main goal was to study the language used by defendants of murder, thus the first corpus we built was the Greek Corpus of Defendants' Testimonies (GCDT) which was constructed from the transcripts that contain the exact words pronounced by the defendants in the hearings in front of the judge (Katranidou & Frantzi, 2016).

This corpus consists of texts which contain only the exact words of the testimonies of the defendants, leaving every other participant's words out of the corpus. In total the corpus consists of 108,403 words from 86 hearings, issued by 124 subjects, all of which are defendants of murder.

Apart from the speech of every defendant, we extracted some metadata from their testimonies, some of which were used later in the quantitative analysis of GCDT. These metadata include the defendant's sex, date of birth, nationality, occupation, place of birth, marital status, number of children and the case verdict. These metadata are written in the beginning of each trial brief and their assortment was made manually. These metadata are depicted in Appendix.

One hundred and ten of them are men and fourteen are women. Ninety-one are native Greek speakers and thirty-three testify through an interpreter. Their average age at the time of the hearing is approximately 38 years. Their level of education is not precisely known. Regarding their occupation, most of them are workers, farmers, builders, freelancers, two are students, four are pensioners and twenty-four are unemployed.

In most of the cases (88.8%) the verdict is condemnatory and only in a few cases (11.2%) has the defendant been acquitted. The acquittals in murders are much rarer than the convictions since the defendant's lawyer usually tries to find extenuating circumstances to reduce the defendant's penalty instead of aiming for an acquittal. The few times that the verdict is not condemnatory are due to lack of clear evidence for the crime.

3.4.2. Greek Corpus of Witnesses' Testimonies

The second corpus we built was the Greek Corpus of Witnesses' Testimonies (GCWT) and its aim was to be used as a reference corpus. This corpus was constructed using the same trial briefs which were used for the extraction of defendants' speech, and contains the exact words pronounced by the witnesses in the hearings in front of the judge (Frantzi & Katranidou, 2017). Similarly, the witnesses are interviewed by the judge, the jury, the prosecutor, and the defense lawyer, thus this corpus consists of transcripts which contain only the exact words of the testimony of the witnesses, ignoring every other word of the transcript. As a witness is defined either a defense or a prosecution witness. In addition, as a witness may testify a police officer, a forensic doctor, a medical examiner, etc. In total the corpus consists of 391,819 words from 86 hearings, issued by 145 subjects.

In case of GCWT we were not interested in extracting the metadata of witnesses, at least in this study, such as their age, sex, occupation, etc., because we analyzed their speech as if they were one person. The only metadata we registered in a file was the characterization of the witness either as prosecution or defense witness. This extra information might be useful in a possible statistical analysis of witnesses' speech in a possible future work.

3.4.3. Pre-trial Corpus

We constructed a third corpus, named pre-GCDT (pretrial - Greek Corpus of Defendants' Testimonies), with testimonies during the criminal investigation of the same defendants whose words were used in GCDT corpus. The aim was to compare the language they use inside the court and before their trial during their interrogation. The criminal investigation takes place after the defendants' arrest and before their trial. In most of the cases the defendants testify in front of the investigator and their testimony is written down verbatim. In this case, the transcription of the spoken words in written text is inevitable. This means that punctuation is used, and that speech is organized into paragraphs, while some features of oral speech are eliminated, such as repetitions, "fills", incomplete phrases, etc.

These testimonies, which are in analog form (i.e., manuscripts), were used to construct the pre-GCDT. Their statements include answers to the interrogator's questions and description of the events. Some testimonies were written by the defendants' counsel using forensic terminology, however in this case the testimony was not included in the corpus.

As we already mentioned, the construction of pre-GCDT was time consuming since the original testimonies were in analog form, and they were converted in digital files. The pre-GCDT concerns 55 of the 124 defendants of the GCDT and intends to compare the defendant's words before and during their testimony in the courtroom. In total this corpus consists of 54,032 words from 52 hearings, issued by 55 subjects.

3.5. Summary

In this chapter we presented three corpora that we managed to create from scratch using as our dataset the testimonies of defendants (GCDT) and of witnesses (GCWT), gathered in real conditions inside a courtroom, and pre-trial interrogations of the defendants (pre-GCDT). The main goal was the separation of the defendants' speech from the total language material we had. Similarly, we managed to isolate the witnesses' speech, and the speech of the defendants during their interrogation. Eventually, our corpora contained the exact words that the subjects testified in direct speech, maintaining the exactness, and meaning of their speech.

Our first research question was answered since we managed to get access to sensitive data and to preserve the anonymity of personal data. The creation of these corpora enriches the research field of Forensic Linguistics in Greece, and particularly the research field of court language corpora which, as we noted above, lacks relevant studies with Greek content. Thus, it is the first digital dataset of such kind carried out on the Greek language, allowing cross-lingual comparisons, stylometric analysis of the laymen involved in the trial briefs, etc.

Apart from the linguistic analysis of the speech of the defendants and witnesses, we extracted useful information regarding the age, nationality, sex, and occupation of the defendants. Any other metadata, for instance, demographic and social data of witnesses, did not concern this research, since the subject of our study was the defendants'

linguistic profile. However, our dataset contains much more information than the one we used in this work, in a digital form, therefore it can be used as the basis for additional discourse analysis, by Forensic Linguistics, law professionals, psychologists, sociologists, etc.

4 STYLOMETRIC PROFILE OF THE DEFENDANTS

Several features for quantifying writing style, a line of research known as stylometry, monopolized the interest of the discourse analysis' researchers. Hence, a great variety of measures including sentence length, word length, word frequencies, and vocabulary richness functions had been proposed and evaluated (Stamatatos, 2009). In this chapter we present the stylometric features we used in our research based on relevant studies, and the results of the stylometric analysis of our corpora using these features.

4.1. Linguistic Features

In order to define the stylometric profile of the defendants' speech inside the court, we measured some sets of stylometric features, namely lexical, syntactic, and content-specific features, which view a text as a sequence of tokens grouped into sentences, each token corresponding to a word, number, or a punctuation mark. Apart from these features, other types of linguistic features can be profiled, such as semantics, pragmatics, information content or item distribution through a text. However, as we depicted in Table 2.1.1, we decided to restrict the current experiments to lexical, syntactic, and content-specific features to demonstrate the overall techniques and methodology for profiling before including every possible type of features. Moreover, these features were chosen because NLP tools can be applied successfully to tasks, such as sentence splitting, POS tagging, text chunking, partial parsing, while on the other hand, more complicated tasks such as semantic or pragmatic analysis cannot be handled adequately, yet, by current NLP technology for unrestricted text (Stamatatos, 2009). The use of other features might be the subject of further research.

4.1.1. Lexical features

Sentence length and word length

We measured the average number of words of every sentence of our corpora and the average number of characters of every word, respectively. Moreover, we used the standard deviation of the sentence length and the standard deviation of word length, which can give information about how the defendants' language might be characterized as simple or comprehensive. Moreover, it is intriguing to see whether the use of longer

sentences from the defendants mean that they use longer words, too or what the correlation is between them.

Lexical richness

The lexical, or also known as vocabulary, richness of a text accounts for how many different word types are used in the text. A typical metric that is used is the Type to Token Ratio (TTR), where 'type' is the total number of the distinct words, and 'token' is the total number of the running words in the text. Another metric is the number of 'hapax legomena', i.e., the words occurring once and the 'dis legomena', i.e., the words occurring twice. The hapax and dis legomena and the ratio of dis legomena to hapax legomena in the text segment, is indicative of the authorship style (Hoover, 2003). Moreover, the more synthetic a language the more different words (Lardilleux, 2007).

Most frequent words

A simple, but effective, method to define a lexical feature for analyzing a text is extracting the most frequent words found in a corpus. The only decision that had to be made was to find the proper number of the frequent words that would be used as a feature. In the earlier studies, sets of at most 100 frequent words were considered adequate to represent the style of an author (Burrows, 1987, 1992). We note that the first dozen of most frequent words of a corpus is usually dominated by function words (articles, prepositions, etc.). Hence, the combination of two or more stylometric features can improve the evaluation of a text analysis.

Word n-grams

Word n-grams, i.e., n-contiguous words, also known as word collocations, have been proposed as textual features (Coyotl-Morales et al., 2006; Peng et al., 2004; Sanderson & Guenter, 2006). In our study we used word 2-grams in order to detect the verbs with negative meaning, because in Greek language the negation is defined with the word 'δεν/δε' or 'μην/μη' in front of a verb which is translated as 'not' and gives a negative meaning to the verb that follows. Thus, it was important for the text analysis of our corpora to find out when a verb had a negative meaning, as it changed the meaning of the phrases.

4.1.2. Syntactic features

The syntactic information is considered more reliable authorial fingerprint in comparison to lexical information. However, this means that a text needs additional processing such as POS-tagging, i.e., marking up a word in the text (corpus) as corresponding to a particular Part of Speech. However, the syntactic measure extraction is a language-dependent procedure since it relies on the availability of a parser able to analyze a natural language with relatively high accuracy, which increases the computational cost. In our case we used a Greek POS tagger, which we will mention further below.

Function and content words

The most common words, i.e., determiners, conjunctions, prepositions, pronouns, modals, qualifiers, and question words, are usually called function words and are found to be among the best features to discriminate an author among others (Argamon & Levitan, 2005). Although some researchers tend to remove these words from their dataset before performing a statistical analysis, it revealed that such words have the potential to indicate not only stylistic but content information as well (Mikros & Argiri, 2007). Content words are those that carry clear meaning such as nouns, verbs, adjectives, and adverbs.

Part of Speech frequencies

Style is also characterized from the POS frequencies (Gamon, 2004; Zhao & Zobel, 2005). Thus, during the linguistic analysis we calculated the frequencies of every category (nouns, verbs, adjectives, and adverbs) of the content words of our corpora, separately, and then we performed comparisons based on these frequencies between the corpora to be analyzed.

Lexical and functional density

Lexical density, a measure of how informative a text is (García & Martin, 2007), evaluates the proportion of content words in the text and is defined as the number of content words divided by the total number of words. Functional density is another metric which gives an indirect measure to rank texts in terms of lexical richness and equals to the ratio of function to content words frequencies in the text (Miranda & Calle, 2007).

4.1.3. Content-specific features

Content-specific features are important keywords and phrases pertaining to certain topics. Given that our texts in question deal with the same topic, i.e., a crime, and all are of the same genre, i.e., testimonies inside a courtroom, we can define certain words frequently used within that topic and that genre.

Keywords

From a statistical point of view, keywords are significantly more frequent words than expected in a sample of texts (Scott & Tribble, 2006). Our goal was to find keywords in our study corpus which had unusual frequency in comparison with a reference corpus. The result of this comparison was the 'keyness' value, which describes the value of a word being a 'key' in its context. Practically this means that the higher the value of keyness, the more unusually frequent the word appears in the study corpus compared to the reference corpus. On the other hand, keyness may have a negative value. That means that with the aid of a software we can also identify words whose frequencies are statistically lower in the study corpus, which are called 'negative keywords', in contrast to positive keywords, which have higher frequencies in the study corpus. Negative keywords are the words of the study corpus that appear quite infrequent compared to the reference corpus.

4.2. Stylometric Analysis

We used Wordsmith Tools v.5[14] (Scott, 1998) for processing our corpora. Wordsmith Tools is a software package used primarily from linguists, in particular for work in the field of corpus linguistics. It is a collection of modules for searching patterns in a language. The software handles many languages including the Greek language. Moreover, we used a Greek POS tagger[15] in order to POS-tag all corpora. We have to note that we found difficulties in detecting an efficient and user-friendly tool for performing POS-

[14] https://lexically.net/wordsmith/

[15] Natural Language Processing Group, Department of Informatics - Athens University of Economics and Business, http://nlp.cs.aueb.gr/software.html

tagging in Greek language. Even the one that we finally used, made several mistakes that we had to correct by hand. The error rate was around 20% and usually some nouns were mistaken for pronouns or adjectives. However, the bulk of the POS-tagging was performed sufficiently.

4.2.1. Internal GCDT comparisons

Initially, having GDCT as the only corpus to be studied, we divided the corpus into individual parts in order to perform internal comparisons. This decision was feasible due to the metadata that we extracted during the corpus preprocessing. Metadata are every side information in text mining applications. The side information is non-textual data. Thus, we came to some conclusions about the linguistic profile of the defendants compared to their age and citizenship. It would be interesting to study the defendants' profile depending on their education as well, but the information we had about their education was inadequate. However, knowing only their occupation, we concluded that most of them had no or elementary education, thus we could not draw any useful conjecture. Similarly, 110 of the defendants are men and 14 are women, thus we could not extract any useful information regarding their sex, because the two categories in our corpus are quite imbalanced, thus a possible comparison would not be scientifically substantiated.

For every category we performed the same measures which included the 100 most frequent words, POS frequencies, lexical density, functional density, hapax and dis legomena and keywords. In order to compare the POS frequencies of the corpora we used the frequency as a percent of the tokens in the text(s) the word list was made from.

By age

The average age of the defendants at the time of the hearing was approximately 38 years. We divided the defendants testimonies into three categories according to their age, i.e., the first category contained ages between 20-34 (44,156 words), the second category ages between 35-49 (43,596 words) and the third category ages above 50 years old (20,651 words).

The Wordsmith WordList tool gave us a list of all the words separately for each category in frequency order. Concerning the 100 most frequent words we came to the conclusion

that at least the first 50 words are common in three categories, including prepositions, conjunctions, articles and some auxiliary verbs such as 'be', 'do' and 'have' in first person singular in past tense. The first content words appear after the 50th position of that list in every category. Thus, in order to measure separately the frequency of the nouns, verbs, adjectives, and adverbs, the corpus went through a POS tagger processing.

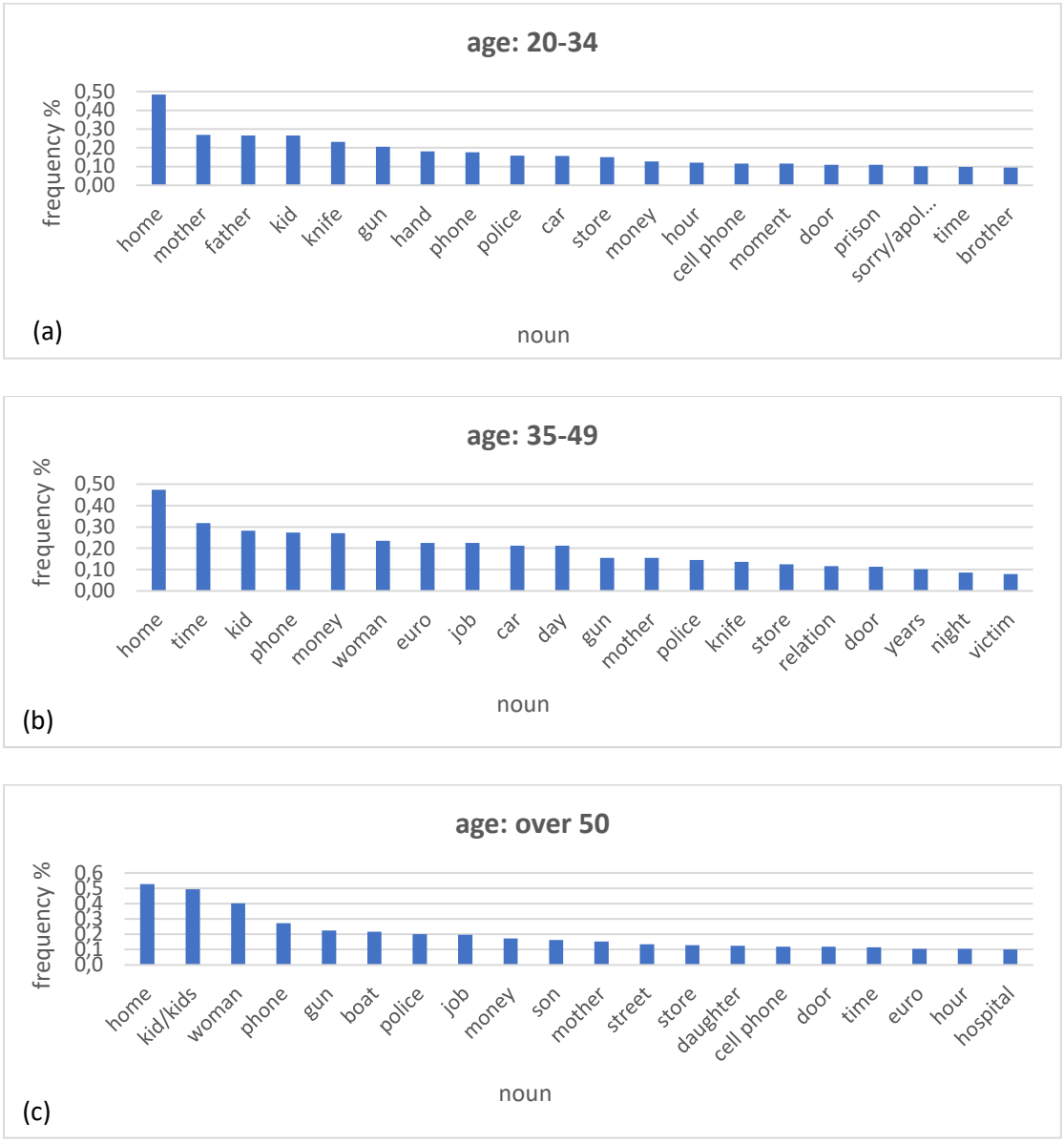


Figure 4.2.1 The first 20 most frequent nouns of three corpora divided by age: (a) 20-34, (b) 35-49 and (c) above 50 years old

Initially, we measured the 20 most frequent nouns of the three categories (Figure 4.2.1). More precisely, these nouns include all cases, i.e., nominative, genitive, accusative,

therefore the frequencies refer to the corresponding lemmas, since in Greek language the cases of a noun differ from one another in the suffix.

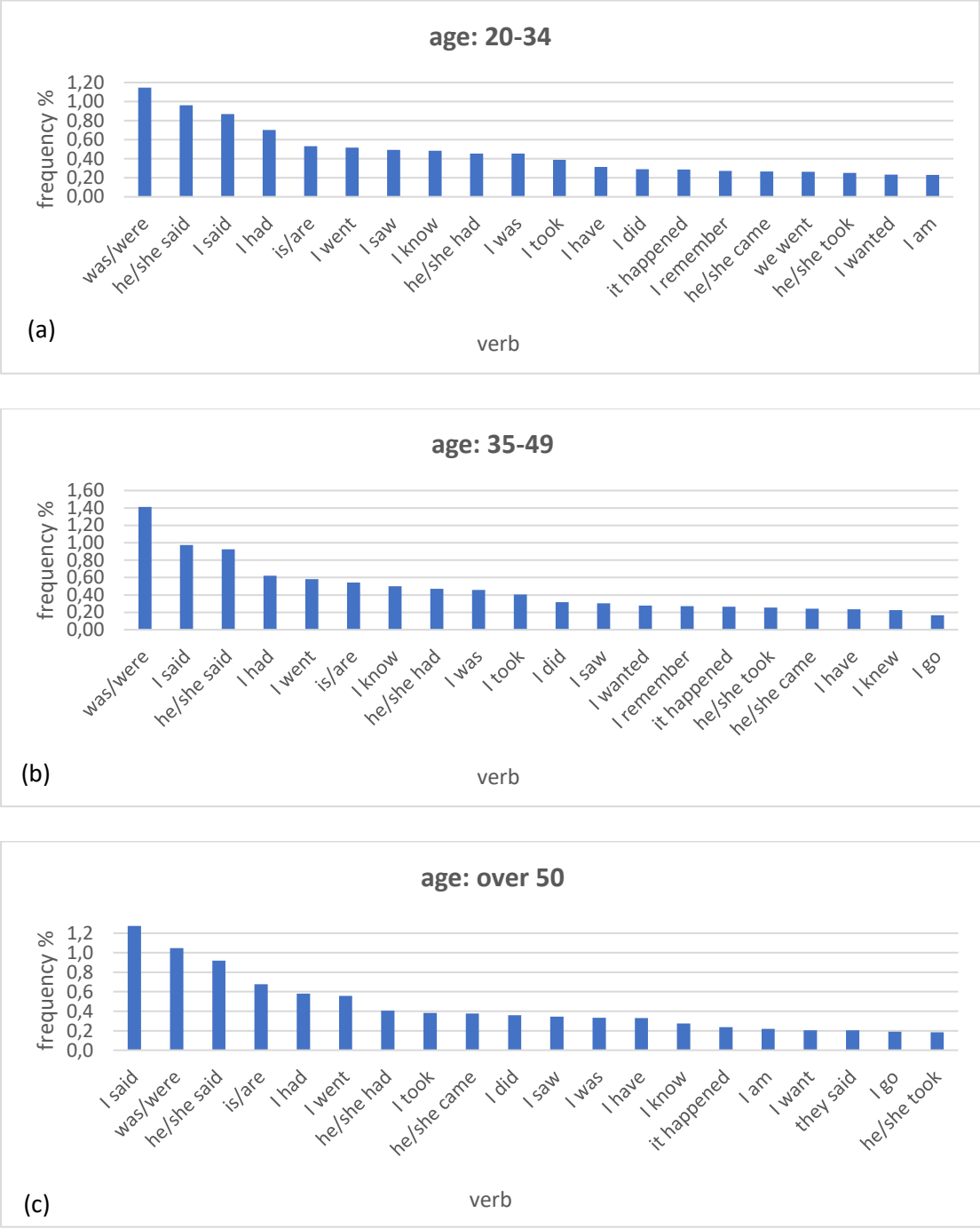


Figure 4.2.2 The first 20 most frequent verbs of three corpora divided by age: (a) 20-34, (b) 35-49 and (c) above 50 years old

It is quite interesting that in all three categories the most frequent noun is the same word (the word 'home'). Also, 10 of the 20 most frequent nouns appear in all three categories, i.e., the words 'home', 'mother', 'kid/kids', 'gun', 'phone', 'police', 'store', 'money', 'door' and 'time'. Moreover, due to the defendants' age, some words in these lists are used only in the category 'above 50', such as the words 'son' and 'daughter', whereas in the category '20-34' the word 'mother' is the second more frequent noun and the word 'father' the third one.

Subsequently, in Figure 4.2.2 we depict the first 20 most frequent verbs in these three corpora. We show that 16 of 20 verbs appear in all three categories with similar frequencies. These verbs are 'was/were', 'I said', 'he/she said', 'is/are', 'I had', 'I went', 'I saw', 'I know', 'he/she had', 'I was', 'I took', 'I have', 'I did', 'it happened', 'he/she came' and 'he/she took'. Apart from the presence of the auxiliary verb 'be' which was quite predictable, the other verbs denote an action of the defendants themselves or of a person involved in the defendants' narration of events. Thus, most of the verbs, except the auxiliary verb 'is/are', appear in the past tense.

In Figure 4.2.3 we depict the 5 most frequent adjectives of the three categories. In general, we noticed that there is limited use of adjectives in the defendants' speech. This is the reason we present only the first five ones since the frequency of the rest of the adjectives tend to zero. As we can see the frequency of these words is significantly reduced compared to the frequency of the nouns and verbs. For instance, the most frequent adjective which is the same word in all three categories is the word 'first' with frequency between 0.096 and 0.134, whereas the most frequent noun which is the word 'home' has frequency between 0.468 and 0.527 and the most frequent verb 'was/were' has frequency between 1.035 and 1.401. Moreover, we noticed that the most frequent adjectives are numerical, such as the words 'first', 'second', 'third' and some that describe quantity or quality, such as the words 'big', 'small', 'many/a lot', 'same', 'good'. Considering the fact that the adjectives which are used are either numerical or elementary, and their frequency is low, it seems that all defendants use simple or no adjectives at all, denoting a poor vocabulary usage, probably due to their low educational level.

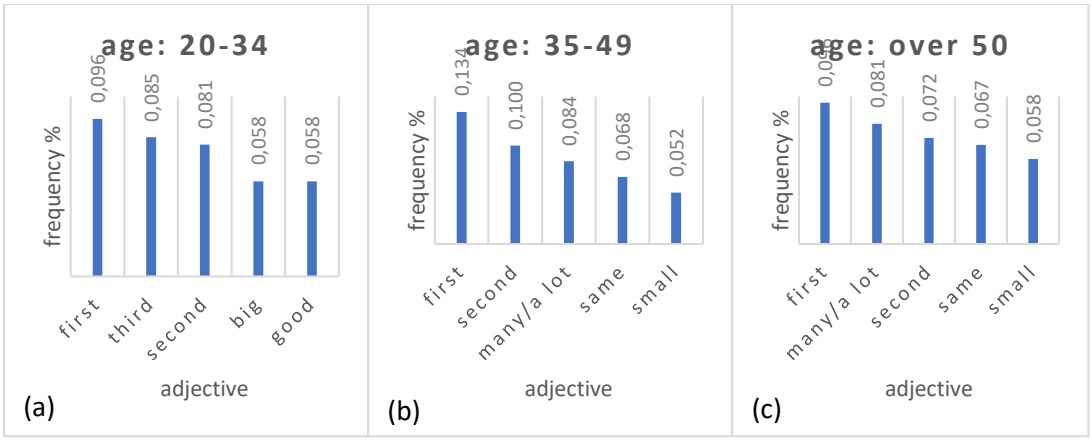


Figure 4.2.3 The first 5 most frequent adjectives of three corpora divided by age: (a) 20-34, (b) 35-49 and (c) above 50 years old

Finally, we measured the frequency of the seven most frequent adverbs of the three categories and the results are shown in Figure 4.2.4. Comparing their frequencies with those of adjectives, we realized that the use of adverbs is greater than the use of adjectives in defendants' speech. The most frequent adverbs which are used in all three categories are the words 'there', 'together', 'nice' and 'inside'. Beyond these seven most frequent adverbs, the rest show reduced frequency, thus they are not depicted in the figures.

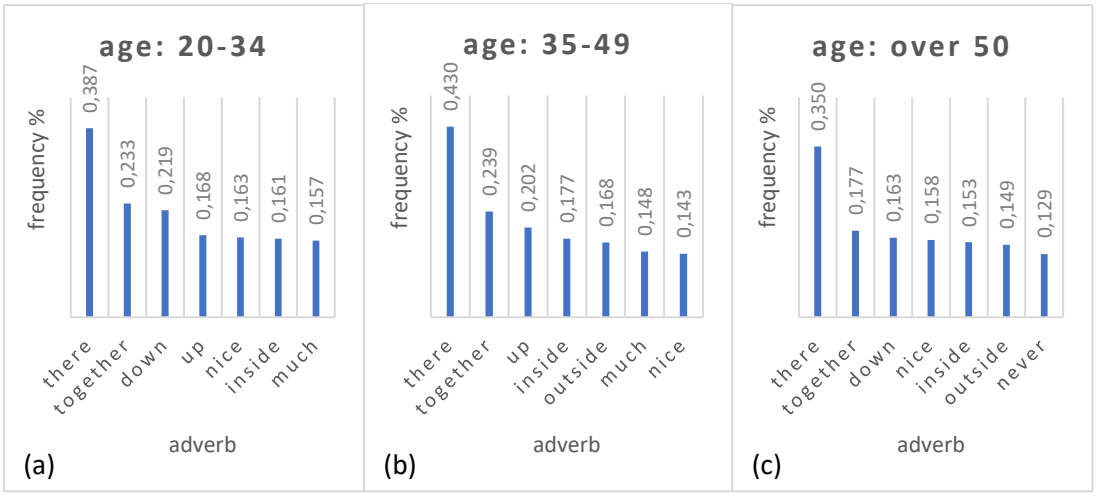


Figure 4.2.4 The first 5 most frequent adverbs of three corpora divided by age: (a) 20-34, (b) 35-49 and (c) above 50 years old

Table 4.2.1 shows the lexical richness, i.e., the TTR and the percentage of hapax and dis legomena of the three categories. In general, hapax legomena are quite common, as

predicted by Zipf's law (Baker et al., 2006), which states that the frequency of any word in a corpus is inversely proportional to its rank in the frequency list. For large corpora, about 40% to 60% of the words are hapax legomena, and another 10% to 15% are dis legomena (Kornai, 2008).

Age	**Tokens**	**Types**	**Hapax Legomena**	**Dis legomena**	**TTR%**
20-34	44156	5226	51.40	16.40	11.84
35-49	43596	5345	52.79	15.88	12.26
> 50	20651	3714	59.01	15.50	17.98

Table 4.2.1 Lexical richness of the three corpora: 'age 20-34', 'age 35-49' and 'age above 50'

In our corpora the percentage of hapax legomena seems to be at least 50% of all tokens and specifically in the category 'above 50' it reaches almost 60%. Numerically speaking, this means that at least 2500 different words in the categories '20-34' and '35-49', and at least 2000 different words in the category 'above 50' occur only once in these corpora. This is explained from the fact that the vocabulary of older people tends to be richer than the younger ones. The higher TTR indicates a higher degree of lexical variation as well. Thus, the category '20-34' has the lowest TTR and hapax legomena, the category '34-49' has a little higher ratio and the category 'above 50' displays the highest ratio both in hapax legomena and TTR. Regarding the frequency of dis legomena the results show that their percentage seems to be between 15.50% and 16.40%. Numerically speaking, this means that at least 850 different words in the categories '20-34' and '35-49', and at least 500 different words in the category 'above 50' occur only twice in these corpora. Arithmetically, dis legomena are more in the younger speakers.

Subsequently, we measured the lexical and the functional density of the three corpora (Table 4.2.2). Lexical density is almost equal, i.e., approximately 45%, for these corpora. The fact that the lexical density of the category 'above 50' is slightly higher, depicts that their testimonies are more informative than the others. This also means that the older people tend to use more descriptive language and more information-bearing content words. Functional density is slightly higher in younger ages (1.28), than in older ones (1.27 and 1.26, respectively), because they tend to use more function words, as one can see also from the function word (FW) frequencies. In the category '20-34' the frequency

of function words is 56.20%, whereas in the categories '35-49' and 'above 50' is 56.03% and 55.77%, respectively.

Age	FW frequency %	Lexical Density %	Functional Density
20-34	56.20	43.80	1.28
35-49	56.03	43.97	1.27
>50	55.77	44.23	1.26

Table 4.2.2 FW frequency, lexical and functional density of the three corpora: 'age 20-34', 'age 35-49' and 'age above 50'

In Table 4.2.3 we depict the average word length and standard deviation (in characters) and the average sentence length and standard deviation (in words) of the three categories. We found that there are slight differences between the speech of 'above 50' defendants and the others. The average sentence length of 'above 50' defendants (9.37 words) is higher than the other two categories (8.07 and 8.04 words, respectively), which seem to have similar results. However, the average word length is similar of all defendants (4.43 and 4.45 characters). Typically, all defendants use one-word or short responses. Moreover, the 'above 50' defendants tend to use more complicated and longer sentences.

Age	Avg word length	Word length st.dev.	Avg sentence length	Sentence length st. dev.
20-34	4.43	2.27	8.07	6.08
35-49	4.45	2.27	8.04	6.00
>50	4.45	2.28	9.37	7.44

Table 4.2.3 Word and sentence length and standard deviation of the three corpora: 'age 20-34', 'age 35-49' and 'age above 50'

Finally, we used the Wordsmith Keywords tool aiming to find unusually frequent words that appear in the study corpus compared to the reference corpus. As a study corpus, we set each of the three corpora on its own, successively, and we defined the other two as the reference corpus. Thus, this process was repeated three times, once for each category. The results are depicted in the following tables. In Table 4.2.4 we depict the 6 words that appear in the category 'age 20-34' unusually frequent and the 2 words that appear quite infrequent compared to the other two categories.

Indicatively, the word 'father' with the highest positive keyness value, is the most unusually frequent word compared to the reference corpus, whereas the word 'woman' which is the lower negative keyword, appears quite infrequent in the study corpus compared to the reference corpus.

N	Keyword	Translation	Keyness
1	πατέρα	father	49.91
2	με	with	40.00
3	τον	him	28.56
4	μπουκάλι	bottle	27.82
5	μάνα	mother	25.92
6	είδα	I saw	24.35
7	ευρώ	euro	-25.83
8	γυναίκα	woman	-44.08

Table 4.2.4 Positive and negative keywords. Study corpus: 'age 20-34', and reference corpora: '35-49' and 'above 50'

Similarly, in Table 4.2.5 we depict the 2 words that appear in the category 'age 35-49' unusually frequent and the 2 words that appear quite infrequent compared to the other two categories. The two most usually frequent words seems to be the words 'euro' and 'money', whereas the words 'him' and 'father' are quite infrequent in the study corpus compared to the reference corpus.

N	Keyword	Translation	Keyness
1	ευρώ	euro	37.17
2	λεφτά	money	24.82
3	πατέρα	father	-28.50
4	τον	him	-37.98

Table 4.2.5 Positive and negative keywords. Study corpus: 'age 35-49', and reference corpora: '20-34' and 'above 50'

Table 4.2.6 depicts the 3 words that appear in the category 'age above 50' unusually frequent and the 2 words that appear quite infrequent compared to the other two categories. The word 'boat' is the most unusually frequent compared to the reference corpus with keyness value equal to 80.93. That is explained because several defendants of

this category were involved in a trial case with illegal transport of migrants by sea. The word 'not' is quite infrequent in this category compared to the other two.

N	Keyword	Translation	Keyness
1	σκάφος	boat	80.93
2	γυναίκα	woman	40.13
3	γιο	son	38.85
4	στη	to (her)	-28.05
5	δεν	not	-30.15

Table 4.2.6 Positive and negative keywords. Study corpus: 'above 50', and reference corpora: '20-34' and '35-49'

We denote that each of the above tables, also, contained some proper nouns, i.e., people names and specific places, which were removed.

By citizenship

Ninety-one of the defendants were native Greek speakers and thirty-three testified through an interpreter. We divided the defendants according to their citizenship in two categories (corpora), i.e., the first category contains the native speakers (66,002 words) and the second category contains the non-native speakers (42,401 words) who testified through an interpreter.

Similarly, using the same tool, we conducted the same measurements as in the previous subsection. Firstly, we measured the 100 most frequent words in both corpora. We concluded that most of them are prepositions, articles, conjunctions, auxiliary verbs and the first content words are shown after the fiftieth word in the native speakers and after the sixtieth word in the non-native speakers. As we expected, this means that non-native speakers use more function words than the native ones, since most of them seem to have an even lower educational level than the Greek defendants.

Like in the previous subsection the two corpora underwent a POS tagger processing. The first 20 most frequent nouns of the two categories are depicted in Figure 4.2.5. It is shown that 13 of 20 nouns are common in both categories. We clarify that the word 'car' seems to appear twice in Figure 4.2.5(a) because in Greek language there are two words that are used with similar meaning, i.e., the word 'αμάξι' and the word 'αυτοκίνητο' and we registered them separately. Besides the word 'car', the nouns 'home', 'kid', 'woman',

'time', 'phone', 'knife', 'money', 'police', 'store', 'euro', 'job' and 'door' are present in both lists, whereas the nouns 'gun', 'mother', 'father', 'moment', 'girl' and 'pills' are only in the frequency list of the native speakers, and the nouns 'hand', 'cell phone', 'person', 'friend', 'bottle', 'years' and 'prison' are shown only in the frequency list of the non-native speakers.

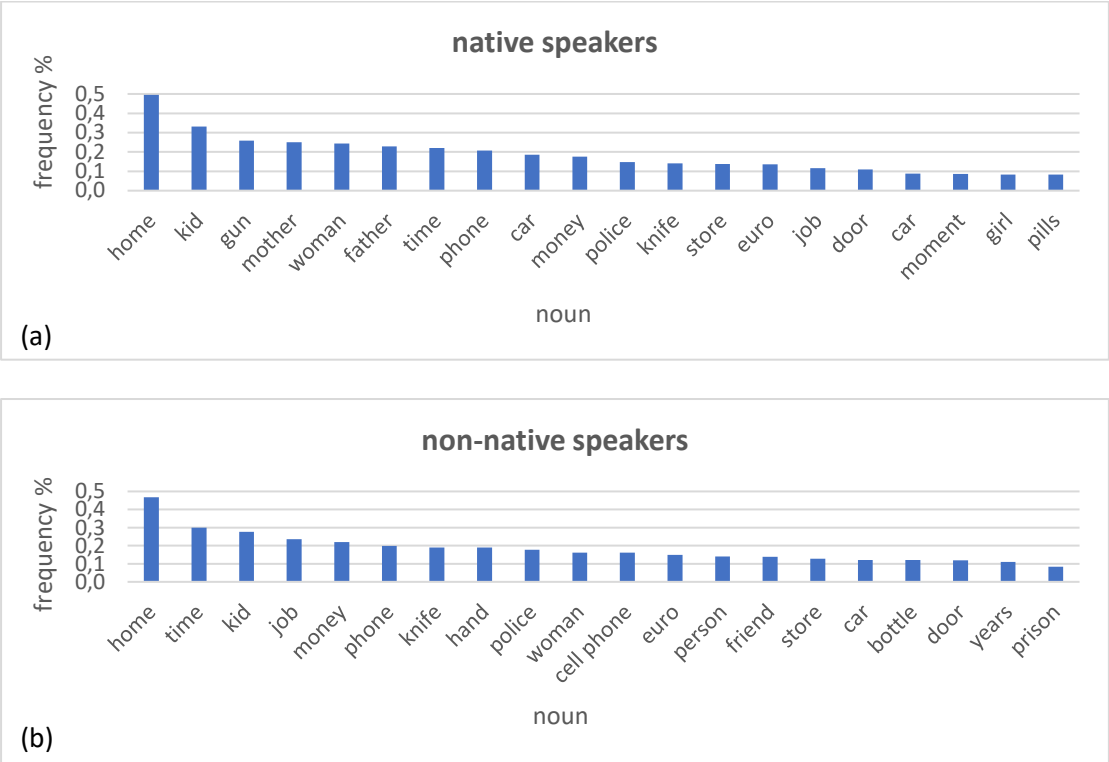


Figure 4.2.5 The first 20 most frequent nouns of two corpora divided by citizenship: (a) native speakers and (b) non-native speakers

We also noticed that the frequencies in the native speakers are greater than the non-native speakers, which means that the native speakers use these nouns more frequently than the non-native ones.

Subsequently, we depict the first 20 most frequent verbs of the two categories (Figure 4.2.6).

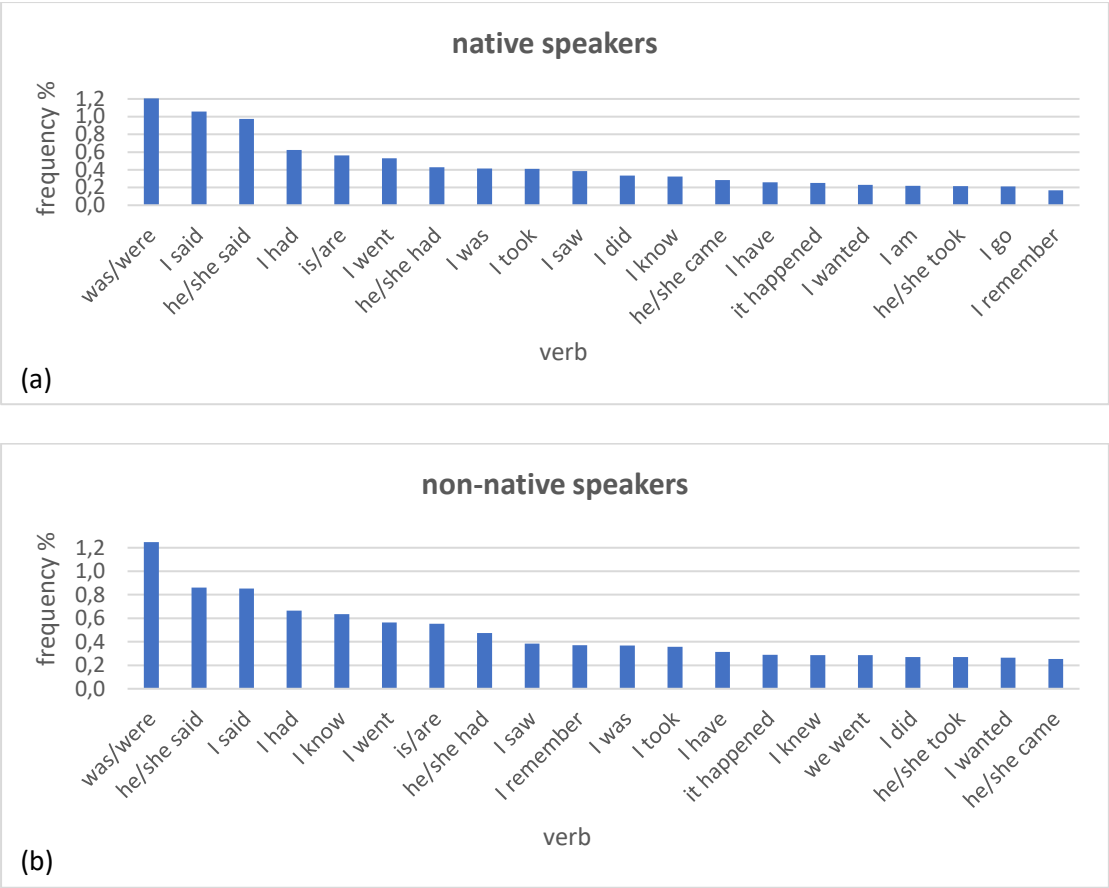


Figure 4.2.6 The first 20 most frequent verbs of two corpora divided by citizenship: (a) native speakers and (b) non-native speakers

It is shown that 18 of 20 verbs are common in both categories. These verbs are 'was/were', 'I said', 'he/she said', 'is/are', 'I had', 'I went', 'I saw', 'I know', 'he/she had', 'I was', 'I took', 'I have', 'I did', 'it happened', 'he/she came', 'he/she took', 'I wanted' and 'I remember'. The verbs in this list are usually used to describe an action of one's person in present or past tense. Also, it is depicted that the verbs 'I know' and 'I remember' are in this list, but they are mostly used with the negative word 'δεν/δε' ('no/not') stating the defendant's 'not knowing' and 'not remembering' of something. This assumption came from the observation of the fifty most frequent 2-grams. In native speakers, the 2-gram 'δε ξέρω' (I don't know) is eighth in the ranking and the 2-gram 'δε θυμάμαι (I don't remember) is forty-second. In non-native speakers, the corresponding 2-grams are third and sixth in the ranking. The verbs that are present only in the native speakers' frequency list are 'I am' and 'I go', whereas the verbs that are present only in the non-native speakers' frequency list are 'I knew' and 'we went'.

Afterwards, we compared the 5 most frequent adjectives in both categories. The frequencies are reduced compared to those of nouns and verbs. The frequency of any other adjective is so low that is unworthy of reference. Most of them are numerical adjectives, such as 'first', 'second' and 'third' and adjectives that denote quality or quantity, such as 'good/nice' and 'many/a lot' (Figure 4.2.7).

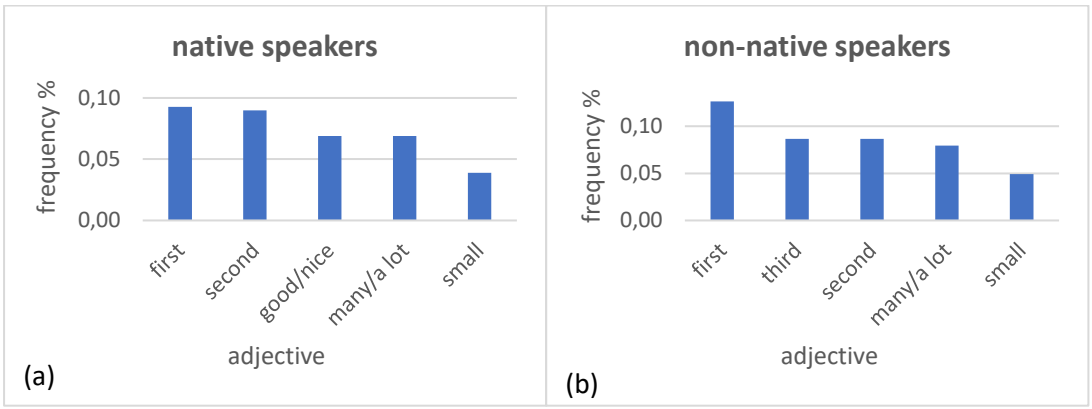


Figure 4.2.7 The first 5 most frequent adjectives of two corpora divided by citizenship: (a) native speakers and (b) non-native speakers

The simplicity of these words show that the vocabulary richness of both corpora is low.

Similarly, we compared the 7 most frequent adverbs of the two corpora (Figure 4.2.8). We have not included any other adverb in the frequency lists because the frequency seems to decrease rapidly after the 7th adverb.

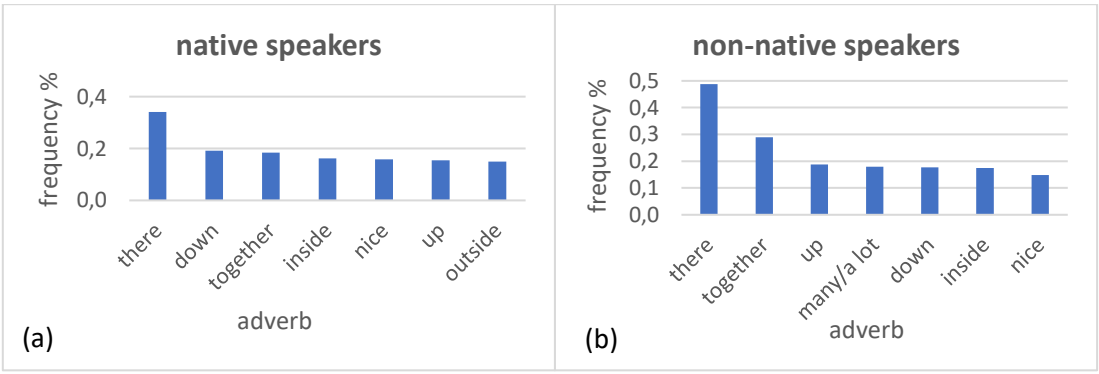


Figure 4.2.8 The first 7 most frequent adverbs of two corpora divided by citizenship: (a) native speakers and (b) non-native speakers

The most frequent adverb in both categories is the word 'there'. Non-native speakers seem to use this adverb more frequently than the native speakers. The other adverbs which are common in both lists are the words 'down', 'together', 'inside' and 'up' and 'nice', whereas the adverb 'outside' is present only in the native speakers' frequency list, and the adverb 'many/a lot' is present only in the non-native speakers' frequency list.

Subsequently, we measured the TTR and the percentage of hapax and dis legomena in both categories (Table 4.2.7). As we mentioned previously, hapax legomena and the TTR signifies a text's lexical richness. A high hapax legomena percentage or TTR indicates a high degree of lexical variation.

Citizenship	**Tokens**	**Types**	**Hapax Legomena**	**Dis legomena**	**TTR%**
native	66002	7517	53.58	15.69	11.39
non-native	42401	4653	49.36	15.94	10.97

Table 4.2.7 Lexical richness of the two corpora: 'native speakers' and 'non-native speakers'

As we expected, it is depicted that native speakers have higher lexical richness in their speech than the non-native speakers, since both the hapax legomena percentage (55.38% vs. 49.36%) and the TTR (11.39% vs. 10.97%) are higher in native speakers. On the contrary, dis legomena are slightly higher in the non-native speakers (15.94% vs. 15.69%).

Then, we measured the lexical and functional density of both corpora (Table 4.2.8). The fact that the lexical density of native speakers is slightly higher (44.3% vs. 43.38%), depicts that their testimonies are more informative than the non-native's speech.

Citizenship	**FW frequency %**	**Lexical Density %**	**Functional Density**
native	55.70	44.30	1.25
non-native	56.62	43.38	1.30

Table 4.2.8 FW frequency, lexical and functional density of the two corpora: 'native speakers' and 'non-native speakers'

Also, it seems that non-native speakers use more function words than the native speakers, since the FW percentage (56.62%) and the functional density (1.30) of non-native speakers are higher than the native ones (55.7% and 1.25, respectively).

In Table 4.2.9, we depict the average and standard deviation of word length, and the average standard deviation of sentence length of both corpora. We found some slight differences between the speech of native and non-native speakers.

Citizenship	Avg word length	Word length st.dev.	Avg sentence length	Sentence length st.
native	4.46	2.31	8.62	6.42
non-native	4.41	2.20	7.79	6.13

Table 4.2.9 Word and sentence length and standard deviation of the two corpora: 'native speakers' and 'non-native speakers'

Both average word and average sentence length of native speakers are higher than of non-native ones (4.46 vs. 4.41 characters, and 8.62 vs. 7.79 words). Thus, native speakers seem to use larger words and more complicated sentences than the non-native ones. Considering the nature of both corpora, i.e., the non-native defendants seem to have lower educational level than the native defendants, the non-native ones tend to use simpler words and shorter sentences.

Finally, we used the Wordsmith Keywords tool to find unusually frequent words that appear in the study corpus compared to the reference corpus. As a study corpus was defined the corpus that contained the non-native speakers and as a reference corpus was defined the corpus that contained the native speakers. We ended up in the previous choice because it is recommended that the reference corpus is greater than the study corpus (Berber-Sardinha, 2000). The results are depicted in Table 4.2.10. As one can see, there are 13 words that appear in the non-native speakers unusually frequent compared to the native speakers, and 9 words, depicted with a negative sign, that appear quite infrequent compared to native speakers. For instance, the words such as 'Greece', 'Greek' and 'Albania' are unusually frequent in non-native speakers compared to native ones, because they probably refer to their country of origin and destination. In contrast, the native Greek defendants rarely referred to their country (Greece). Moreover, the noun 'boss' is another unusually frequent word in non-native speakers compared to native ones since most of them work as unskilled or in low-skilled jobs and the 'employer' is usually called 'boss'. Although the Greek defendants of our corpus are low-skilled workers, the use of the word 'boss' is not very common among them. On the

other hand, the words 'she', 'her', 'gun' and 'mother' are the most infrequent words in non-native speakers compare to native ones.

N	Keyword	Translation	Keyness
1	ξέρω	I know	51.57
2	Ελλάδα	Greece	50.56
3	όχι	no	46.18
4	ναι	yes	43.18
5	θυμάμαι	I remember	40.41
6	μπουκάλια	bottles	37.53
7	ήξερα	I knew	32.88
8	πορτοφόλι	wallet	28.71
9	ήμουνα	I was	28.21
10	Αλβανία	Albania	26.94
11	θείος	uncle	24.48
12	αφεντικό	boss	24.40
13	ελληνικά	Greek	24.40
14	ότι	that	-25.12
15	πατέρας	father	-28.64
16	η	the (she)	-33.15
17	θα	will	-47.98
18	μου	me	-49.87
19	μάνα	mother	-59.29
20	όπλο	gun	-60.07
21	την	she	-139.21
22	της	her	-169.87

Table 4.2.10 Positive and negative keywords. Study corpus: 'non-native speakers', and reference corpus: 'native speakers'

4.2.2. GCDT vs Greek general language corpora

After the internal comparisons of our corpus that we described in the previous subsection, our next goal was to measure some basic linguistic features of GCDT compared to the Greek general language.

Comparison with a created corpus

As a reference corpus we used a Greek general language corpus (GGLC), that we created from two sources, the Portal for the Greek language[16] which contains text corpora from the field of journalism in electronic format, and the Clarin:el[17] which contains, among others, text corpora from various sources.

The corpus that we created contains in total 518,024 words, and particularly 188,859 words of the corpus are published in the newspapers 'Makedonia' and 'Ta Nea', and 329,165 words of the corpus are derived from a speech corpus of answers to the interviews for research conducted in 1986-87.

The features we measured were the most frequent words, POS frequencies, lexical density, functional density, hapax and dis legomena and keywords. In order to compare the POS frequencies of the corpora we used the frequency as a percent of the tokens in the text(s) the word list was made from. Comparing these features with those of the Greek general language would give us information regarding special characteristics of the testimonies' language which has its own particularities.

First, we measured the most frequent words in both corpora (Table 4.2.11). The Wordsmith WordList tool gave us a list of all the words in GCDT and the GGLC in frequency order. In both corpora, the top of this list is occupied by function words, such as 'and', 'the', 'to', 'not', 'with', 'that', etc. The nominative pronoun 'I' appears in the first 15 most frequent words in GCDT, since the defendants refer to themselves in their testimonies, while it does not appear at all in the corresponding list of GGLC. The auxiliary verb 'is/are' appears in the first 15 most frequent words in GGLC, while it does not appear at all in the corresponding list of GCDT. The 15 most frequent words in the list take up approximately one third of the GCDT and one quarter of the GGLC.

[16] Centre for the Greek language, project "Portal for the Greek language and language education https://www.greek-language.gr/

[17] Central inventory of language resources and services https://inventory.clarin.gr/

	GCDT			GGLC		
s/n	**Word**	**Freq. %**	**Cumulative freq. %**	**Word**	**Freq.%**	**Cumulative freq.%**
1	and	4.06	4.06	and	3.33	3.33
2	the	3.68	7.74	to	2.85	6.18
3	to	3.39	11.13	the	2.32	8.50
4	not	3.3	14.43	is/are	2.28	10.78
5	me	2.71	17.14	not	1.70	12.49
6	with	2.25	19.4	the (she)	1.60	14.09
7	him	1.92	21.31	that	1.54	15.63
8	her	1.7	23.01	from	1.49	17.12
9	that	1.5	24.51	the	1.48	18.59
10	into	1.47	25.98	she	1.24	19.84
11	he	1.43	27.41	with	1.11	20.95
12	these	1.39	28.8	for	1.06	22.01
13	I	1.38	30.18	of (him)	0.99	22.99
14	him	1.31	31.5	of (her)	0.98	23.97
15	for	1.23	32.73	will	0.91	24.88

Table 4.2.11 Most frequent words in GCDT and in GGLC

Table 4.2.12 shows the percentage of word types with frequency one and two in the corpus, namely the hapax and dis legomena. It is depicted that in GCDT the hapax legomena take up almost 50% of the word types, while in GGLC they occupy just over 51%. The TTR in GCDT, i.e., the number of distinct words (types) is just 8.7% of the total number of words (tokens). Similarly, the TTR of GGLC is 7.93%. This means that GCDT is more lexically rich than GGLC, or in other words the defendants use more unique words, compared to the general language.

Corpora	**Tokens**	**Types**	**Hapax Legomena**	**Dis legomena**	**TTR%**
GCDT	108403	9440	49.61	15.40	8.70
GGLC	518024	41101	51.78	15.96	7.93

Table 4.2.12 Lexical richness of GCDT and GGLC

Subsequently, we measured the frequencies of content words (CW) and the FW frequencies of both corpora. The lexical and the functional density of the two corpora differ (Table 4.2.13). For instance, the fact that the lexical density of GGLC is larger (50.66%) than GCDT (44.30%), depicts that the general language is more informative than the defendants' speech. This is justified from the fact that the sources of GGLC

include among others scientific articles, police reports, judicial reports, reviews etc., which tend to use more formal speech, more descriptive language, and more information-bearing content words. Thus, functional density in GGLC is less than 1 (0.97), since function words are less than content words, whereas in GCDT functional density is greater than 1 (1.26) since function words are more than content words.

Corpora	FW frequency %	Lexical Density %	Functional Density
GCDT	55.70	44.30	1.26
GGLC	49.33	50.66	0.97

Table 4.2.13 FW frequency, lexical and functional density of GCDT and GGLC

In Table 4.2.14 we depict the average word length and word length standard deviation (in characters) and the average sentence length and sentence length standard deviation (in words). Having made the appropriate measurements, we found that there are differences between the defendants' speech and the general language.

Corpora	Avg word length	Word length st.dev.	Avg sentence length	Sentence length st. dev.
GCDT	4.44	2.27	8.27	6.32
GGLC	4.89	2.83	24.98	1783.76

Table 4.2.14 Word and sentence length and standard deviation of GCDT and GGLC

There is a small difference in word length between the two corpora. Defendants' speech seems to use words with less characters (4.44) than the general language (4.89). However, there is a great difference in the average sentence length and sentence length standard deviation since the average sentence length for defendants (8.27 words) is shorter than that of general language (24.98 words). Considering the nature of GCDT, its low standard deviation is justified from the fact that the corpus is derived from testimonies inside a courtroom and apart from some descriptive speech pieces, it contains mainly responses. Typically, defendants use one-word or short responses. On the other hand, GGLC includes, apart from speech corpus, written texts published in newspapers. Therefore, the GGLC tend to use more complicated and longer sentences.

In order to perform a more qualitative content analysis, we used an approach based on keywords derived analyses. We used the Word Smith KeyWords tool to compare the

word list extracted from our study corpus, GCDT, to a word list extracted from the reference corpus, GGLC. We took a list with words that are significantly more frequent in GCDT than in GGLC. Table 4.2.15 depicts the list of the first 25 positive keywords, i.e., the keywords with maximum positive keyness.

N	Keyword	Translation	Keyness
1	μου	to me	3167.31
2	είπε	he/she said	2721.10
3	είπα	I said	2242.80
4	εγώ	I	1755.05
5	δεν	not	1690.73
6	τον	him	1311.08
7	ήταν	was/were	1276.20
8	είχα	I had	1152.85
9	πήρα	I took	1060.71
10	είδα	I saw	1007.33
11	πήγα	I went	972.41
12	σπίτι	home	868.56
13	έκανα	I did	711.70
14	ήμουν	I was	699.52
15	τηλέφωνο	phone	664.52
16	ήρθε	he/she came	662.33
17	στο	in	651.56
18	με	with	616.06
19	πήρε	he/she took	603.95
20	θυμάμαι	I remember	585.63
21	όπλο	gun	583.75
22	ήξερα	I knew	550.29
23	αστυνομία	police	549.29
24	μαχαίρι	knife	548.22
25	μετά	after	489.26

Table 4.2.15 First 25 positive keywords. Study corpus: GCDT, and reference corpus: GGLC

The field 'keyness' stands for the value of the log-likelihood statistics. Practically this means that the higher the value of keyness, the more unusually frequent the word appears in GCDT compared to GGLC. The list mainly consists of verbs in the first person, singular number, past tense. They are used to describe an action or a feeling of

the defendant before, during and after the event in question. Indicatively, the most unusually frequent words of GCDT compared to GGLC are the verbs 'he/she said', 'I said', 'I had', 'I took', 'I saw', 'I went', 'I did', 'I was', 'I knew', etc. The nouns 'home', 'phone', 'police', 'knife' seem to be quite frequent in GCDT compared to GGLC.

Table 4.2.16 depicts the first 25 negative keywords, i.e., the keywords with the lower negative value of keyness, i.e., words of GCDT that appear quite infrequent compared to the reference corpus.

N	Keyword	Translation	Keyness
1	είναι	is/are	-1634.69
2	τη	the (she)	-903.04
3	πούμε	say (we)	-814.42
4	ας	let	-751.21
5	περιοχή	area	-580.20
6	έχει	has	-562.80
7	των	of (them)	-518.28
8	η	she	-497.82
9	εδώ	here	-468.51
10	και	and	-447.91
11	υπάρχει	there is	-447.18
12	οι	the (they)	-440.76
13	της	of her	-379.44
14	που	where	-374.88
15	έχουν	they have	-327.24
16	πιο	more	-324.37
17	πολύ	much	-303.07
18	στη	at	-299.95
19	τώρα	now	-265.78
20	κέντρο	center	-254.19
21	δηλαδή	namely	-231.69
22	υπάρχουν	there are	-230.74
23	τους	their	-205.06
24	νομίζω	I think	-196.58
25	κόσμος	world	-188.27

Table 4.2.16 First 25 negative words. Study corpus: GCDT, and reference corpus: GGLC

The lower the value of keyness, the more unusually frequent the word appears in the reference corpus of GGLC compared to our study corpus GCDT. The list consists of verbs in present tense, such as 'is/are', 'say', 'there is/there are', 'I think', nouns such as 'area', 'center', 'world' and the rest of the list consists of prepositions and pronouns.

Comparison with published corpora

There are several research projects in Greece which are designed for the qualitative analysis of the Greek language and the linguistic communication. Parts of these corpora are available online and can be used for quantitative analysis.

As a reference corpus we used three Greek general language corpora, successively, parts of which are posted on the internet, namely the Hellenic National Corpus (HNC)[18] (Hatzigeorgiu et al., 2000), the Corpus of Greek Texts (CGT)[19] (Goutsos, 2003, 2010) and the Corpus of Spoken Greek (CSG)[20] (Pavlidou, 2012).

At the time of writing, HNC was currently the biggest written corpus of Modern Greek, consisting of 62,041 texts and 62,435,379 words derived from written language material, such as books, newspapers, journals, etc. CGT is the first electronic corpus of Greek that was created with the aim of providing a resource for linguistic research in a wide range of both written and spoken Modern Greek genres. At the time of writing, it consisted of 26,031 texts and 29.511.849 words which had come from written texts. CSG is a set of digital files, which is updated and enriched according to the research project's affordances and needs. It consisted of 1.8 million words which has been drawn from naturally occurring circumstances of spoken communication. Part of the transcribed material is available and can be used freely online, consisting of 671,543 words which included 40 everyday conversations among family and friends, 145 telephone calls and 17 television interviews with politicians.

---

[18] Hellenic National Corpus, Institute for Language and Speech Processing, ATHENA Research & Innovation Information Technology, http://hnc.ilsp.gr

[19]Corpus of Greek Texts, University of Athens' program "Kapodistrias", http://www.sek.edu.gr/

[20]Corpus of Spoken Greek, Institute of Modern Greek Studies, Manolis Triandaphyllidis Foundation, part of the Greek Talk-in-interaction and Conversation Analysis research project, http://corpus-ins.lit.auth.gr/

We measured POS frequencies, the most frequently used words, the most frequent function words, and the most frequent word 2-grams. We used the frequency as a percent of the tokens in the text(s) the word list was made from. By using the frequency as a percent, we are given the capability of comparing the frequency of specific words in GCDT and a reference corpus. We examined the twenty most frequent nouns and verbs (Figure 4.2.9), and the ten most frequent adjectives, adverbs, and pronouns of GCDT and we compared their frequency with the frequency of their appearance in the three reference corpora, successively (Figure 4.2.10). The analysis of the measurements is described in the following paragraphs.

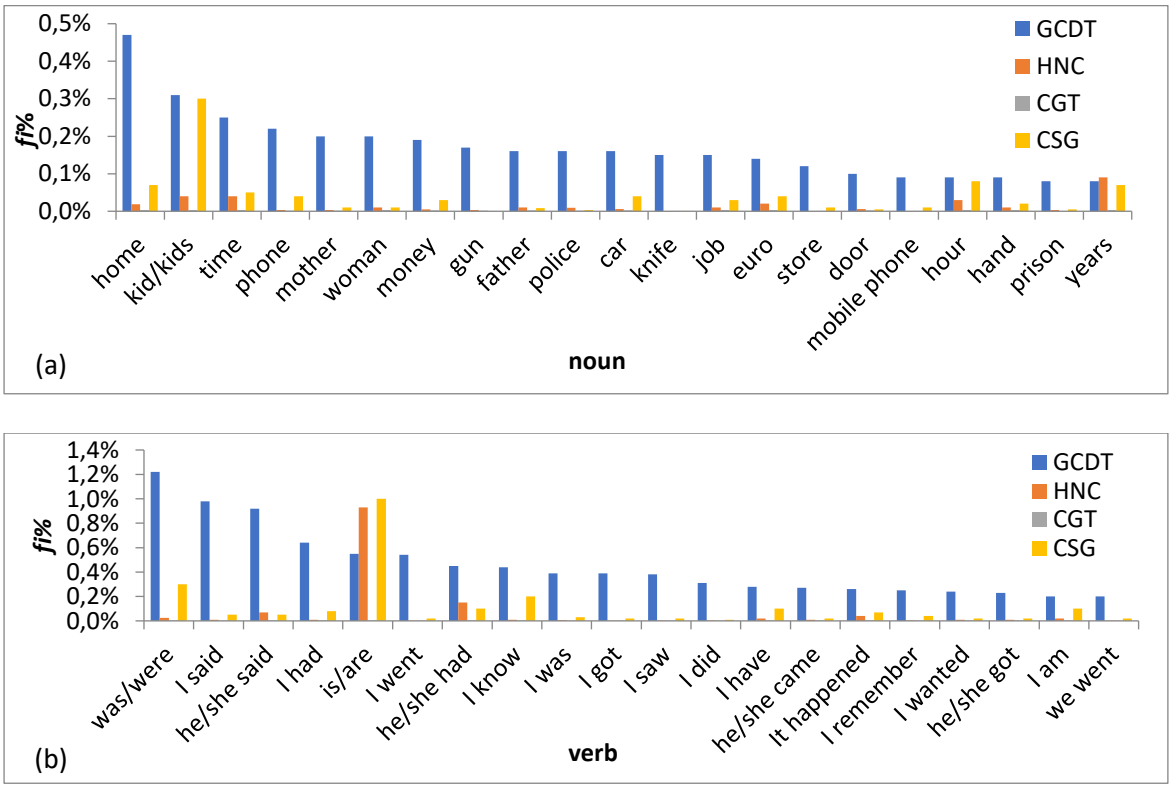


Figure 4.2.9 The 20 most frequent (a) nouns, and (b) verbs in GCDT and in the reference corpora

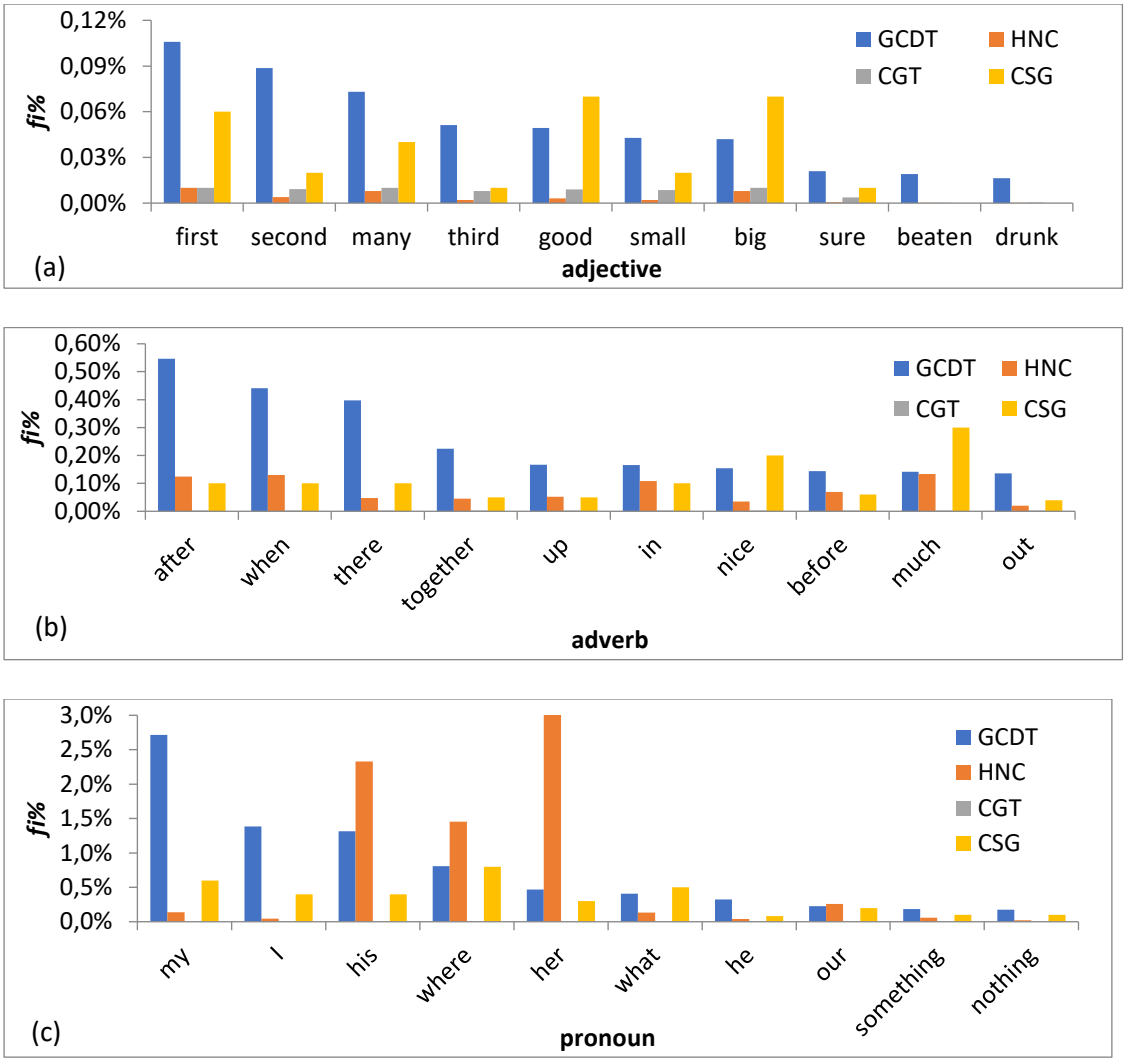


Figure 4.2.10 The 10 most frequent (a) adjectives, (b) adverbs, and (c) pronouns in GCDT and in the reference corpora

Vs. Hellenic National Corpus

Firstly, we examined the twenty most frequent nouns and verbs of GCDT, and we compared their frequency with HNC. The results showed a large variance in the frequencies of occurrence of words between the two corpora, not only for nouns where we would expect a higher frequency of occurrence in GCDT for specific words such as 'knife', 'money', 'gun', 'police', 'prison', but for other nouns such as 'telephone' and 'mother'. Apart from the noun 'years', none of the other nineteen most frequent nouns in GCDT is as frequent in HNC but, on the contrary, they present a much lower frequency of occurrence (Figure 4.2.9a).

Similarly, apart from the verb 'to be' ('is/are'), which is significantly less frequent compared to HNC corpus, the rest of the most frequently used verbs in GCDT, in present and past tense, are much rarer in HNC. It is worth noting that among the twenty most frequent verbs, fifteen are used in the past tense, since the defendants' testimonies describe a past action, i.e., 'was/were', 'I said', 'he/she said' and only five of them are used in present tense, relating to the hearing procedure: 'is/are', 'I know', 'I have', 'I remember', 'I am' (Figure 4.2.9b).

At this point, we noticed that the verbs 'ξέρω' ('I know') and 'θυμάμαι' ('I remember') are mostly used with the negative word 'δεν/δε' ('no/not') stating the defendant's 'not knowing' and 'not remembering' of something. This assumption came from the observation of the twenty most frequent 2-grams. In particular, the 2-gram 'δε ξέρω' (I don't know) is fourth in the ranking and the 2-gram 'δε θυμάμαι (I don't remember) is fifteenth. In Table 4.2.17, we illustrate two indicative examples of the use of the verbs "know' and 'remember'.

Example 1: "… δε ξέρω πώς έγινε, έστριψα τη στροφή και την έριξα. …"
Translation: "… I don't know how that happened, I turned, and I threw her …"
Example 2: "… Δε θυμάμαι. Μαλώσαμε. Αυτή φώναζε …"
Translation: "… I don't remember. We argued. She was yelling …"

Table 4.2.17 Example of the use of the verbs 'I know' and 'I remember' in GCDT

After verbs and nouns, we extracted the frequency lists of adjectives, adverbs, and pronouns, and compared their frequencies of occurrence to those in HNC. Regarding the use of adjectives, we noticed that, among the ten most frequent ones, there are adjectives such as 'first', 'second', 'third', 'many', 'good', 'small', 'big' and 'sure', which appear in HNC with much lower frequency, and adjectives such as 'beaten' and 'drunk' which are considerably infrequent in HNC (Figure 4.2.10a).

Adverbs seem to be used more frequently in GCDT than in HNC, since the defendants' language tends to be descriptive. The adverbs 'after', 'when', 'there', 'together', 'up', 'in', 'nice', 'before', 'much' and 'out' are the ten most frequently used. Apart from the adverb

'much', the rest present a much higher frequency of appearance compared to HNC (Figure 4.2.10b).

Regarding pronouns, the two most frequent ones in GCDT, 'my' and 'I', are a lot rarer in HNC. However, the pronouns 'his', 'where' and 'her', have much lower frequencies compared to those in HNC (Figure 4.2.10c).

Vs. Corpus of Greek Texts

Comparing the appearance of the 20 most frequent nouns and verbs of GCDT in the CGT, we noticed that there is no relevance between them, since the corresponding words in the latter corpus tend to zero frequency. Similar results are derived from the comparison of the appearance of the 10 most frequent adverbs and pronouns of CGDT in the reference corpus CGT. However, regarding the use of adjectives, we noticed that adjectives such as 'first', 'second', 'third', 'many', 'good', 'small', 'big' and 'sure', appear in CGT with clearly lower frequency than GCDT, but with similar or greater frequency than HNC. Adjectives such as 'beaten' and 'drunk' are considerably infrequent in CGT as in HNC.

Vs. Corpus of Spoken Greek

Concerning the comparison with CSG, we noticed considerable differences in contrast to the other two reference corpora. For instance, nouns such as 'kid/kids', 'hour' and 'years' seem to have similar frequency both in GCDT and CSG. Moreover, the auxiliary verb 'be' in present tense (i.e., 'is/are') presents much greater frequency in CSG compared to GCDT, and verbs such as verbs 'know' and 'have' and the auxiliary verb 'be' in past tense appear more frequent compared to the other two reference corpora.

Regarding the use of adjectives, we noticed that, among the ten most frequent, there are adjectives such as 'first', 'second', 'many', 'small' and 'sure' which appear in CSG, with the adjectives 'good', and 'big' presenting greater frequency in CSG compared to CGDT. However, adjectives such as 'beaten' and 'drunk' are considerably infrequent in CSG as in other two reference corpora.

Some of the most frequent adverbs in GCDT are used more frequent than in CSG such as 'after', 'when', 'there', 'together', 'up', 'in', 'before', and 'out'. However, the adverbs 'nice 'and 'much' present a much higher frequency of appearance in CSG.

Regarding pronouns, the three most frequent ones in GCDT, 'my', 'I' and 'his' are a lot rarer in CSG.

Concluding, we noticed that the CSG reference corpus presented more similarities with GCDT, regarding the POS frequencies, than the other two reference corpora. This is explained partially from the fact that CSG stems from oral speeches and conversations, likewise GCDT. In contrast, the other two reference corpora, HNC and CGT, which derived from written texts, seem to have more differences in the use of speech, since written text can be significantly more precise. Written words can be chosen with greater deliberation and thought, and a written argument can be extraordinarily sophisticated, intricate, and lengthy. On the other hand, speeches can also be precise, but precision in oral communication comes only with a great deal of preparation and compression. Once spoken, words cannot be retracted.

4.2.3. GCDT vs. GCWT

The previous reference corpora consist exclusively of written language material or transcriptions from oral speeches and aim to be representative of the Greek general language. However, the defendants use specific vocabulary during the trial procedure. To achieve more accurate statistical results and to be methodologically correct, we constructed a reference corpus with similar stylometric features to our study corpus. The new reference corpus which derived from witnesses' testimonies related to murder cases, namely GCWT, have been constructed, as mentioned in previous section, from the transcriptions of the court spoken language during the trial procedure. The size of the GCWT is four times greater than the study corpus, quite close to the ideal size of a reference corpus (Berber-Sardinha, 2000; Koppel et al., 2002).

First, we measured the most frequent words in both corpora (Table 4.2.18). The Wordsmith WordList tool gave us a list of all the words in GCDT and GCWT in frequency order. As we expected, in both corpora, the top of this list is occupied by function words, such as 'and', 'the', 'to', 'not', 'with', 'that', etc., with the word 'and' holding the 4% of the total corpus size. The most frequent 15 words in the list take up approximately one third of the corpus, in both cases.

	GCDT			GCWT		
s/n	**Word**	**Freq.**	**Cumulative freq. %**	**word**	**Freq.%**	**Cumulative freq.%**
1	and	4.06	4.06	and	3.93	3.93
2	the	3.68	7.74	the	3.52	7.45
3	to	3.39	11.13	not	2.89	10.34
4	not	3.3	14.43	to	2.27	12.61
5	me	2.71	17.14	him	2.2	14.81
6	with	2.25	19.4	that	2.06	16.87
7	him	1.92	21.31	he	2.04	18.91
8	her	1.7	23.01	my	1.95	20.86
9	that	1.5	24.51	was	1.93	22.78
10	into	1.47	25.98	her	1.8	24.59
11	he	1.43	27.41	of	1.77	26.36
12	these	1.39	28.8	with	1.75	28.11
13	I	1.38	30.18	into	1.55	29.66
14	him	1.31	31.5	from	1.36	31.02
15	for	1.23	32.73	she	1.23	32.25

Table 4.2.18 Most frequent words in GCDT and in GCWT

Table 4.2.19 shows the percentage of word types with frequency one and two in the corpus, namely the hapax and dis legomena. It is depicted that in GCDT the hapax legomena take up almost 50% of the word types, while in GCWT they occupy almost 46%. The TTR in GCDT, i.e., the number of distinct words (types) is just 8.7% of the total number of words (tokens). Similarly, the TTR of GCWT is 5.66%. This means that GCDT is more lexically rich than GCWT, or in other words the defendants use more unique words, compared to the total number of words, than the witnesses.

Corpora	**Tokens**	**Types**	**Hapax Legomena**	**Dis legomena**	**TTR%**
GCDT	108403	9440	49.61	15.40	8.70
GCWT	391819	22177	45.96	15.47	5.66

Table 4.2.19 Lexical richness of GCDT and GCWT

Subsequently, we measured the frequencies of content words (CW) and the FW frequencies of both corpora (Table 4.2.20). The lexical and the functional density of both corpora, also shown in the same table, are almost equal (approximately 45%). The fact that the lexical density of GCWT is slightly larger, depicts that the witnesses' testimonies are more informative than the defendants' speech. This is justified from the fact that

GCWT, apart from testimonies of laymen, it also contains testimonies from specialized witnesses, such as forensic pathologists and police officers, who tend to use more descriptive language and more information-bearing content words due to their profession and educational level.

Corpora	FW frequency %	Lexical Density %	Functional Density
GCDT	55.7	44.3	1.26
GCWT	54.1	45.9	1.18

Table 4.2.20 FW frequency, lexical and functional density of GCDT and GCWT

However, both corpora have low lexical density compared to the typical lexical density of written texts since they are derived from transcriptions of spoken language and are made of special language material. Their lexical density matches the results described in relevant research, where it is mentioned that spoken texts tend to have a lower lexical density (near 45%) than written ones (above 50%) (Johansson, 2008; Ure, 1971).

On the other hand, as function words are inversely proportional to content words, we noticed that FW frequency and functional density is higher in GCDT (55.7% and 1.26, respectively), than in GCWT (54.1% and 1.18, respectively) since the defendants seem to use more function words in their speech than the witnesses.

In Table 4.2.21 we depict the average word length and word length standard deviation (in characters) and the average sentence length and sentence length standard deviation (in words). Having made the appropriate measurements, we found that there are slight differences between the defendants' and the witnesses' speech.

Corpora	Avg word length	Word length st.dev.	Avg sentence length	Sentence length st. dev.
GCDT	4.44	2.27	8.27	6.32
GCWT	4.64	2.54	8.76	6.46

Table 4.2.21 Word and sentence length and standard deviation of GCDT and GCWT

There is a small difference in word length between the two corpora. Witnesses seem to use larger words (4.64 characters) more frequently than the defendants (4.44 characters). The average sentence length for defendants (8.27 words) is shorter than that of witnesses (8.76 words), as is the standard deviation (6.32 words) for defendants compared to

witnesses (6.46 words). Considering the nature of both corpora, the low standard deviations are not surprising. Both corpora derived from testimonies inside a courtroom and apart from some descriptive speech pieces, they contain responses. Typically, defendants and witnesses use one-word or short responses. Moreover, defendants' educational level is lower than the witnesses', using simpler words and shorter sentences.

In order to perform a more qualitative content analysis, we used an approach based on keywords derived analyses.

N	Keyword	Translation	Keyness
1	είπα	I said	645.33
2	πήρα	I took	472.89
3	είχα	I had	443.14
4	έκανα	I did	421.26
5	πήγα	I went	407.38
6	ήθελα	I wanted	405.01
7	να	to	396.66
8	εγώ	I	315.24
9	χτύπησα	I hit	246.74
10	μου	my	222.60
11	πάω	I go	213.33
12	κάνω	I do	175.85
13	μπορούσα	I could	147.74
14	ήμουν	I was	146.71
15	έφυγα	I left	131.47
16	φοβήθηκα	I got scared	129.23
17	έβαλα	I put	128.37
18	θα	will	124.94
19	έπαιρνα	I was taking	114.53
20	με	with	113.13
21	σκέφτηκα	I thought	105.39
22	σκοτώσω	kill	94.96
23	πάμε	we go	94.92
24	ναι	yes	91.39
25	πήγαινα	I was going	89.39

Table 4.2.22 First 25 positive keywords. Study corpus: GCDT, and reference corpus: GCWT

We used the Word Smith KeyWords tool to compare the word list extracted from our study corpus, GCDT, to a word list extracted from the reference corpus, GCWT. We took a list with words that are significantly more frequent in GCDT than in GCWT. Table 4.2.22 depicts the list of the first 25 positive keywords, i.e., the keywords with maximum positive keyness. The higher the value of keyness, the more unusually frequent the word appears in GCDT compared to GCWT.

The list mainly consists of verbs in the first person, singular number, past tense. They are used to describe an action or a feeling of the defendant before, during and after the event in question. Indicatively, the most unusually frequent words of GCDT compared to GCWT are the verbs 'I said', 'I took', 'I had', 'I did', 'I went', 'I wanted', 'I hit', 'I left', 'I was', 'I got scared', 'I thought', etc.

Table 4.2.23 depicts the first 25 negative keywords, i.e., the keywords with the lower negative value of keyness, i.e., words of GCDT that appear quite infrequent compared to the reference corpus. The lower the value of keyness, the more unusually frequent the word appears in the reference corpus of GCWT compared to our study corpus GCDT.

In contrast to the Table 4.2.22, the list consists of verbs in the third person, singular number, past tense, since the witnesses are used to describe an action of someone else. Indicatively, the most unusually frequent words of GCWT compared to GCDT are the verbs 'he/she had', 'was/were', 'they had', 'he/she did', referring to the defendants, and the verb 'was found', referring to the victim. Moreover, four of the twenty-five words of the keyword list are different stems of the same lemma of the word 'defendant', i.e., 'defendant (he)', 'defendant (she)', 'defendant's (genitive case)', 'the defendant (accusative case)'.

At this point we should clarify that we did not proceed to stemming, since we were interested in the frequency of the POS separately, as well as the use of different tenses. Another word that appears in this list is the word 'victim', since defendants seems to rarely refer to that term, and the word 'duty' which is used frequently by police officers who testify as witnesses.

N	Keyword	Translation	Keyness
1	είχε	he/she had	-323.64
2	κατηγορούμενος	defendant (he)	-289.07
3	ήταν	was/were	-262.05
4	μας	us	-215.34
5	ο	the (he)	-176.18
6	η	the (she)	-170.91
7	κατηγορούμενο	the defendant	-148.77
8	ότι	that	-146.84
9	του	his	-115.63
10	θύμα	victim	-105.16
11	κατηγορουμένου	defendant's	-94.11
12	κατηγορούμενη	defendant (she)	-92.63
13	γνωρίζω	I know	-89.82
14	γιος	son	-86.98
15	είχαν	they had	-74.82
16	βρέθηκε	was found	-69.81
17	είναι	is/are	-66.38
18	της	her	-64.35
19	από	from	-60.16
20	βρήκαμε	we found	-58.67
21	υπηρεσία	duty	-58.60
22	οι	the (they)	-55.79
23	άκουσα	I heard	-53.41
24	των	their	-50.94
25	έκανε	he/she did	-50.31

Table 4.2.23 First 25 negative words. Study corpus: GCDT, and reference corpus: GCWT

4.2.4. GCDT vs. pre-GCDT

As we mentioned in a previous section, the pre-GCDT concerns 55 of the 124 defendants of the GCDT, for whom we had their testimonies in front of an interrogator before their trial. In order to compare the style of the defendants during their testimony in front of a judge and in front of an interrogator, we constructed a smaller corpus which is a part of GCDT and included only the 55 corresponding testimonies of the defendants in front of the judge. To define the stylometric profile of GCDT and pre-GCDT corpora, we measured some sets of stylometric features.

Firstly, we measured the average word length and the average sentence length, and the standard deviation of word length and sentence length, which are displayed in Table 4.2.24.

Corpora	**Avg word length**	**Word length st.dev.**	**Avg sentence length**	**Sentence length st.dev.**
GCDT (part)	4.44	2.25	8.64	7.28
pre-GCDT	4.74	2.59	16.67	10.62

Table 4.2.24 Descriptive statistics of GCDT (part) and of pre-GCDT

The average word length (per testimony) measured in characters in pre-GCDT is 4.74 and in the part of GCDT is 4.44. The word length standard deviation is 2.59 and 2.35, respectively. It seems that in their interrogation before their trial, the defendants use slightly larger words and with more variety than inside the court. The length of defendants' sentences is more "spread out" during their interrogation, since the average sentence length in words in pre-GCDT is 16.67 which is almost twice that of the GCDT's which is 8.64. Both corpora are derived from testimonies in front of an interrogator or a judge, so apart from some descriptive speech parts, they also contain responses. Typically, defendants use one-word responses or short sentences. Moreover, the defendants tend to use simpler words and shorter sentences inside the court, whereas in front of the interrogator the defendants' speech seem to be more spontaneous and unplanned. The latter is also explained by the fact that in front of the interrogator, defendants are emotionally charged due to the fact that the crime and their arrest are recent, whereas inside the court their speech is more structured since they usually have time to plan and edit their speech.

Table 4.2.25 shows the number of word types with frequency one and two in the corpora, namely the hapax and dis legomena. The proportion of hapaxes reflects the quantity of different words used in the text and describes the richness of the vocabulary. In both corpora, hapax legomena take up at least 50% of the word types and dis legomena proportion is approximately 16%. The TTR is 10.5% and 11.18% in part of GCDT and the pre-GCDT, respectively. This means that defendants' speech in front of an interrogator is more lexically rich than their speech inside a courtroom. This means that the defendants use more distinct words in front of the interrogator than inside the

court. On the other hand, dis legomena are slightly more frequently used inside a courtroom than during their interrogation.

Corpora	Tokens	Types	Hapax Legomena	Dis legomena %	TTR%
GDCT (part)	48289	5108	50.78	16.51	10.50
pre-GCDT	54032	6412	51.30	15.60	11.18

Table 4.2.25 Lexical richness of GCDT (part) and pre-GCDT

Both corpora underwent POS-tagging, and the results are shown in Table 4.2.26. Lexical density is 44.16% and 45.66% in the part of GCDT and in pre-GCDT, respectively. On the other hand, the FW frequency of the part of GCDT is 55.84% and of the pre-GCDT is 54.34%, and functional density of the part of GCDT is 1.264 and of the pre-GCDT is 1.189. The fact that the lexical density of pre-GCDT is slightly larger in conjunction with the larger number of function words in the part of GCDT, depicts that the defendants' testimonies in the interrogation are more informative than the defendants' speech inside the court, since they use more content words.

Corpora	FW frequency %	Lexical Density %	Functional Density
GDCT (part)	55.84	44.16	1.264
pre-GCDT	54.34	45.66	1.189

Table 4.2.26 FW frequency, lexical and functional density of GCDT (part) and pre-GCDT

We also used an approach based on keywords derived analysis in order to discover significant words in these corpora, comparing the word list extracted from the part of GCDT with the word list extracted from pre-GCDT. Table 4.2.27 depicts the list of the first 20 positive keywords.

We note that the reference corpus (pre-GCDT) is larger in words than the study corpus (part of GCDT), hence it satisfies the requirement for a word list to be accepted as reference corpus by the Wordsmith tool to be larger than the study corpus. These keywords are unusually frequent in the speech of defendants inside the courtroom compared to their speech during the interrogation phase. Indicatively, some of these words are 'I said', 'he/she said', 'not', 'it was', 'yes', 'no', 'I will', 'I knew', 'I wanted', etc.

N	Keyword	Translation	Keyness
1	είπα	I said	406.66
2	είπε	he/she said	235.36
3	δεν	not	233.57
4	εγώ	I	232.81
5	ήταν	it was	151.82
6	ναι	yes	97.81
7	θα	I will	83.93
8	ήξερα	I knew	79.41
9	όχι	no	77.02
10	ξέρω	I know	60.66
11	πήρα	I took	53.9
12	πήγα	I went	51.74
13	τι	what	47.21
14	ήθελα	I wanted	44.66
15	ήμουν	I was	43.69
16	μην	not	43.39
17	είπαν	they said	40.64
18	λεφτά	money	40.64
19	τα	the	38.6
20	πήρε	he/she took	37.36

Table 4.2.27 First 25 positive keywords

Table 4.2.28 depicts the first 20 negative keywords, i.e., those with the lower negative value of keyness, which appear quite infrequent compared to the reference corpus. Indicatively, some keywords that appear unusually frequent in the pre-trial testimonies compared to the trial testimonies are the words 'about', 'from', 'in', 'where', 'while', 'nothing', etc.

Comparing these two tables we conclude that the defendants use some specific verbs in past tense inside the court, whereas in the pre-trial phase the defendants use more adverbs and prepositions.

N	Keyword	Translation	Keyness
1	περίπου	about	-135.66
2	από	from	-123.34
3	στη	in	-111.11
4	προσθέσω	add	-109.64
5	όπου	where	-107.11
6	οποίο	which	-96.01
7	τη	the (she)	-95.91
8	της	her	-94.08
9	οδό	road	-81.47
10	ενώ	while	-71.37
11	τίποτε	nothing	-67.55
12	οποίος	who (he)	-63.73
13	καθώς	as	-52.95
14	οποία	who (she)	-52.18
15	και	and	-49.83
16	βρίσκεται	is located	-48.43
17	συνέχεια	continuity	-48.39
18	γνωρίζω	I know	-45.07
19	του	his	-44.47
20	σας	your	-42.09

Table 4.2.28 First 25 negative keywords

4.3. Results

The stylometric analysis that our corpora underwent revealed some interesting conclusions. Firstly, we deduced that defendants' testimonies follow specific linguistic patterns. Among the most frequent nouns that are used are words such as 'home', 'gun', 'phone', 'police', and 'money', i.e., nouns relevant with a crime that has been committed. Defendants that belong to the category 'above 50', use words 'son' and 'daughter', whereas those of the category '20-34' use the word 'mother' and 'father', according to their marital status which agrees with their age.

The most frequent verbs that defendants use, are those which describe their actions in the past tense, such as 'I/he/she said', 'I went', 'I saw', 'I/he/she had', 'I was', 'I/he/she took', 'it happened', 'he/she came', and the verbs 'I don't know' and 'I don't remember'. Moreover, we noticed limited use of adjectives in the defendants' speech and use of basic

adverbs. We could characterize their vocabulary as poor, due the nature of their speech, but also due to the fact that the majority of them have low educational level.

Typically, all defendants use one-word or short responses. However, the older defendants, comparing to younger ones, tend to use more complicated words and longer sentences, and they also use more descriptive language and more information-bearing content words. The same assumption was derived for native speakers comparing to non-native ones. Non-native speakers seem to have even lower educational level than the native ones, thus their vocabulary can be described as elementary. Moreover, we noticed that non-native speakers use specific words, such as the word 'boss', which are absent in native speakers' speech.

Comparing GCDT with Greek general language corpora stemming from written texts, we noticed a large variance in the frequencies of occurrence of words between them, not only for nouns where we would expect a higher frequency of occurrence in GCDT for specific words such as 'knife', 'money', 'gun', 'police', 'prison', but for other nouns used such as 'telephone' and 'mother'. Adjectives such as 'beaten' and 'drunk' seemed to be considerably infrequent in colloquial language comparing to GCDT, whereas basic adverbs such as 'after', 'there', 'together', 'in', 'out', etc., were used more frequently in defendant's speech since the nature of the testimony language tends to be descriptive. However, comparing GCDT with Greek general language corpora stemming from oral speeches and conversations, we noticed more similarities in contrast to the other reference corpora. For instance, nouns such as 'kid/kids', 'hour' and 'years' seem to have similar frequency with GCDT, the verb 'be' in present tense presents much greater frequency compared to GCDT and the verb 'know' and auxiliary verbs such as, 'have' and 'be' in past tense appear more frequent compared to the other two reference corpora. This can be partly explained from the fact that the reference corpus that stems from oral speeches consists of more colloquial words likewise GCDT. In contrast, the other two reference corpora, which derived from written texts, seem to have more structured speech, since the words that are used in written texts can be significantly more precise, sophisticated, elaborate, and complex.

From the comparison of GCDT with GCWT we concluded that defendants use more unique words than the witnesses, whereas witnesses' testimonies are more informative

than the defendants' speech. The latter is justified by the fact that GCWT contains testimonies from specialized witnesses, such as forensic pathologists and police officers, who tend to use more descriptive language and more information-bearing content words, compared to the defendants' speech whose average educational level is lower than the witnesses' and thus they tend to use simpler words and shorter sentences. We confirmed that defendants use verbs in the first person in singular number at past tense, which are used to describe an action or a feeling of the defendant before, during and after the event in question. In contrast, witnesses use verbs in the third person in singular number at past tense, since they describe an action of someone else. Moreover, witnesses use unusually frequently, comparing to defendants, the word 'defendant' and the word 'victim', since defendants seems to rarely refer to these two terms, and the word 'duty' which is used frequently by police officers who testify as witnesses. However, apart from their differences, defendants seem to have more stylometric features in common with witnesses than with general language. For instance, FW frequency is a little higher in GCDT than in GCWT, and much higher than in GGLC, since the defendants seem to use a little more function words in their speech than the witnesses but much more function words than in the general language. Also, the value of keyness is much higher between GCDT and GGLC keywords, than between GCDT and GCWT keywords, which denotes that defendants use more unusually frequent words compared to general language than to witnesses.

The final comparison verified our assumption that the style of defendants' speech differs depending on whether they testify during the interrogation phase or inside the courtroom. Thus, in the interrogation phase the defendants use slightly larger words more frequently and with more variety than inside the court. The length of defendants' sentences is more "spread out" in front of an interrogator than in front of a judge. Typically, the defendants tend to use simpler words and shorter sentences inside the court, whereas during their interrogation the defendants' speech seem to be more spontaneous and unplanned. The latter is also explained by the fact that in front of the investigator, defendants are emotionally charged due to the fact that the crime and their arrest are recent, whereas inside the court their speech is more structured since they usually have time to plan and edit their speech. Moreover, the defendants' testimonies in the interrogation are more informative than the defendants' speech inside the court.

4.4. Summary

In this chapter we fulfilled our second objective of our study, i.e., the quantification of the defendants' speech. In other words, we measured several linguistic features of their testimonies, i.e., lexical, syntactic, and content-specific, and we observed that their speech follows some linguistic patterns. We confirmed that the stylometric features they use, differ from those of the general language, due to the nature of the testimony language. Moreover, we noticed that the speech of the defendants accused of murder differs from each other depending on their demographic and social characteristics. For instance, age and nationality plays a decisive role in the way they speak. Similarly, defendants' speech in front of a judge inside a courtroom differs with that before their trial during their interrogation, since inside a courtroom their speech is more structured and their psychological state is calmer, compared to that during their interrogation where they are more anxious and their speech more unprompted. Finally, comparing the defendants' language inside the court with that of the witnesses', we denoted several similarities, since the style of both corpora belongs to the same genre, but also several differences due to the fact that the average education level of witnesses is higher, and their psychological state is more stable and rational than the defendants who are accused of a crime and try to defend themselves.

5 THE GCDT MACHINE LEARNING CLASSIFIER

So far, among others, we created a corpus from defendants' testimonies, a corpus from witnesses' testimonies and a corpus from pre-trial testimonies. From these corpora we calculated several standard stylometric variables such as hapax legomena, dis legomena, lexical density, functional density, average word and sentence length, word and sentence standard deviation, and most frequent words. The statistical analysis of these data showed that most of them can characterize the linguistic profile of our study corpus. Therefore, the most effective of these variables were used to train a machine learning classifier.

In this section we present our text classifier, namely GCDT classifier, whose output answers the question of whether a testimony belongs to a murderer or not. Briefly, we loaded GCDT classifier with testimonies of both guilty and not guilty persons, the classifier's algorithm found correlations between testimony and verdict, and in case we gave a new testimony to the classifier it could predict whether the testimony belonged to a murderer or not.

5.1. Description of GCDT classifier

The first step in creating our classifier was deciding which features of our dataset were important, and how to encode these features. Selecting relevant features and deciding how to encode them for a learning method can have an enormous impact on the learning method's ability to extract a good model. Usually, there are limits to the number of features that should be used with a given learning algorithm. If too many features are provided, then the algorithm will have a higher chance of relying on idiosyncrasies of the training data that cannot generalize well to new examples. This problem is known as overfitting and can be especially problematic when working with small training sets.

In view of the foregoing, we chose the features of the standard stylometric variables that we have already calculated during our corpora statistical analysis, that played a decisive role in characterizing the stylometric profile of the defendants. Hence, the linguistic features which were used as training data are the number of words, hapax legomena, dis legomena, number of content words, number of function words, lexical density, function

words' frequency, functional density, average length of words in characters, average length of sentences in words, average standard deviation of words and the average standard deviation of sentences.

We defined as guilty the defendants whose final verdict was convicting (111 defendants), and as not guilty those whose verdict was acquittal (13 defendants). Due to the fact that the innocent defendants were few in relation to the guilty ones, in the category of not guilty we included all the witnesses, too. By making this assumption, we managed to balance the number of the two target classes, considering the fact that the speech of both defendants and witnesses has similar stylistic features. However, due to their role in the judicial process, defendants' speech tends to be apologetic and said in the first person, which is not the case with witnesses. This assumption might cost us in terms of the classifier's accuracy, but we chose to test our classifier's efficiency.

In total our training data consists of a matrix of 269 rows and 12 features. Every row represents either a murderer or a witness. Namely, 124 rows signify the defendants, and the rest 145 rows indicate the witnesses, both prosecution and defense. From the 124 defendants 111 have found guilty and the other 13 were found not guilty. Thus, the 13 records of the not guilty defendants were added in the number of witnesses. Therefore, in total the matrix contains 111 records of guilty and 158 records of not guilty persons (Figure 5.1.1).

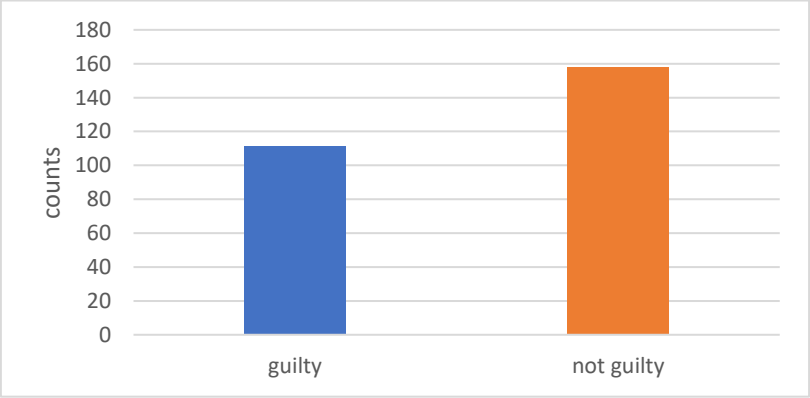


Figure 5.1.1 Number of guilty (1) and not guilty persons (0)

Every cell of this matrix contains a value. The columns of the matrix which contains features such as hapax legomena, dis legomena, number of words, function words, and

content words have values that indicate the number of their appearance. The columns which contain features such as lexical density, functional density, frequency of function words, average word length, average sentence length, word length standard deviation and sentence length standard deviation have values which denote a ratio. However, all values have undergone normalization, therefore they ended up ranging between 0 and 1. Thus, all the values of the matrix are decimal numbers. The value of the target class is either 1 or 0, depending on whether the text belongs to a guilty or to a not guilty person.

Figure 5.1.2 depicts a sample of the training data that were used in GCDT classifier.

text	hapax	dis	num_of_words	content words	lexical_density	function words	function words freq	functional_density	Avg_word_length	Word_length_std_dev	Avg_sentence_length	Sentence_length_std_dev	guilty
d111	0.0142431	0.0095020	0.0071672	0.0100668	0.42	0.006	0.58	1.4	4.44	2.19	6.57	4.55	1
d112	0.0143962	0.0117378	0.0059425	0.0085804	0.43	0.005	0.57	1.33	4.29	2.12	7.71	6.51	0
d113	0.0238916	0.0178862	0.0081509	0.0140530	0.51	0.005	0.49	0.95	4.95	2.71	7.96	5.89	1
d114	0.0225133	0.0285062	0.0175266	0.0262818	0.45	0.013	0.55	1.24	4.57	2.34	6.51	3.71	1
d115	0.0398194	0.0268293	0.0190323	0.0291194	0.45	0.014	0.55	1.2	4.7	2.51	7.97	4.8	1
d116	0.0398194	0.0424798	0.0234089	0.0365513	0.46	0.017	0.54	1.16	4.68	2.46	10.93	8.69	1
d117	0.0298645	0.0368904	0.0218631	0.0335110	0.46	0.016	0.54	1.2	4.6	2.41	9.28	10.5	1
d118	0.0246574	0.0212399	0.0144750	0.0209444	0.43	0.011	0.57	1.33	4.65	2.59	11.3	12.35	1
d119	0.0385942	0.0536587	0.0350733	0.0522258	0.44	0.026	0.56	1.26	4.51	2.3	7.79	5.14	1
d120	0.0390536	0.0352135	0.0201967	0.0297951	0.44	0.015	0.56	1.28	4.66	2.49	8.38	5.09	1
d121	0.0271078	0.0318598	0.0143344	0.0214173	0.44	0.011	0.56	1.25	3.82	2.8	5.86	4.21	1
d122	0.0294051	0.0318598	0.0146356	0.0232415	0.47	0.01	0.53	1.12	4.29	2.76	8.02	6.39	1
d123	0.0324681	0.0268293	0.0147962	0.0240522	0.48	0.01	0.52	1.07	4.95	2.69	10.84	7.53	1
d124	0.0306303	0.0240346	0.0130897	0.0197282	0.45	0.01	0.55	1.23	4.77	2.66	8.82	5.46	1
wp1	0.1464130	0.1425311	0.1468783	0.1214099	0.25	0.148	0.75	3.07	4.85	2.9	11.09	9.37	0
wp2	0.0863775	0.0877544	0.0621763	0.0956010	0.46	0.045	0.54	1.19	4.79	2.81	10.31	10.08	0
wp3	0.0379816	0.0363314	0.0204376	0.0314841	0.46	0.015	0.54	1.18	4.7	2.56	6.72	4.39	0

Figure 5.1.2 Training data features of GCDT classifier (sample)

We added the first row only for explanatory reasons. The first cell of every row has a code name, declaring the serial number of every observation (text). For instance, 'd111' belongs to the testimony of the 111th defendant and 'wp1' belongs to the testimony of the 1st witness of prosecution. The other columns denote the number of hapax legomena ('hapax'), the number of dis legomena ('dis), the number of words ('#words'), the number of content words ('#CW'), the lexical density ('LD'), the number of function words ('#FW'), the frequency of function words ('FW freq'), the average word length in characters ('AvgWordLen'), the word length standard deviation ('WordLenStd'), the average sentence length in words ('AvgSentenceLen'), and the sentence length standard deviation ('SentenceLenStd'). Finally, the last column denotes the value of the target class ('guilty'). Therefore, in total the matrix has 14 columns.

Figure 5.1.3 shows the frequency distribution of word length (in characters) in every guilty and in every not guilty person.

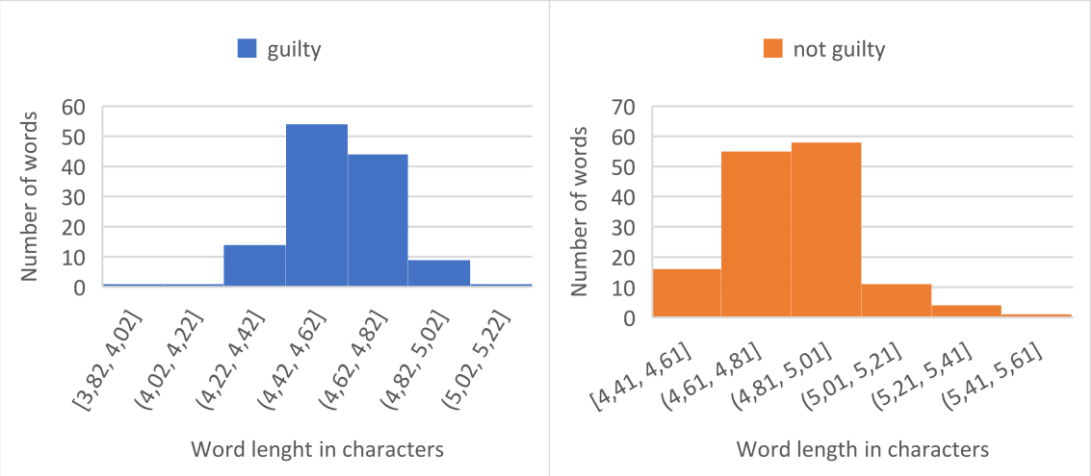


Figure 5.1.3 Frequency distribution of word length of every text

The bulk of the words in the category 'guilty' is between 4.42 to 4.82 characters per word, whereas in the category 'not guilty' is between 4.61 to 5.01 characters per word.

Similarly, Figure 5.1.4 shows the frequency distribution of sentence length (in words) in every text of guilty and not guilty person. For instance, in the guilty defendants' texts, 12 sentences consist of 4-6 words, 58 sentences consist of 6-8 words, 38 sentences consist of 8-10 words, etc. The majority of sentences in the category 'guilty' is between 6 to 10 words per sentence, whereas in the category 'not guilty' is between 6.4 to 10.4 words per sentence.

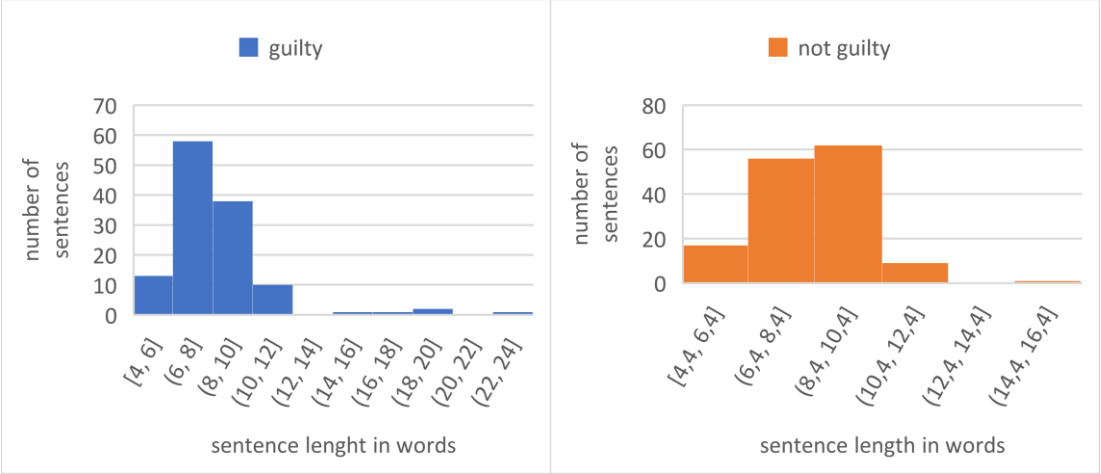


Figure 5.1.4 Frequency distribution of sentence length of every text

As we forementioned in Subsection 4.2.3, witnesses seem to use larger words more frequently than the defendants. Even if we do not have the necessary metadata related to the educational level of witnesses, we assume that the use of larger words and sentences relies on the fact that among them there are more specialized scientists, such as forensic surgeons, police officers, etc. Therefore, the defendants' educational level, on average, is lower than the witnesses', since the metadata of defendants show that regarding their occupation, most of them are workers, farmers, builders, freelancers or unemployed. That explains their tendency to use simpler words and shorter sentences compared to witnesses. The fact that among witnesses there are more specialized scientists makes their testimony more descriptive, since their goal is to give a report of what happened, and their answers are not summarized in one-word responses.

5.2. Evaluation of GCDT classifier

The classification algorithm we built considers each training text sample as a unit that contributes separately to the attribution model. In other words, each text sample of known 'class' is an instance of the problem in question. In detail, each text sample of the training corpus is represented by a vector of features, described in Section 4.1, and the classification algorithm is trained using the set of text instances of known class in order to develop the attribution model. Then, this model will be able to estimate the true class of an unseen text. The training of a machine learning classifier is a demanding job. The purpose of training a classifier is to acquire ability to give the desired set of outputs after

a set of inputs. The GCDT classifier algorithm is built in Python 3.7[21] using the SciPy 1.4[22] platform and the Scikit-learn 0.22[23] tool for predictive data analysis.

Training data is randomly selected from our matrix features data set. To evaluate our model, we must reserve a portion of the annotated data for the test set. If the test set is too small, then our evaluation may not be accurate. However, making the test set larger usually means making the training set smaller, which can have a significant impact on performance if a limited amount of annotated data is available.

Therefore, we split our initial dataset in two parts, 80/20, i.e., 80% of the data is used as training and validation dataset for the algorithms' evaluation, and the rest 20% of the data is held back and it is used as the testing dataset in the prediction phase of GCDT classifier.

5.2.1. Validation

In order to evaluate our model, we used the metric of accuracy. Accuracy is the ratio of the number of correctly predicted instances divided by the total number of instances in the dataset multiplied by 100 to give a percentage. In order to check which classification algorithm performs well in our problem or what configurations to use, we tested six different algorithms that we described in Subsection 2.2.5, which are the most common algorithms used for classification problems as ours. Thus, the algorithms we tested were Logistic Regression (LR), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (k-NN), Classification and Regression Trees (CART), Gaussian Naive Bayes (NB), Support Vector Machines (SVM). The first two are linear algorithms and the rest are nonlinear.

To estimate the accuracy of our model we used stratified k-fold cross validation in the part of the dataset that we reserved, namely the 80% of the dataset. Particularly, we used stratified 10-fold cross validation. This technique split the training dataset into 10 parts,

[21] https://www.python.org/

[22] https://www.scipy.org/

[23] https://scikit-learn.org/stable/

trained the algorithm on 9 parts and tested it on the remaining 1 part. This procedure was repeated for all combinations of train-test splits. Stratified means that each fold or split of the dataset has the same proportion of observations with a given categorical value, such as the class outcome value, as we described in Subsection 2.2.8.

All algorithms undergone hyperparameter tuning in order to achieve their best performance using the GridSearchCV[24] library of Scikit-Learn. Indicatively, the hyperparameters optimization and the evaluation results for every algorithm is depicted in Table 5.2.1. These results are before making predictions on the test dataset.

Classification algorithm	**Hyperparameters tuning**	**Accuracy mean**	**Accuracy standard deviaton**
Logistic Regression	solver= 'newton-cg' penalty='l2' C = 10 multi_class='ovr'	0.78	0.08
Linear Discriminant Analysis	solver='svd'	0.76	0.08
k-Nearest Neighbors	metric= 'manhattan' n_neighbors= 19 weights= 'uniform'	0.71	0.09
Classification and Regression Trees	criterion= 'gini' max_depth= 3 max_features=6 min_samples_leaf= 6	0.72	0.06
Gaussian Naïve Bayes	var_smoothing= 5.336699231206313e-07	0.72	0.06
Support Vector Machines	C= 50 gamma= 'scale' kernel= 'rbf'	0.66	0.05

Table 5.2.1 Accuracy of classification algorithms

We created a plot of the models evaluation results and compared the spread and the mean accuracy of each model. There is a statistical population of accuracy measures for

[24] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

each algorithm because each algorithm was evaluated 10 times via 10-fold cross validation. A useful way to compare the samples of results for each algorithm is to create a box and whisker plot (or boxplot) for each distribution and compare the distributions.

Briefly, a box and whisker plot is a graph that presents information from a five-number summary. It is especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set. It is also very useful when large numbers of observations are involved and when two or more data sets are being compared. It is a way of summarizing a set of data measured on an interval scale. It is often used in explanatory data analysis. This type of graph is used to show the shape of the distribution, its central value, and its variability. The ends of the box are the upper and lower quartiles, so the box spans the interquartile range. The median is marked by a line inside the box, and the whiskers are the two lines outside the box that extend to the highest and lowest observations.

In our case (Figure 5.2.1), we see that five out of six box and whisker plots achieve maximum score of accuracy more than 80%. The two best algorithms seem to be LR and LDA, with maximum values of accuracy equal to 0.87 (87% accuracy) and 0.86 (86% accuracy) respectively. The maximum value of accuracy is the largest number of the set. The median value of LR equals to 0.81 (81% accuracy) and the median value of LDA equals to 0.76 (76% accuracy). This means that there are exactly 50% of the elements less than the median and 50% of the elements greater than the median. The minimum value, i.e., the smallest number of the set, of LR is 0.73 (73% accuracy) and the minimum value of LDA is 0.68 (68% accuracy).

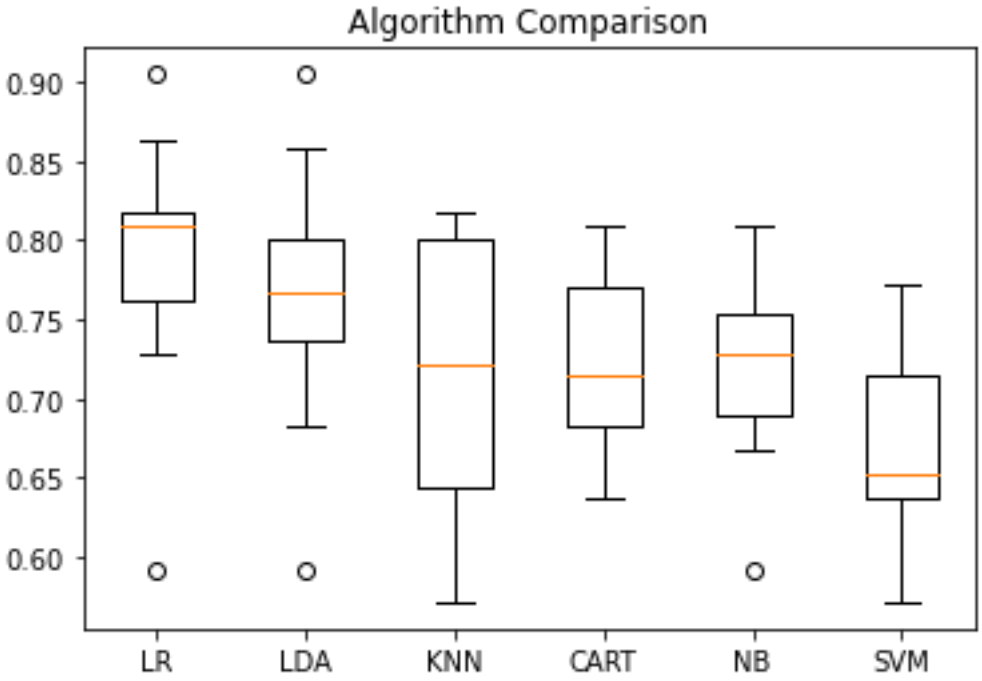


Figure 5.2.1 Box and whisker plot of classification algorithms

5.2.2. Prediction

The results in the previous paragraph showed the accuracy of the machine learning algorithms before testing them in our test set. However, LDA gave the best results after calculating the accuracy of our model on the test dataset, i.e., 92.5% (≈ 93%) accuracy. That was the final check on the accuracy of the best model. It was valuable to keep a test set in case there was an error during training, such as overfitting to the training set or a data leak. Both issues would result in an overly optimistic result. Therefore, we fitted the model on the training dataset and made predictions on the test dataset. Also, we saved the model into a file so we can load it later to make predictions on new data. We evaluated the predictions by comparing them to the expected results in the test set. Then, we calculated the classification accuracy, and we measured the values of a confusion matrix and a classification report.

Figure 5.2.2 is the output of GCDT classifier. As we can see, they are depicted overall accuracy, the confusion matrix, and the classification report. These metrics can give us information about the quality of predictions from our classification algorithm.

```
Accuracy: 0.9259259259259259

Confusion matrix:

[[30  3]

 [1  20]]

Classification report:

class       precision   recall   F1-score   support

0.0            0.97      0.91       0.94         33

1.0            0.87      0.95       0.91         21

Accuracy                            0.93         54

Macro avg      0.92      0.93       0.92         54

Weighted avg   0.93      0.93       0.93         54
```

Figure 5.2.2 Output of the GCDT classifier's evaluation metrics: Accuracy, Confusion Matrix and Classification Report

Accuracy: According to the output of our classifier, accuracy is 0.925 or about 93% on the hold out dataset.

Confusion matrix: The confusion matrix provides an indication of the six errors made. As we described in Subsection 2.2.7, the confusion matrix is a way of tabulating the number of misclassifications. In Table 5.2.2 we depict in more detail the values of confusion matrix.

		Predicted class	
		0 (not guilty)	**1 (guilty)**
Actual class	**0 (not guilty)**	30	3
	1 (guilty)	1	20

Table 5.2.2 Confusion matrix of GCDT classifier

The elements in the main diagonal of the confusion matrix, show the number of correct classifications for each class, i.e., 30 correct predictions of class 0 (not guilty) and 20

correct predictions of class 1 (guilty). The off-diagonal elements provide the misclassifications, for example, 3 wrong predictions of the class 0, which were misclassified as 1, and 1 wrong prediction of the class 1, which was misclassified as 0.

Classification report: Among others, the output of our classifier can show us the classification report which provides for each class values of Precision, Recall, F1-score, and Support. These metrics are calculated by using TP, FP, TN, and FN which were described thoroughly in Table 2.2.2. In case we assume that class 0 (not guilty) is the correct result, then TP equals to 30, FP equals to 1, TN equals to 20 and FN equals to 3. Therefore, the values of the metrics of the classification report, for class=0, are derived from the following formulas:

- Precision = $\frac{\mathrm{TP}}{\mathrm{TP+FP}} = \frac{30}{30+1} = 0.97$
- Recall = $\frac{\mathrm{TP}}{\mathrm{TP+FN}} = \frac{30}{30+3} = 0.91$
- F1-score = $2 * \frac{(Recall*Precision)}{(Recall+Precision)} = 2 * \frac{(0.91*0.97)}{(0.91+0.97)} = 0.94$
- Support value = the number of samples of the true response that lie in that class = 33 samples are not guilty (0)

In case we assume that class 1 (guilty) is the correct result then the TP are equal to 20, FP are equal to 3, TN are equal to 30 and FN are equal to 1. Therefore, the values of the metrics of the classification report, for class=1, are derived from the following formulas:

- Precision =$\frac{\mathrm{TP}}{\mathrm{TP+FP}} = \frac{20}{20+3} = 0.87$
- Recall =$\frac{\mathrm{TP}}{\mathrm{TP+FN}} = \frac{20}{20+1} = 0.95$
- F1-score =$2 * \frac{(Recall*Precision)}{(Recall+Precision)} = 2 * \frac{(0.95*0.87)}{(0.95+0.87)} = 0.91$

- Support value = the number of samples of the true response that lie in that class = 20 samples are guilty (1)

These results mean that GCDT classifier can predict with 91% accuracy the correct result in case a testimony belongs to a guilty speaker.

Therefore, the overall accuracy of our classifier equals to 92.5% (≈ 93%), as it is the mean value of the individual accuracies 94% and 91%. The total Support equals to 54, since it is the number of samples of the true response that lie in both classes. The macro average is simply the average of the values of the respective classes without considering the proportion for each label in the dataset. The classification report metrics are explained below, where P_{c0} and $P_{c1,}$ are the precision values for class 0 and class 1, R_{c0} and R_{c1} are the recall values for class 0 and class 1, $F1_{c0}$ and $F1_{c1}$ are the F1-score values for class 0 and class 1. Thus:

- macro average precision = $\frac{Pc0 + Pc1}{2} = \frac{0.97+0.87}{2} = 0.92$
- macro average recall = $\frac{Rc0 + Rc1}{2} = \frac{0.91+0.95}{2} = 0.93$
- macro average F1-score = $\frac{F1c0 + F1c1}{2} = \frac{0.94+0.91}{2} = 0.93$

The weighted average returns the average considering the proportion for each label in the dataset. The classification report metrics are explained below, where P_{c0} and $P_{c1,}$ are the precision values for class 0 and class 1, R_{c0} and R_{c1} are the recall values for class 0 and class 1, $F1_{c0}$ and $F1_{c1}$ are the F1-score values for class 0 and class 1, and c0 and c1 are the number of instances in class 0 and class 1, respectively. Thus:

- weighted average precision = $\frac{[(Pc0*c0)+ (Pc1*c1)]}{c0+c1} = \frac{[(0.97*30)+ (0.87*20)]}{30+20} = 0.93$
- weighted average recall = $\frac{[(Rc0*c0)+ (Rc1*c1)]}{c0+c1} = \frac{[(0.91*30)+ (0.95*20)]}{30+20} = 0.93$
- weighted average F1-score = $\frac{[(F1c0*c0)+ (F1c1*c1)]}{c0+c1} = \frac{[(0.94*30)+ (0.91*20)]}{30+20} = 0.93$

It is important to note that since the two classes of our classifier are imbalanced, we should consider the weighted average values and not the macro average ones. However, as one can see the results show 93% accuracy in both cases.

ROC and AUC-ROC: The default value of the threshold on which we got the confusion matrix was 0.50. This means that all values equal or greater than the threshold are mapped to one class and all other values are mapped to another class. However, classification problems that have class imbalance, as in our case, using the default threshold can result in poor performance. Instead of constructing several confusion matrices for every threshold, to evaluate the prediction skills of our model, we used the diagnostic tool of ROC curve and AUC-ROC which consolidate the information from several confusion matrices into a single graph. In other words, the ROC graph summarizes all of the confusion matrices that each threshold produces.

Figure 5.2.3 depicts the ROC curve and the AUC-ROC value of GCDT classifier. The y-axis (TPR) shows the proportion of not-guilty (class 0) samples that were correctly classified. The x-axis (FPR) shows the proportion of guilty (class 1) samples that were incorrectly classified. The dashed diagonal line represents a no skill classifier that cannot discriminate between the classes and would predict a random class or a constant class in all cases (AUC = 0.50).

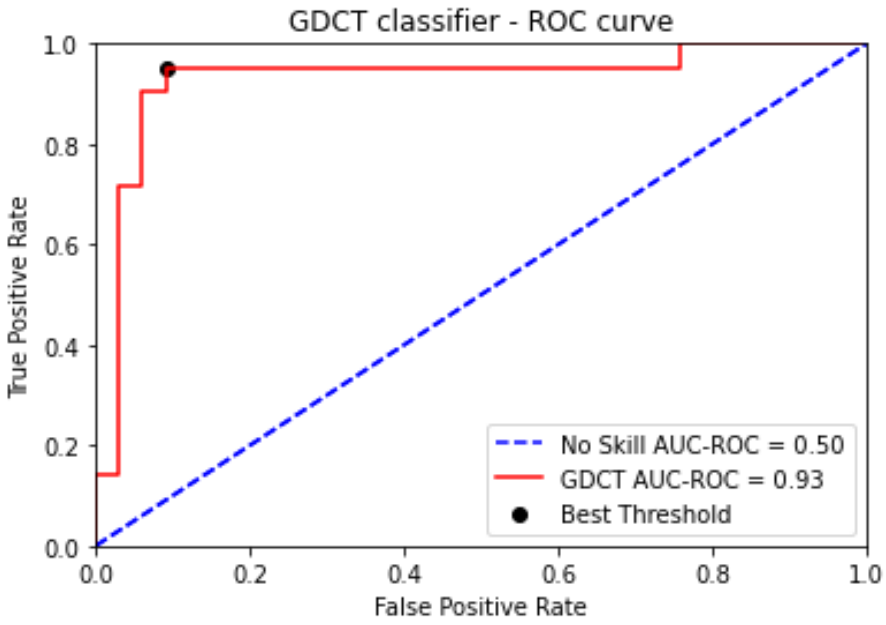


Figure 5.2.3 ROC curve of GCDT classifier

Any point on the dashed line means that the proportion of correctly classified not guilty samples is the same as the proportion of incorrectly classified samples that are guilty.

Any point located further to the left of the dashed line means that the proportion of correctly classified samples that were not guilty (TP) is greater than the proportion of the samples that were incorrectly classified as not guilty (FP). If a new point is even further to the left, means that the new threshold further decreases the proportion of the samples that were incorrectly classified as not guilty. For instance, the best threshold depicted with a black dot is represented by the point (0.091, 0.95) which has correctly classified 95% of the not guilty samples and 90.9% (100%-9.1%) of the samples that were guilty. In other words, this threshold resulted in 9.1% FP.

Moreover, we calculated the AUC-ROC graph. It seems that our classifier is a skillful model, since it is represented by a curve that bows up to the top left of the plot and the AUC-ROC is 0.93.

PR and AUC-PR: Furthermore, we present the PR curve which can better characterize a binary classifier, than a ROC curve, in case a dataset is imbalanced, as in our case. This assumption arises from the fact that our dataset contains 111 records of one class (guilty) and 158 records of the other class (not guilty). This means that the 'not guilty' class is 42.3% greater than the 'guilty' class. Imbalanced classification refers to classification predictive modeling problem, where the number of examples in the training dataset for each class label is not balanced. That is, where the class distribution is not equal or close to equal and is instead biased or skewed. PR curve is more useful in our case because Precision does not include the number of TN in its calculation and is not affected by the imbalance. Figure 5.2.4 depicts the PR curve of GDCT classifier compared to a no skill model.

The PR curve is constructed by calculating and plotting the precision against the recall for GCDT classifier at a variety of thresholds. It visualizes how the choice of a threshold affects the classifier's performance and can help us select the best threshold for our problem. As one can see, a PR curve of a no skill model is a horizontal line with a precision that is proportional to the number of positive cases (class = 1) in the dataset. In our case the ratio of positive cases in the dataset is equal to 111/269=0.41, since 111 are the 'guilty' cases and 269 are the total records of the dataset. This classifier would simply predict that all instances belong to the positive class.

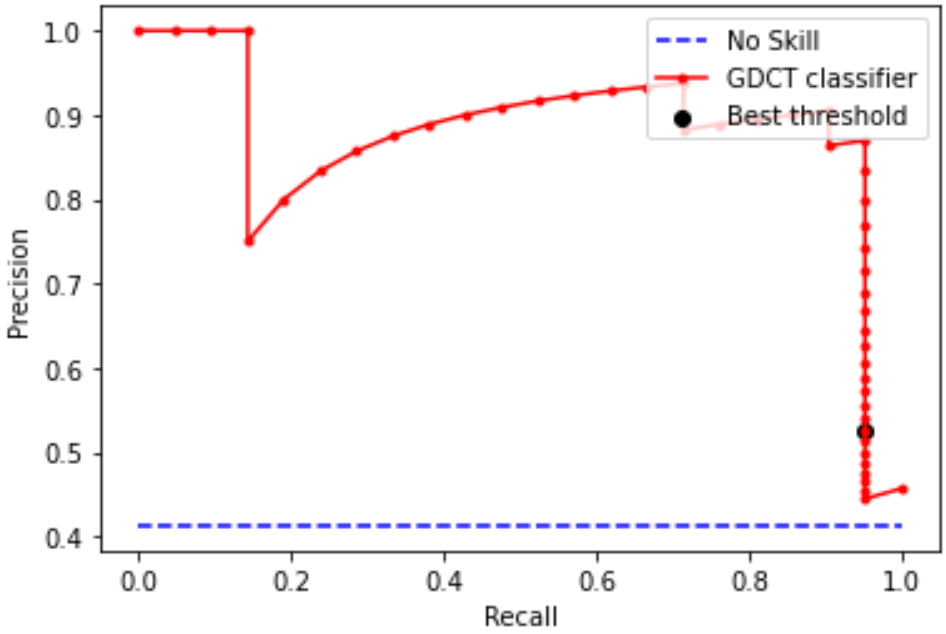


Figure 5.2.4 PR curve of GDCT classifier

Furthermore, the PR curve of GDCT classifier illustrates a model with sufficient skills, since a skillful model is represented by a curve that bows towards a coordinate of (1,1). As we expected, at thresholds with low recall, the precision is correspondingly high, and at very high recall, the precision begins to drop.

Much like ROC curves, we can summarize the information of a PR curve with a single value, that of AUC-PR metric. In a perfect classifier, AUC-PR equals to 1 and a no skill classifier in our dataset would equal to 0.41. A classifier that provides some predictive value will fall between a no skill and a perfect classifier. Figure 5.2.5 shows the AUC-PR value for GDCT classifier which equals to 0.881.

```
GDCT classifier: AUC-PR=0.881
```

Figure 5.2.5 Output of AUC-PR of GDCT classifier

5.3. Summary

In this chapter we managed to fulfill our final and most ambitious goal, which was the construction, from scratch, of a machine learning classifier capable of predicting, with almost 93% of accuracy, whether a written text was the transcribed spoken words of a murderer or not. The accuracy that the forementioned classifier achieved can answer the

question that we had set in the beginning of our study, regarding the reliability of the classifier's results and the trustiness that can offer as an auxiliary and complementary tool in the investigation and interrogation procedure.

The 93% accuracy means that in 9.3 out of 10 times the GCDT classifier can predict correctly that a recorded testimony belongs to a murderer or not. The measurements showed that GCDT classifier can predict correctly with 97% accuracy that a testimony belongs to a not guilty person, and with 87% accuracy that a testimony belongs to a guilty person. The remaining proportion corresponds to the false prediction that either a testimony of a murderer belongs to a non-murderer or testimony of a non-murderer belongs to a murderer. Although the false rate does not seem to be negligible, the first false case is somehow more 'innocent' than the second one. That is explained from the fact that even if a judge blindly relied on the prediction of our model, it would be less serious to acquit a murderer than to accuse an innocent person. Of course, our model does not replace a judge's decision but can offer to the trial procedure an additional tool in evaluating a murderer's testimony.

The forementioned accuracy of GCDT classifier was achieved by constructing a training dataset using the most efficient stylometric features which have been proven to characterize the corpus in question. The features' selection was made by keeping a balance between accuracy and complexity. The number of the used features are 12. Theoretically, by increasing the number of the features used, the accuracy of the model would increase, but so would the complexity of the model, and vice versa.

Thus, we concluded that the use of the aforementioned features performed well in terms of accuracy and complexity, since a 93% accuracy is more than acceptable considering the short time complexity of our model in tandem with the assurance of algorithmic fairness and interpretability.

6 DISCUSSION

The major findings of our thesis are the creation of corpora derived from testimonies inside a courtroom in Greece. The volume of the linguistic material and the difficulty of collecting it are two factors that give extra value to our research. These corpora are a good basis for further research in the field of Greek Forensic Linguistics which lacks linguistic material gathered in real life conditions, i.e., inside a courtroom in front of a judge or in front of an interrogator. A relevant Greek doctoral thesis which concerns written and spoken courtroom discourse in military justice, aims to simplify the written legal texts and identifies the most important language problems of the spoken courtroom discourse by analyzing the texts of criminal court decisions, their written records and the testimonies given before the pre-trial authorities (Kapopoulos, 2021). Compared to the aforementioned research which has focused on the language problems identified in the legal procedures, our study focuses on the stylometric features of the defendants in relation to the general language and the language of witnesses.

Our dataset, which consist of criminal court decisions, stems from transcriptions of spoken testimonies into writing form. The process of transcription may contain corrections in grammar and syntax, thus there is an error rate during the process of transcriptions which cannot be verified without the corresponding audio recordings.

Another assumption is the unavoidable loss of precision on the defendant's speech in case where an interpreter was interfered. Even though there is a consensus that the quality of interpretation needs to be improved, it is clear that interpretation will never be perfect and completely accurate. However, courts should be particularly aware of the problems created by interpretation and try to correct them. Taping court proceedings is already being followed in Greece, but we have not been able to access this data.

Our findings suggest that defendants use specific stylometric features during their testimony inside a courtroom. The comparison of the linguistic features of the defendants' testimonies with those of the general language shows clear differences between them. These differences are predictable given that defendants use a specific vocabulary relevant to their case during their testimony. This conclusion is reinforced by the fact that the majority of the defendants for murder belong to a specific social class,

something that is confirmed by the metadata we collected from the proceedings of the trial. In particular, the defendants of the cases we examined are workers, farmers, builders, freelancers or unemployed, therefore their linguistic features are characterized by limited and poor vocabulary. Comparing the stylometric features of defendants' testimonies with those of the witnesses' testimonies, there seem to be several similarities but also several differences, as well. The similarities are justified due to the fact that the speech of defendants and witnesses has similar stylistic features, since both of them testify inside a courtroom for the same cases. On the other hand, the differences in the speech of defendants compared to the speech of witnesses are due to their role in the judicial process since defendants' speech tends to be apologetic and said in the first person.

One of the dissertation's achievements is the development of a computational tool which can give support in testimony evaluation before the criminal court decides whether a defendant is guilty or not, by classifying a defendants' testimony according to its linguistic characteristics in the category of 'guilty' or 'not guilty. The implementation of our text classifier that has the ability to predict whether a testimony has been said by a guilty defendant or not, is based on the separation of testimonies into guilty and not guilty, including witnesses to the 'not guilty' category. Due to the nature of the cases' accusation, the percentage of innocent people charged with murder in relation to the guilty ones is very low. Given these limitations, the only realistic approach was to study defendants against witnesses. The assumption that witnesses were considered innocent defendants yielded surprisingly good results when the corresponding lexical data were used as inputs to our classification model. One might assume that the good performance of our classification model lies in this assumption. Nevertheless, of the 12 stylistic features used to train our model, none of them have clear differences in value between defendants and witnesses. Therefore, we suppose that the accuracy of almost 93% of our classification model is due to the correct combination of the training data that were used.

The big question that arises from our experiments is whether we have really managed to distinguish guilty from innocents or defendants from witnesses. To avoid this problem, we performed measurements where the training set was consisted only of guilty and witnesses, while the test set consisted of guilty and innocents in equal proportions. Since

the innocents are about 10% of all the defendants, the test set should include all the innocents and another 10% of guilty defendants. Such an experiment should show whether the classifier is really capable of distinguishing the innocents from the guilty defendants despite being trained to use witnesses instead of innocents. However, this experiment faces two main limitations that have to be confronted. Firstly, the size of the test set compared to the training set is small, which affects the reliability of the model's accuracy. Thus, the evaluation measurements were performed several times. Secondly, the fact that the set of innocents should be kept constant, affects the randomness of the algorithm. The only solution was to keep the set of innocents constant and change only the set of the guilty defendants. Inevitably, the limitation of having only 13 innocent defendants and the inability of testing our model with several sets of innocents, gives unreliable results. Apparently, the average accuracy of the guilty / innocent classification tends to be lower than the initial evaluation measurements. Indicatively, we mention that after seven several measurements where the set of innocents was kept constant and changing only the set of the guilty defendants, the accuracy of our model was ranging from 0.5 to 0.6, i.e., 50% to 60% of accuracy. Therefore, based on the available data, it is not possible to reliably identify a guilty from an innocent defendant since the use of witnesses in the training set results in focusing more on the differences between guilty defendants and witnesses than on those between guilty and innocents.

Lastly, we should mention that the role of the human factor, i.e., the judgment of the judge / interrogator / investigator, in the judicial process is irreplaceable and by no means will it be replaced by a machine learning tool. Furthermore, it is worth noting the limitations of any classification model that need to be considered in terms of neutralizing bias before engaging in standard decision making processes. There are several examples of machine learning failures and algorithmic bias, such as in facial recognition where, apart from inevitable concerns about privacy, ethics and human rights, there are issues of accuracy as well (Fussey & Murray, 2019). For instance, it was found that leading facial-recognition software packages performed much worse at identifying the gender of women and people of color than at classifying male, white faces (Buolamwini & Gebru, 2018). A case of a machine learning failure is that of a wrongfully accused by a facial recognition algorithm (Hill, 2020). Another interesting research regarding algorithms for

predicting recidivism showed that the widely used commercial risk assessment software COMPAS, which can assess a criminal defendant's likelihood of committing a crime, is no more accurate or fair than predictions made by people with little or no criminal justice expertise (Dressel & Farid, 2018). Machine learning specialists of Amazon uncovered a problem regarding an AI recruiting tool that showed bias against women and forced to remove the tool (Dastin, 2008). An organization, called the Algorithmic Justice League (https://www.ajl.org/), was created having as a mission 'to raise public awareness about the impacts of AI, equip advocates with empirical research, build the voice and choice of the most impacted communities, and galvanize researchers, policy makers, and industry practitioners to mitigate AI harms and biases'.

Therefore, researchers should be very skeptical about introducing tools for decision making processes into the market, especially in the case of evaluating a defendants' testimony. Thus, we should emphasize that GCDT classifier offers a stepping stone in the creation of a supporting computational tool in the process of evaluating the testimonies of a defendant and it is not intended to replace the judgment of a human.

7 CONCLUSIONS

In this dissertation we created a Greek corpus from testimonies of defendants for murder inside the courtroom and we studied their linguistic profile. Using tools and practices from Natural Language Processing and Machine Learning we quantified the defendants' speech inside the courtroom by measuring several linguistic features of their testimonies and we concluded that demographic characteristics, such as age and nationality, play a decisive role in the way they testify. Compared to the general Greek language, we detected several differences confirming that the stylometric features that defendants use differ from those of the general language. Moreover, defendants' speech differs depending on whether they testify inside a courtroom or during their interrogation. Comparing defendants' stylometric features with those of witnesses, we denoted several similarities but also several differences. Finally, having a sufficient number of testimonies, both defendants and witnesses, we developed a machine learning text classifier, and we examined the accuracy of predicting whether a written testimony belongs to a murderer or not. Our classifier, based on testimonies of defendants and witnesses, can characterize a person who testifies, as guilty or not, with 93% accuracy.

7.1 Contribution

This research embarked on a challenging task of investigating testimonies arising in specific legal context of Greek courtrooms' examination. To our knowledge, this is the first study in Greek language that reported the use of the language of testimony at such a high level as the Court. Unlike previous studies, which focused mainly on the discourse analysis of legal texts, this research brought to the fore the interactive performance of defendants and witnesses inside the Greek courtroom.

Our first contribution concerns the construction of a corpus which contained the testimonies of defendants accused of murder gathered in real life conditions, i.e., inside a courtroom in front of a judge. In tandem with that corpus, we constructed a second one which contained testimonies from witnesses of the same trial cases, and a third one involving the testimonies of the defendants before their trial, in front of an interrogator.

All three corpora are constructed by scratch using trial briefs from a Greek court and introduce a good source for further research in the field of Forensic Linguistics.

Subsequently, after testing the use of the most popular stylometric features in our dataset, borrowed from the research field of author identification, we managed to construct the linguistic profile of the defendants accused of murder who testify inside a courtroom. We demonstrated that this approach was feasible and useful since it proved that defendants use specific linguistic patterns in their testimonies, as we initially assumed. This conclusion was reached after a quantitative analysis of the stylometric characteristics of the speech of the defendants compared to the general language, to a language with similar stylometric characteristics, such as the language that the witnesses use inside a court, and to the pre-interrogations of the same defendants before their trial.

Finally, we applied text classification techniques in our dataset, borrowed from the field of machine learning and text mining, and we showed that a prediction model can be implemented which can predict whether a text has been written or said by a murderer, or not, with 93% accuracy. Regarding the prediction of the category of a testimony, the results of our model demonstrate that stylometric techniques, such as those previously used for author identification, can be used for training a classification model and can be effective even when the communication takes place in natural environments, attempting to classify oral speech. Our model achieves high accuracy and precision at identifying both testimonies of guilty and not guilty persons correctly. However, we should mention that the true achievement of this study was the development of a tool which can give support in testimony evaluation without replacing any of the judicial procedures.

7.2 Future work

The model presented in this dissertation can achieve almost 93% in accuracy and correspondingly high values in precision, recall, F1-score and AUC, i.e., in the main classification evaluation measures, on our dataset. Although this study is the first attempt in Greece that deals with the analysis of real-life testimonies of defendants inside a court, these values represent a remarkable performance to predict the category of a testimony. However, the model's performance rate also shows that there is further room for improvement. The current work uses mainly low-level stylometric features (lexical and

syntactic). This ensures that they are essentially language-independent and efficient. In such cases, an interesting future work direction could focus on a richer feature set, comprising low-level and high-level (semantic and structural) features, that can be adapted in every testimony separately. Moreover, as it is already mentioned in this thesis, it is essential that a balanced dataset, i.e., with an equal amount of data at each classification category, should be introduced.

Given that the GCDT classifier provides remarkable results when specific stylometric features from defendants' testimonies are considered as training set, it could be interesting to further enrich the pool of legal text classification classifiers considering several versions of the same approach with different fixed features settings. Thus, another future work dimension could be to explore the linguistic profile of another group of defendants, i.e., accused of rape or terrorism, in order to expand our existing model. This indicates the evaluation of the effects when a change occurs in the genre of the dataset and the confirmation or not of our model functions.

Finally, an interesting future work involves a research study regarding the implementation of such predictions models in real life environments that concern judicial procedures and the effects of their application. This includes the investigation of whether a model prediction can be used to facilitate the judicial or investigative process and whether safe conclusions can be drawn. Such research might go beyond the limits of our research field, since it is likely to fall within the remit of sociologists and behaviorists as it should be investigated to what extent a learning machine can gain the trust of the judiciary in order to be applied to the legal procedures.

BIBLIOGRAPHY

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group Web forum messages. In *IEEE Intelligent Systems*, *20(*5), 67-75. https://doi.org/10.1109/MIS.2005.81

Abbasi, A., Chen, H., & Nunamaker, J. F. (2008). Stylometric identification in electronic markets: Scalability and robustness. *Journal of Management Information Systems*, *25*(1), 49-78. https://doi.org/10.2753/MIS0742-1222250103

Adelsbach, A., & Sadeghi, A.-R. (2003). Advanced techniques for dispute resolving and authorship proofs on digital works. In *Security and Watermarking of Multimedia Contents V*, (Vol. 5020, pp. 677-688). SPIE. https://doi.org/10.1117/12.477295

Aggarwal, C. C. (2015). Mining text data. In *Data mining* (pp. 429-455). Springer, Cham.

Ainsworth, P. B. (2001). *Offender profiling and crime analysis*. Devon: Willan.

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European court of human rights: A natural language processing perspective. *PeerJ Computer Science 2,* e93. https://doi.org/10.7717/peerj-cs.93

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., D., E., B., J., & Kochut, K. (2017). Text Summarization Techniques: A Brief Survey. *International Journal of Advanced Computer Science and Applications*, *8*(10), 397-405. https://doi.org/10.14569/ijacsa.2017.081052

Alruily, M., Ayesh, A., & Zedan, H. (2014). Crime profiling for the Arabic language using computational linguistic techniques. *Information Processing and Management*, *50*(2), 315-341. https://doi.org/10.1016/j.ipm.2013.09.001

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012). The 'K' in K-fold cross validation. In *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)* (pp. 441-446). i6doc. com publ.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52*(2), 119-123. https://doi.org/10.1145/1461928.1461959

Argamon, S., & Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing* (pp. 1-3).

Ariani, M. G., Sajedi, F., & Sajedi, M. (2014). Forensic Linguistics: A Brief Overview of the Key Elements. *Procedia - Social and Behavioral Sciences*, *158,* 222-225. https://doi.org/10.1016/j.sbspro.2014.12.078

Baker, P., Hardie, A., & McEnery, T. (2006). *Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Balakrishnama, S., Ganapathiraju, A., & Picone, J. (1999). Linear discriminant analysis for signal processing problems. In *Proceedings IEEE Southeastcon'99. Technology on the Brink of 2000 (Cat. No. 99CH36300)* (pp. 78-81). IEEE. https://doi.org/10.1109/SECON.1999.766096

Beil, F., Ester, M., & Xu, X. (2002). Frequent term-based text clustering. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 436-442). https://doi.org/10.1145/775047.775110

Bennell, C., Jones, N. J., & Taylor, A. (2011). Determining the authenticity of suicide notes: Can training improve human judgment? *Criminal Justice and Behavior*, *38*(7), 669-689. https://doi.org/10.1177/0093854811405146

Berber-Sardinha, T. (2000). Comparing corpora with WordSmith Tools: How large must the reference corpus be? *WCC '00 Proceedings of the Workshop on Comparing Corpora*, *9,* 7-13. https://doi.org/10.3115/1117729.1117731

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(10), 281-305. https://dl.acm.org/doi/10.5555/2188385.2188395

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1998). When is "nearest neighbor" meaningful? In: *Beeri C., Buneman P. (eds) Database Theory — ICDT'99. ICDT 1999. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg,1540,* 217-235. https://doi.org/10.1007/3-540-49257-7_15

Bhatia, V. K., Flowerdew, J., & Jones, R. H. (2007). *Advances in discourse studies (1st ed.)*. Routledge. https://doi.org/10.4324/9780203892299

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., & Lang, M. (2017). mlrMBO: A modular framework for model-based optimization of expensive black-box functions. *ArXiv preprint arXiv:1703.03373*.

Boella, G., Di Caro, L., & Humphreys, L. (2011). Using classification to support legal knowledge engineers in the Eunomos legal document management system. In *Fifth International Workshop on Juris-Informatics (JURISIN)*.

Boyer, M., & Lapalme, G. (1985). Generating paraphrases from meaning-text semantic networks. *Computational Intelligence*, *1*(1), 103-117. https://doi.org/10.1111/j.1467-8640.1985.tb00063.x

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. In *ACM Computing Surveys*, *49*(2), 1-50. https://doi.org/10.1145/2907070

Broussalis, G., Markopoulos, G., & Mikros, G. (2012). Stylometric profiling of the Greek Legal Corpus. *Selected Papers of the 10th International Conference of Greek Linguistics, 10,* 167-176.

Bucholtz, M. (2007). Variation in transcription. *Discourse Studies*, *9*(6), 784-808. https://doi.org/10.1177/1461445607082580

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research, 81,* 77-91.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1301-1309.

Burrows, J. F. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, *2*(2), 61-70. https://doi.org/10.1093/llc/2.2.61

Burrows, J. F. (1992). Not unles you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, *7*(2), 91-109. https://doi.org/10.1093/llc/7.2.91

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *ACM International Conference Proceeding Series*, *148*, 161-168. https://doi.org/10.1145/1143844.1143865

Chambers, J. K., & Trudgill, P. (1998). *Dialectology*. Cambridge University Press. https://doi.org/10.1017/cbo9780511805103

Chaski, C. E., & D, P. (2005). Who' s At The Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal*, *4*(1), 1-13.

Chifflet, P. (2015). Questioning the validity of criminal profiling: an evidence-based approach. *Australian and New Zealand Journal of Criminology*, *48*(2), 238-255. https://doi.org/10.1177/0004865814530732

Choudhary, S. K. (2018). Significance of Forensic Stylistics in Fixing Authorship of Handwriting. *Journal of Forensic Sciences & Criminal Investigation*, *7*(4), 69-87. https://doi.org/10.19080/jfsci.2018.07.555718

Copi, I., Cohen, C., & Flage, D. (2016). *Essentials of Logic*. Routledge. https://doi.org/10.4324/9781315389028

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1023/A:1022627411411

Cotterill, J. (2002). 'Just One More Time …': Aspects of Intertextuality in the Trials of O. J. Simpson. In: *Language in the Legal Process.* (pp. 147-161) Palgrave Macmillan, London. https://doi.org/10.1057/9780230522770_9

Cotterill, J. (2003). *Language and power in court: A linguistic analysis of the O.J. Simpson trial.* Springer. https://doi.org/10.1057/9780230006010

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, *25*(4), 431-447. https://doi.org/10.1093/applin/25.4.431

Coyotl-Morales, R. M., Villaseñor-Pineda, L., Montes-y-Gómez, M., & Rosso, P. (2006). Authorship attribution using word sequences. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4225,* 844-853. https://doi.org/10.1007/11892755_87

Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.com*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davis, J. A. (1996). *Crime scene investigative analysis: Elements of profiling*. San Diego, CA: Author.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *ACM International Conference Proceeding Series*, *148*, 233-240. https://doi.org/10.1145/1143844.1143874

Deepu, S., Pethuru, R., & Rajaraajeswari, S. (2016). A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction. In *Proceedings of the 1st International Conference on Innovations in Computing & Networking (ICICN16), Bangalore, India* (pp. 12-13).

Dickerson, R. (1965). *The fundamentals of legal drafting*. Toronto: Little, Brown and Company.

Douglas, J. E., Ressler, R. K., Burgess, A. W., & Hartman, C. R. (1986). Criminal profiling from crime scene analysis. *Behavioral Sciences & the Law*, *4*(4), 401-421. https://doi.org/10.1002/bsl.2370040405

Dowden, C., Bennell, C., & Bloomfield, S. (2007). Advances in Offender Profiling: A Systematic Review of the Profiling Literature Published Over the Past Three Decades. In *Journal of Police and Criminal Psychology*, 22(1), 44-56. https://doi.org/10.1007/s11896-007-9000-9

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Dumas, B. K. (2007). Review of *Forensic Linguistics: An Introduction to Language in the Justice System* by John Gibbons. *International Journal of Speech Language and the Law*, *11*(1). https://doi.org/10.1558/ijsll.v11i1.146

El-Waned, S. A., Elfatatry, A., & Abougabal, M. S. (2007). A new look at software plagiarism investigation and copyright infringement. In *2007 ITI 5th International Conference on Information and Communications Technology* (pp. 315-318). IEEE. https://doi.org/10.1109/ITICT.2007.4475669

Farzindar, A., & Lapalme, G. (2004). Legal Text Summarization by Exploration of the Thematic Structure and Argumentative Roles. *In Text Summarization Branches Out Conference Held in Conjunction with ACL 2004*, 27-34.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fisher, G., Bugliosi, V., Cochran, J. L., Rutten, T., Cooley, A., Bess, C., Rubin-Jackson, M., Byrnes, T., Darden, C. A., Walter, J., Dershowitz, A. M., Schiller, L., Willwerth, J., Shapiro, R. L., Warren, L., & Toobin, J. (1997). The O. J. Simpson Corpus. *Stanford Law Review*, *49*(4), 971-1019. https://doi.org/10.2307/1229341

Fitzpatrick, E., & Bachenko, J. (2009). *Building a forensic corpus to test language-based indicators of deception*. Language and Computers, 71(1), 183–196.

Fitzpatrick, E., & Bachenko, J. (2012). Building a data collection for deception research. *Proceedings of the Workshop on Computational Approaches to Deception Detection*, 31-38.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Categorization. *Journal of Machine Learning Research, 3*, 1289-1305.

Fornaciari, T., & Poesio, M. (2013). Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*, *21*(3), 303-340. https://doi.org/10.1007/s10506-013-9140-4

Frantzeskou, G., Stamatatos, E., Gritzalis, S., & Katsikas, S. (2006). Effective identification of source code authors using byte-level information. *Proceedings of the 28th International Conference on Software Engineering*, *2006*, 893-896. https://doi.org/10.1145/1134285.1134445

Frantzi, K. T. (2005). *Γλωσσικά και μη χαρακτηριστικά των προκηρύξεων της 17Ν* [Linguistic and non-linguistic characteristics of the announcements of 17N]. *Μελέτες για την ελληνική γλώσσα, 25*, 639-650.

Frantzi, K. T. (2007). Extracting stylistic distances from texts for forensic linguistics purposes. *Proceedings of Digital Humanities 2007. Association for Digital Humanities, 4-8 June 2007, University of Illinois, Urbana-Champaign*, 67-69.

Frantzi, K. T. (2009). *Δικαστική Γλωσσολογία: Αυτόματη Εξαγωγή Υφολογικών Γλωσσικών Χαρακτηριστικών* [Forensic Linguistics: Automatic extraction of linguistic-stylistic features]. *Selected papers on theoretical and applied linguistics, 18*, 505-512.

Frantzi, K. T., & Katranidou, A. K. (2017). A Corpus-Based Analysis of the Language Used by Defendants of Homicide in Court. *World Journal of Social Science Research*, *4*(2), 164-174. https://doi.org/10.22158/wjssr.v4n2p164

Fraser, H. (2003). Issues in transcription: Factors affecting the reliability of transcripts as evidence in legal cases. *Speech, Language and the Law*, *10*(2), 203-226. https://doi.org/10.1558/sll.2003.10.2.203

Fraser, H. (2014). Transcription of indistinct forensic recordings: Problems and solutions from the perspective of phonetic science. *Language and Law/Linguagem e Direito, 1*(2), 5-21.

French, P., & Stevens, L. (2013). *Forensic speech science*. In R. Knight and M. Jones, Eds., Bloomsbury Companion to Phonetics. London: Continuum.

Fussey, P., & Murray, D. (2019). *Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology*. Univ. Essex.

Galatolo, R. (2006). Active voicing in court. In E. Holt & R. Clift (Eds.), *Reporting Talk: Reported Speech in Interaction* (Studies in Interactional Sociolinguistics, 195-220), Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486654.009

Galatolo, R., & Mizzau, V. (2005). Quoting dialogues and the construction of the narrative point of view in legal testimony: The role of prosody and gestures. *Studies in Communication Sciences 5*, 217–232.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics (COLING '04*). Association for Computational Linguistics, USA, 611–es. https://doi.org/10.3115/1220355.1220443

García, A. M., & Martín, J. C. (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing*, *22*(1), 49-61. https://doi.org/10.1093/llc/fql048

Gerolympou, K. (1999). *Preliminary study of the use of intimidation in a criminal court setting*. Master Thesis for the Department of Forensic Psychology. Manchester: Manchester Metropolitan University.

Gibbons, J. (1999). Language and the law. *Annual Review of Applied Linguistics*, *19*, 156-173.

Gonçalves, T., & Quaresma, P. (2005). Evaluating preprocessing techniques in a text classification problem. *São Leopoldo, RS, Brasil: SBC-Sociedade Brasileira de Computação.*

Goodrich, P. (1987). *Legal discourse: Studies in linguistics, rhetoric and legal analysis*. London: Macmillan.

Goutsos, D. (2003). Corpus of Greek Texts: design and implementation. *In Proceedings of the sixth International Conference on Greek Linguistics*, University of Crete, 18-21 September 2003. CD-ROM publication. (In Greek.)

Goutsos, D. (2010). The Corpus of Greek Texts: a reference corpus for Modern Greek. *Corpora*, *5*(1), 29-44. https://doi.org/10.3366/cor.2010.0002

Grant, T., & Baker, K. (2001). Identifying reliable, valid markers of authorship: A response to Chaski. *Speech, Language and the Law*, *8*(1), 66-79. https://doi.org/10.1558/sll.2001.8.1.66

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

Gries, S. T. (2016). *Quantitative Corpus Linguistics with R: A Practical Introduction* (2nd ed.). Routledge. https://doi.org/10.4324/9781315746210

Guyon, I., Saffari, A., Dror, G., & Cawley, G. (2010). Model selection: Beyond the bayesian/frequentist divide. *Journal of Machine Learning Research*, *11*(3), 61-87.

Hachey, B., & Grover, C. (2006). Extractive summarisation of legal texts. *Artificial Intelligence and Law*, *14*(4), 305-345. https://doi.org/10.1007/s10506-007-9039-z

Hammond, M. (2007). *Programming for Linguists: Perl for Language Researchers*. John Wiley & Sons. https://doi.org/10.1002/9780470752234

Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of Data Mining (Adaptive Computation and Machine Learning). In *Journal of the American Statistical Association*, *98*(461).

Harris, S. (2001). Fragmented Narratives and Multiple Tellers: Witness and Defendant Accounts in Trials. *Discourse Studies*, *3*(1), 53-74. https://doi.org/10.1177/1461445601003001003

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari, E., Papageorgiou, H., & Demiros, I. (2000). Design and implementation of the online ILSP Greek corpus. *2nd International Conference on Language Resources and Evaluation, LREC 2000*.

Heffer, C. (2005). *The Language of Jury Trial: A Corpus-Aided Analysis of Legal-Lay Discourse*. Springer. https://doi.org/10.1057/9780230502888

Heselwood, B. (2014). *Phonetic Transcription in Theory and Practice*. Edinburgh University Press. https://doi.org/10.3366/edinburgh/9780748640737.001.0001

Hill, K. (2020, June 24). Wrongfully accused by an algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Hollien, H. (2012). About forensic phonetics. *Linguistica*, *52*(1), 27-53. https://doi.org/10.4312/linguistica.52.1.27-53

Holmes, D. I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, *13*(3), 111-117. https://doi.org/10.1093/llc/13.3.111

Holt, E., & Clift, R. (2006). *Reporting Talk: Reported Speech in Interaction* (Vol. 24). Cambridge University Press. https://doi.org/10.1017/CBO9780511486654

Hoover, D. L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, *37*(2), 151-178. https://doi.org/10.1023/A:1022673822140

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression: Third Edition* (Vol. 398). John Wiley & Sons. https://doi.org/10.1002/9781118548387

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, *20*(1), 19-62.

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. In *International conference on artificial intelligence: Methodology, systems, and applications* (pp. 77-86). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11861461_10

Hovland, D.L. (1993). Errors in Interpretation: Why Plain Error is Not Plain, *11(2) Law & Ineq.,* 473. https://scholarship.law.umn.edu/lawineq/vol11/iss2/6/

Iegorova, A. (2018). *Examining Forensic Testimonies as a genre of narrative discourse*. (Unpublished doctoral dissertation). Aristotle University of Thessaloniki, Faculty of Philosophy, School of Italian Language and Literature.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, *4*(8), 966-974.

Jackson, P., Al-Kofahi, K., Tyrrell, A., & Vachher, A. (2003). Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, *150*(1-2), 239-290. https://doi.org/10.1016/S0004-3702(03)00106-1

Jensen, K. E. (2014). Linguistics in the digital humanities: (computational) corpus linguistics. *MedieKultur: Journal of Media and Communication Research*, *30*(57), 20-p. https://doi.org/10.7146/mediekultur.v30i57.15968

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In: *Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998*. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1398, 137-142, Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0026683

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers in Linguistics*, *53*(0), 61-79.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *In Science*, 349(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Juola, P. (2008). Authorship Attribution. *In: Foundations and Trends in Information Retrieval 1*, 234–334.

Juszczak, P., Tax, D. M. J., & Duin, R. P. W. (2002). Feature scaling in support vector data description. In *Proc. ASCI* (pp. 95-102). Citeseer.

K. Sai Prasad, Dr. S. P. (2020). A Theoretical Review on Data Mining and Machine Learning Techniques for Data Analysis. *International Journal of Advanced Science and Technology, 29*(11s), 1220-1226.

Kapopoulos, C.F. (2021). *Νόμος και γλώσσα στη στρατιωτική δικαιοσύνη: μελέτη της γλώσσας στα στρατιωτικά δικαστήρια* [Law and language in military justice: a study of language in military courts]. (Διδακτορική Διατριβή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Σχολή Φιλοσοφική, Τμήμα Φιλολογίας. Τομέας Γλωσσολογίας). Retrieved from National Archive of PhD Theses, http://hdl.handle.net/10442/hedi/50540

Katranidou, A., & Frantzi, K. T. (2016). The Greek Corpus of Defendants' Testimonies: frequent use of infrequent words. *European Journal of Humanities and Social*, *3*, 25-29. https://doi.org/10.20534/ejhss-16-3-25-29

Katsampos S. (2015). *Εφαρμογή Διαχείρισης Πληροφορίας σε Πηγές Νομικών Κειμένων* [Information management application in legal text sources]. (Μη δημοσιευμένη

μεταπτυχιακή διπλωματική εργασία). Πανεπιστήμιο Θεσσαλίας, Σχολή Θετικών Επιστημών. Retrieved from https://ir.lib.uth.gr/

Kennedy, G. (2014). *An Introduction to Corpus Linguistics*. Routledge. https://doi.org/10.4324/9781315843674

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *Proceedings of 2014 Science and Information Conference, SAI 2014,* 372-378. https://doi.org/10.1109/SAI.2014.6918213

Kim, S. B., Rim, H. C., Yook, D. S., & Lim, H. S. (2002). Effective Methods for Improving Naive Bayes Text Classifiers. In: *Ishizuka M., Sattar A. (eds) PRICAI 2002: Trends in Artificial Intelligence. PRICAI 2002. Lecture Notes in Computer Science, vol 2417*, 414-423, Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45683-X_45

Kirk, J. (1996). English Corpus Linguistics: Studies in Honour of Jan Svartvik. Edited by Karin Aijmer and Bengt Altenberg. London: Longman, 1991. *Journal of English Linguistics, 24*(3), 250-258. https://doi.org/10.1177/007542429602400308

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, *17*(4), 401-412. https://doi.org/10.1093/llc/17.4.401

Kornai, A. (2008). *Mathematical Linguistics*, Springer.

Kotsalis L, & Margaritis M. P. (2007). *Δικαστική ψυχολογία* [Forensic psychology]. Αθήνα: Σάκκουλας.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160*(1), 3-24. https://doi.org/10.31449/inf.v31i3.148

Koutsogoula, E. (2014). *Ανάλυση δικαστικών αποφάσεων για αξιώσεις κατασκευαστικών έργων – Εφαρμογές Τεχνικής Νοημοσύνης* [Analysis of court decisions on construction claims – Applications of Artificial Intelligence]. (Μη δημοσιευμένη μεταπτυχιακή διπλωματική εργασία). Τομέας Προγραμματισμού και Διαχείρισης Τεχνικών Έργων, Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Πολιτικών Μηχανικών. Retrieved from http://dspace.lib.ntua.gr

Lang, M., Bischl, B., & Surmann, D. (2017). batchtools: Tools for R to work on batch systems. *The Journal of Open Source Software*, *2*(10), 135. https://doi.org/10.21105/joss.00135

Lardilleux, A., & Lepage, Y. (2009). Hapax legomena: Their contribution in number and efficiency to word alignment. In: *Vetulani Z., Uszkoreit H. (eds) Human Language Technology. Challenges of the Information Society. LTC 2007*. Lecture Notes in Computer Science, vol 5603, 440-450, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04235-5_38

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining* (Vol. 4). John Wiley & Sons. https://doi.org/10.1002/0471687545

Lee, J. (2010). Interpreting reported speech in witnesses' evidence. *Interpreting. International Journal of Research and Practice in Interpreting*, *12*(1), 60-82. https://doi.org/10.1075/intp.12.1.03lee

Leech, G. (2014). *The state of the art in corpus linguistics* (pp. 20-41). Routledge. https://doi.org/10.4324/9781315845890

Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. How to represent texts in input space? *Machine Learning*, *46*(1–3), 423-444. https://doi.org/10.1023/A:1012491419635

Lidsky, L. B. (2000). Silencing John Doe: Defamation & discourse in cyberspace. *Duke Law Journal*, *49*(4), 855. https://doi.org/10.2307/1373038

Lim, H. S. (2004). Improving kNN based text classification with well estimated parameters. In: *Pal N.R., Kasabov N., Mudi R.K., Pal S., Parui S.K. (eds) Neural Information Processing. ICONIP 2004*. Lecture Notes in Computer Science, vol 3316, 516-523, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30499-9_79

Linder, D. (2011). The Oklahoma City Bombing and the Trial of Timothy McVeigh. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1030565

Linelu, P., Gustavsson, L., & Juvonen, P. (1988). Interactional dominance in dyadic communication: A presentation of initiative-response analysis. *Linguistics*, *26*(3), 415-442. https://doi.org/10.1515/ling.1988.26.3.415

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14-23. https://doi.org/10.1002/widm.8

M, H., & M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1. https://doi.org/10.5121/ijdkp.2015.5201

Madigan, D., Genkin, A., Lewis, D., Argamon, S.E., Fradkin, D., & Ye, L. (2005). Author Identification on the Large Scale. In *Proceedings of the Meeting of the Classification Society of North America*.

Madsen, R. E., Sigurdsson, S., Hansen, L. K., & Larsen, J. (2004). Pruning the vocabulary for better context recognition. *IEEE International Conference on Neural Networks - Conference Proceedings*, *2,* 1439-1444. https://doi.org/10.1109/IJCNN.2004.1380163

Maley, Y. (1987). The language of legislation. *Language in Society*, *16*(1), 25-48. https://doi.org/10.1017/S0047404500012112

Maley, Y. (1994). The language of the law. In *John Gibbons (ed.), Language and the Law. Longman.* 11-50.

Matoesian, G. (2005). Nailing down an answer: Participations of power in trial talk. In *Discourse Studies, 7*(6), 733-759. https://doi.org/10.1177/1461445605055424

McEnery, T. (2019). *Corpus Linguistics*. Edinburgh University Press.

McEnery, T., & Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

McEnery, T., & Wilson, A. (2003). Corpus Linguistics. *The Oxford handbook of computational linguistics*, 448-463.

McMenamin, G. R. (1993). *Forensic stylistics*. Amsterdam, Elsevier.

McMenamin, G. R. (2002). Forensic Linguistics Advances in Forensic Stylistics. *CRC Press LLC 2002*, *1*.

Mikros, G. K. (2013). Authorship Attribution and Gender Identification in Greek Blogs. *Methods and Applications of Quantitative Linguistics*, *21*, 21-32

Mikros, G. K., & Argiri, E. K. (2007). Investigating topic influence in authorship attribution. *CEUR Workshop Proceedings*, *276*.

Miranda, G. A., & Calle, M. J. (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing, 22*(1), 49-66.

Moens, M. F., Uyttendaele, C., & Dumortier, J. (1999). Information extraction from legal texts: The potential of discourse analysis. *International Journal of Human Computer Studies*, *51*(6), 1155-1171. https://doi.org/10.1006/ijhc.1999.0296

Mokros, A., & Alison, L. J. (2002). Is offender profiling possible? Testing the predicted homology of crime scene actions and background characteristics in a sample of rapists. *Legal and Criminological Psychology*, *7*(1), 25-43. https://doi.org/10.1348/135532502168360

Montañés, E., Quevedo, J. R., & Díaz, I. (2003). A wrapper approach with Support Vector Machines for Text Categorization. In: *Mira J., Álvarez J.R. (eds) Computational Methods in Neural Modeling. IWANN 2003*. Lecture Notes in Computer Science, vol 2686, 230-237, Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44868-3_30

Muller, D. A. (2000). Criminal Profiling: Real Science or Just Wishful Thinking? *Homicide Studies*, *4*(3), 234-264. https://doi.org/10.1177/1088767900004003003

Niforas, N. (2016). *Χρήση τεχνικών εξόρυξης από κείμενο (text mining) στην ταξινόμηση νομοθετικών διατάξεων* [Use of text mining techniques in the classification of legal provisions]. (Μη δημοσιευμένη μεταπτυχιακή διπλωματική εργασία). Πανεπιστήμιο Πατρών, Τμήμα Διοίκησης Επιχειρήσεων. Retrieved from http://hdl.handle.net/10889/9676

Nini, A. (2013). Codal variation theory as a forensic tool. In Bridging the Gap(s) between Language and the Law: *Proceedings of the 3rd European Conference of the International Association of Forensic Linguists. Porto: Faculdade de Letras da Universidade do Porto*, 31-41.

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

Olsson, J. (2004). *Forensic Linguistics*: *An Introduction to Language, Crime and the Law*. Bloomsbury.

Olsson, J., & Luchjenbroers, J. (2013). *Forensic linguistics*. London: A&C Black.

Palau, R. M., & Moens, M. F. (2009). Argumentation mining: The detection, classification and structure of arguments in text. *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 98-107. https://doi.org/10.1145/1568234.1568246

Pavelec, D., Justino, E., & Oliveira, L. S. (2007). Author identification using stylometric features. *Inteligencia Artificial*, *11*(36), 59-65. https://doi.org/10.4114/ia.v11i36.892

Pavlidou, Th.-S. (2012). The Corpus of Spoken Greek: Goals, Challenges, Perspectives. *KOLREC Proceedings, Workshop 18 (Best Practices for Speech Corpora in Linguistic Research)*, 23-28.

Peng, F., Schuurmans, D., & Wang, S. (2004). Augmenting naive Bayes classifiers with statistical language models. *Information Retrieval*, *7*(3), 317-345. https://doi.org/10.1023/b:inrt.0000011209.19643.e2

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., & Burzo, M. (2015). Deception detection using real-life trial data. *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, 59-66. https://doi.org/10.1145/2818346.2820758

Pérez-Rosas, V., & Mihalcea, R. (2015). Experiments in open domain deception detection. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 112-1125. https://doi.org/10.18653/v1/d15-1133

Pitsogiannis P.I. (1983). *Αναχριτική* [Investigation]. Θεσσαλονίκη: Σάκκουλας

Precht, K., Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, *20*(1), 1934-1965.

Rangel, F., Rosso, P., Montes-Y-Gómez, M., Potthast, M., & Stein, B. (2018). Overview of the 6th Author Profiling Task at PAN 2018: Multimodal gender identification in Twitter. *Working Notes Papers of the CLEF,* 1-38.

Rieber, R., & Stewart, W. (1990). *The Language Scientist as Expert in the Legal Setting (eds).* New York Academy of Sciences, 606 Number 6.

Rota M. P. (2014). *Η σκιαγράφηση του ψυχο-κοινωνικού προφίλ του εγκληματία στη σύγχρονη ανακριτική* [The sketching of the psycho-social profile of the criminal in the modern investigation]. (Διδακτορική Διατριβή, Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών. Τμήμα Κοινωνιολογίας. Τομέας Εγκληματολογίας, Αθήνα, Ελλάδα). Retrieved from National Archive of PhD Theses, http://hdl.handle.net/10442/hedi/35227

Russel, S., & Norvig, P. (2012). Artificial intelligence—a modern approach 3rd Edition. In *The Knowledge Engineering Review*, *11*(1), 78-79. https://doi.org/10.1017/S0269888900007724

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3): e0118432. https://doi.org/10.1371/journal.pone.0118432

Sanderson, C., & Guenter, S. (2006). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. *COLING/ACL 2006 - EMNLP 2006: 2006 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 482-491).

Santaniello, L. (2018). If an Interpreter Mistranslates in a Courtroom and There is No Recording, Does Anyone Care?: The Case for Protecting LEP Defendants' Constitutional Rights, *Nw. JL & Soc. Pol'y*, *14*, 91.

Schneider, K. M. (2005). Techniques for improving the performance of naive bayes for text classification. In: *Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2005*. Lecture Notes in Computer Science, vol 3406, 682-693, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30586-6_76

Scott, M. (1998). *WordSmith Tools Version 3*. Oxford: Oxford University Press.

Scott, M. (2001). *Comparing corpora and identifying keywords, 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, 47-67.

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

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. In *ACM Computing Surveys, 34*(1), 1-47. https://doi.org/10.1145/505282.505283

Shams, R., & Mercer, R. E. (2016). Supervised classification of spam emails with natural language stylometry. *Neural Computing and Applications*, *27*(8), 2315-2331. https://doi.org/10.1007/s00521-015-2069-7

Shapero, J. (2011). *The language of suicide notes*. University of Birmingham. Ph.D.

Shuy, R. W. (1993). *Language crimes: the use and abuse of language evidence in the courtroom* (p. 24). Cambridge, MA: Blackwell.

Shuy, R. W. (2008). Discourse Analysis in the Legal Context. In *The Handbook of Discourse Analysis,* 437-452. https://doi.org/10.1002/9780470753460.ch23

Simpson, P. (2004). *Stylistics* (1st edn). Routledge, London.

Smith, G. P. (1966). [Review of *The Fundamentals of Legal Drafting*, by R. Dickerson]. *Michigan Law Review, 64*(4), 767–770. https://doi.org/10.2307/1287027

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Solan, L. M. (2013). *The Language of Judges*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226767895.001.0001

Soucy, P., & Mineau, G. W. (2003). Feature selection strategies for text categorization. In *Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 505-509). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44886-1_41

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538-556. https://doi.org/10.1002/asi.21001

Stamatatos, E., Kokkinakis, G., & Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26*(4), 471-495. https://doi.org/10.1162/089120100750105920

Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M. (2016). Clustering by authorship within and across documents. In *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al.* (pp. 691-715).

Stammatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., & Stein, B. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. *CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings*, *1391*(31).

Stygal, G. (1994). *Trial Language*. Philadelphia: John Benjamins.

Sulea, O. M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P., & Van Genabith, J. (2017). Exploring the use of text classification in the legal domain. *CEUR Workshop Proceedings*, *2143*. arXiv preprint arXiv:1710.09306

Svartvik, J. (2007). Corpus Linguistics 25+years on. In *Corpus Linguistics 25 years on* (pp. 9-25). Brill.

Tabachnick, B. G., L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5, pp. 481-498). Boston, MA: pearson.

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G., & Tambouratzis, D. (2004). Discriminating the Registers and Styles in the Modern Greek Language-Part 2: Extending the Feature Vector to Optimize Author Discrimination. *Literary and Linguistic Computing*, *19*(2), 221-242. https://doi.org/10.1093/llc/19.2.221

Tkačuková, T. (2010). Representing oneself: Cross-examination questioning: Lay people as cross-examiners. In *The Routledge Handbook of Forensic Linguistics. London and New York: Routledge*, 333-346.

To, V., Fan, S., & Thomas, D. (2013). Lexical density and readability: A case study of English textbooks. *Internet Journal of Language, Culture and Society*, *37*, 61-71.

Trial of the Six. (1976). Εκδόσεις Αθήναι.

Trosborg, A. (1997). *Rhetorical strategies in legal language: discourse analysis of statutes and contracts* (Vol. 424). Gunter Narr Verlag.

Tse, C. A. (1998). Is the Simultaneous Mode Feasible and Desirable in Court Interpreting? The Hong Kong Experience and Experiment. *Division of Language Studies City University of Hong Kong.*

Turell, M., Coulthard, M., & Johnson, A. (2008). *An Introduction to Forensic Linguistics: Language in Evidence*.

Ure, J. (1971). Lexical density and register differentiation. In *G. Perren and J.L.M. Trim (eds), Applications of Linguistics, London: Cambridge University Pres,* 443-452.

Vel, O. De. (2000). Mining e-mail authorship. In *Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000).*

Vrij, A., Winkel, F., & Akehurst, L. (1997). Police officers' incorrect beliefs about nonverbal indicators of deception and its consequences. In *Nijboer, J. and Reijntjes, J., editors, Proceedings of the first world conference on new trends in criminal investigation and evidence*, 221–238. Koninklijke Vermande, Lelystad, the Netherlands.

Voultepsis, I. (1975). *Οι δίκες της Χούντας – Πλήρη πρακτικά: Δίκη Πρωταιτίων 21ης Απριλίου 1967* [The trials of the Junta]. Δημοκρατικοί Καιροί.

Walker, A. G. (1986). The Verbatim Record: the myth and the reality. In *S. Fisher, A. D. Todd (Eds.*) Discourse and Institutional Authority, 205-222. Norwood, NJ: Erlbaum.

Wilson, A. (2019). A Brief Introduction to Supervised Learning. Retrieved from https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590

Witten, I. H., Frank, E., & Geller, J. (2002). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. *SIGMOD Record*, *31*(1), 76-77. https://doi.org/10.1145/507338.507355

Wyner, A., & Milward, D. (2008). Legal text-mining using linguamatics' I2E. *Presentation at Workshop on Natural Language Engineering of Legal Argumentation*, *Florence, Italy as part of JURIX 2008*.

Wyner, A., Mochales-Palau, R., Moens, M. F., & Milward, D. (2010). Approaches to text mining arguments from legal cases. *In: Francesconi E., Montemagni S., Peters W., Tiscornia D. (eds) Semantic Processing of Legal Texts*. Lecture Notes in Computer Science, vol 6036, 60-79, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12837-0_4

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, *44*(1), 48-59. https://doi.org/10.1177/0165551516677946

Yadav, S., & Shukla, S. (2016). Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. *Proceedings - 6th International Advanced Computing Conference, IACC 2016*, 78-83. https://doi.org/10.1109/IACC.2016.25

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95-112. https://doi.org/10.1016/j.jeconom.2015.02.006

Zhao, Y., & Zobel, J. (2005). Effective and scalable authorship attribution using function words. In: *Lee G.G., Yamada A., Meng H., Myaeng S.H. (eds) Information Retrieval Technology. AIRS 2005*. Lecture Notes in Computer Science, vol 3689,174-189, Springer, Berlin, Heidelberg. https://doi.org/10.1007/11562382_14https://doi.org/10.1007/11562382_14

Zheng, A., Shelby, N., & Volckhausen, E. (2019). Evaluating Machine Learning Models. *Machine Learning in the AWS Cloud.*

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification

techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), 378-393. https://doi.org/10.1002/asi.20316

APPENDIX

Metadata of defendants

Α/Α	ΓΕΝΟΣ	ΕΤΟΣ ΓΕΝΝΗΣΗΣ	ΕΘΝΙΚΟΤΗΤΑ	ΕΠΑΓΓΕΛΜΑ	ΤΟΠΟΣ ΓΕΝΝΗΣΗΣ	ΕΓΓΑΜΟΣ/ΑΓΑΜΟΣ	ΤΕΚΝΑ	ΕΤΥΜΗΓΟΡΙΑ	ΕΤΟΣ
1	ΑΝΔΡΑΣ	1993	ΕΛΛΗΝΙΚΗ	ΦΟΙΤΗΤΗΣ ΑΚΑΔΗΜΙΑΣ ΕΜΠΟΡΙΚΟΥ ΝΑΥΤΙΚΟΥ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2014
2	ΑΝΔΡΑΣ	1980	ΕΛΛΗΝΙΚΗ	ΑΝΕΡΓΟΣ	ΓΙΑΝΝΙΤΣΑ ΠΕΛΛΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2014
3	ΑΝΔΡΑΣ	1955	ΕΛΛΗΝΙΚΗ-ΓΕΩΡΓΙΑΝΗ	ΟΔΗΓΟΣ	ΣΟΧΟΥΜΙ ΓΕΩΡΓΙΑΣ	ΕΓΓΑΜΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2014
4	ΑΝΔΡΑΣ	1983	ΕΛΛΗΝΙΚΗ	ΕΜΠΟΡΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2014
5	ΑΝΔΡΑΣ	1965	ΒΟΥΛΓΑΡΙΚΗ	ΕΡΓΑΤΗΣ	ΣΟΦΙΑ ΒΟΥΛΓΑΡΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2014
6	ΑΝΔΡΑΣ	1991	ΕΛΛΗΝΙΚΗ	ΣΙΔΕΡΑΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΕΓΓΑΜΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2014
7	ΑΝΔΡΑΣ	1994	ΕΛΛΗΝΙΚΗ	ΑΝΕΡΓΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2014
8	ΑΝΔΡΑΣ	1990	ΕΛΛΗΝΙΚΗ	ΒΟΗΘΟΣ ΨΥΚΤΙΚΟΥ	ΘΕΣΣΑΛΟΝΙΚΗ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2014
9	ΑΝΔΡΑΣ	1986	ΜΑΡΟΚΙΝΗ	ΕΡΓΑΤΗΣ	ΤΙΡΑΝΑ ΑΛΒΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2014
10	ΑΝΔΡΑΣ	1974	ΑΛΒΑΝΙΚΗ	ΕΛΑΙΟΧΡΩΜΑΤΙΣΤΗΣ	ΠΟΓΡΑΔΕΤΣ ΑΒΛΑΝΙΑΣ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2014
11	ΑΝΔΡΑΣ	1966	ΕΛΛΗΝΙΚΗ-ΡΩΣΙΚΗ	ΕΡΓΑΤΗΣ	ΚΑΥΚΑΣΟ ΡΩΣΙΑΣ	ΔΙΑΖΕΥΓΜΕΝΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
12	ΑΝΔΡΑΣ	1967	ΕΛΛΗΝΙΚΗ-ΓΕΩΡΓΙΑΝΗ	ΟΙΚΟΔΟΜΟΣ	ΓΕΩΡΓΙΑ	ΔΙΑΖΕΥΓΜΕΝΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
13	ΑΝΔΡΑΣ	1986	ΡΟΥΜΑΝΙΚΗ	ΑΝΕΡΓΟΣ	ΓΑΛΑΤΣΙ ΡΟΥΜΑΝΙΑΣ	ΑΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
14	ΑΝΔΡΑΣ	1988	ΡΟΥΜΑΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΜΠΡΑΙΛΑ ΡΟΥΜΑΝΙΑΣ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
15	ΑΝΔΡΑΣ	1984	ΡΟΥΜΑΝΙΚΗ	ΟΙΚΟΔΟΜΟΣ	ΜΠΡΑΙΛΑ ΡΟΥΜΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
16	ΑΝΔΡΑΣ	1989	ΡΟΥΜΑΝΙΚΗ	ΑΝΕΡΓΟΣ	ΦΟΞΑΝ ΡΟΥΜΑΝΙΑΣ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
17	ΑΝΔΡΑΣ	1972	ΠΑΚΙΣΤΑΝΙΚΗ	ΕΡΓΑΤΗΣ ΣΕ ΑΓΡΟΤΙΚΕΣ ΕΡΓΑΣΙΕΣ	ΠΑΚΙΣΤΑΝ	ΕΓΓΑΜΟΣ	4	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
18	ΘΗΛΥ	1957	ΕΛΛΗΝΙΚΗ	ΟΙΚΙΑΚΑ	ΠΕΔΙΝΟ ΚΙΛΚΙΣ	ΕΓΓΑΜΟΣ	4	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
19	ΑΝΔΡΑΣ	1951	ΒΟΥΛΓΑΡΙΚΗ	ΟΙΚΟΔΟΜΟΣ	Μπότεγκραβ Βουλγαρίας	ΔΙΑΖΕΥΓΜΕΝΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
20	ΑΝΔΡΑΣ	1986	ΒΟΥΛΓΑΡΙΚΗ	ΕΡΓΑΤΗΣ	Γκρίεζα Βουλγαρίας	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
21	ΑΝΔΡΑΣ	1973	ΕΛΛΗΝΙΚΗ	ΟΙΚΟΔΟΜΟΣ	ΓΕΩΡΓΙΑ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
22	ΑΝΔΡΑΣ	1991	ΕΛΛΗΝΙΚΗ	ΜΙΚΡΟΠΩΛΗΤΗΣ	ΒΕΡΟΙΑ	ΑΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
23	ΑΝΔΡΑΣ	1980	ΕΛΛΗΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΚΑΒΑΛΑ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
24	ΑΝΔΡΑΣ	1968	ΕΛΛΗΝΙΚΗ	ΤΕΧΝΙΤΗΣ	ΣΤΟΥΤΓΚΑΡΔΗ ΓΕΡΜΑΝΙΑΣ	ΔΙΑΖΕΥΓΜΕΝΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
25	ΑΝΔΡΑΣ	1979	ΕΛΛΗΝΙΚΗ-ΚΑΖΑΚΣΤΑΝ	ΥΔΡΑΥΛΙΚΟΣ	ΚΑΖΑΚΣΤΑΝ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
26	ΑΝΔΡΑΣ	1988	ΤΥΝΗΣΙΑΚΗ	ΒΑΦΕΑΣ	ΜΟΝΑΣΤΗΡΙ ΤΥΝΗΣΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
27	ΑΝΔΡΑΣ	1989	ΡΟΥΜΑΝΙΚΗ	ΑΝΕΡΓΟΣ	ΕΛΡΑΣΙ ΡΟΥΜΑΝΙΑ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
28	ΑΝΔΡΑΣ	1966	ΕΛΛΗΝΙΚΗ-ΚΙΡΓΙΣΤΑΝ	ΗΛΕΚΤΡΟΛΟΓΟΣ ΟΔΗΓΟΣ ΕΡΓΑΤΗΣ	ΚΙΡΓΙΣΤΑΝ	ΔΙΑΖΕΥΓΜΕΝΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
29	ΑΝΔΡΑΣ	1965	ΕΛΛΗΝΙΚΗ	ΑΥΤΟΚΙΝΗΤΙΣΤΗΣ	ΜΑΡΟΥΣΙ ΑΤΤΙΚΗΣ	ΕΓΓΑΜΟΣ	2	ΑΘΩΩΤΙΚΗ	2013
30	ΑΝΔΡΑΣ	1957	ΕΛΛΗΝΙΚΗ	ΚΑΤΑΔΥΤΙΚΕΣ ΕΡΓΑΣΙΕΣ	ΠΑΛΑΙΟΧΩΡΙ ΧΑΛΚΙΔΙΚΗΣ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
31	ΑΝΔΡΑΣ	1987	ΕΛΛΗΝΙΚΗ	ΦΟΙΤΗΤΗΣ ΤΕΙ ΠΡΕΒΕΖΑΣ	ΒΕΡΟΙΑ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
32	ΘΗΛΥ	1961	ΕΛΛΗΝΙΚΗ	ΑΓΡΟΤΙΣΣΑ ΣΥΝΤΑΞΙΟΥΧΟΣ	ΠΑΛΑΙΟΓΡΑΤΣΑΝΟ ΚΟΖΑΝΗΣ	ΕΓΓΑΜΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
32	ΘΗΛΥ	1961	ΕΛΛΗΝΙΚΗ	ΑΓΡΟΤΙΣΣΑ ΣΥΝΤΑΞΙΟΥΧΟΣ	ΠΑΛΑΙΟΓΡΑΤΣΑΝΟ ΚΟΖΑΝΗΣ	ΕΓΓΑΜΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
33	ΑΝΔΡΑΣ	1981	ΕΛΛΗΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΠΡΕΒΕΖΑ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2013
34	ΑΝΔΡΑΣ	1939	ΕΛΛΗΝΙΚΗ	ΣΥΝΤΑΞΙΟΥΧΟΣ ΙΚΑ	ΚΟΚΚΙΝΙΑ ΚΙΛΚΙΣ	ΔΙΑΖΕΥΓΜΕΝΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2012
35	ΑΝΔΡΑΣ	1972	ΕΛΛΗΝΙΚΗ	ΓΕΩΡΓΟΣ ΟΙΚΟΔΟΜΟΣ	ΚΑΤΕΡΙΝΗ ΠΙΕΡΙΑΣ	ΔΙΑΖΕΥΓΜΕΝΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2012
36	ΑΝΔΡΑΣ	1982	ΑΛΒΑΝΙΚΗ	ΕΡΓΑΤΗΣ ΣΕ ΑΓΡΟΤΙΚΕΣ ΕΡΓΑΣΙΕΣ	ΚΟΥΤΟΥΡΜΑΝ ΑΛΒΑΝΙΑΣ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2012
37	ΘΗΛΥ	1980	ΕΛΛΗΝΙΚΗ	ΥΠΑΛΛΗΛΟΣ ΚΑΝΤΙΝΑΣ	ΛΑΚΚΙΑ ΦΛΩΡΙΝΑΣ	ΕΓΓΑΜΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2012
38	ΑΝΔΡΑΣ	1986	ΑΛΒΑΝΙΚΗ	ΕΡΓΑΤΗΣ	ΝΤΟΥΡΕΣ ΑΛΒΑΝΙΑΣ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2012
39	ΑΝΔΡΑΣ	1971	ΜΠΟΥΡΚΙΝΑ ΦΑΣΟ	ΕΡΓΑΤΗΣ	ΜΠΟΝ ΓΚΑΜΠΕ ΜΠΟΥΡΓΚΙΝ	ΕΓΓΑΜΟΣ	4	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2012
40	ΑΝΔΡΑΣ	1980	ΕΛΛΗΝΙΚΗ	ΑΓΡΟΤΗΣ	ΚΙΛΚΙΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2012
41	ΑΝΔΡΑΣ	1989	ΡΟΥΑΝΤΑ	ΕΠΙΧΕΙΡΗΜΑΤΙΑΣ ΠΟΔΟΣΦΑΙΡΙΣΤΗ	ΡΟΥΑΝΤΑ	ΑΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2012
42	ΑΝΔΡΑΣ	1982	ΝΙΓΗΡΙΑ	ΜΙΚΡΟΠΩΛΗΤΗΣ Κ ΠΟΔΟΣΦΑΙΡΙΣΤ	ΝΙΓΗΡΙΑ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2012
43	ΑΝΔΡΑΣ	1986	ΝΙΓΗΡΙΑ	ΜΙΚΡΟΠΩΛΗΤΗΣ	ΝΙΓΗΡΙΑ	ΕΓΓΑΜΟΣ	0	ΑΘΩΩΤΙΚΗ	2012
44	ΑΝΔΡΑΣ	1984	ΒΟΥΛΓΑΡΙΚΗ	ΕΡΓΑΤΗΣ	ΜΠΟΤΕΒΓΚΡΑΝΤ ΒΟΥΛΓΑΡΙΑΣ	ΑΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2012
45	ΘΗΛΥ	1981	ΕΛΛΗΝΙΚΗ ΓΕΩΡΓΙΑΝΗ	ΟΙΚΙΑΚΑ	ΤΥΦΛΙΔΑ ΓΕΩΡΓΙΑΣ	ΕΓΓΑΜΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
46	ΑΝΔΡΑΣ	1942	ΕΛΛΗΝΙΚΗ	ΑΓΡΟΤΗΣ	ΤΥΡΟΛΟΗ ΣΕΡΡΩΝ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
47	ΑΝΔΡΑΣ	1974	ΕΛΛΗΝΙΚΗ	ΜΑΓΕΙΡΑΣ	ΓΑΖΩΡΟ ΣΕΡΡΩΝ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
48	ΑΝΔΡΑΣ	1964	ΕΛΛΗΝΙΚΗ	ΑΓΡΟΤΗΣ	ΚΡΩΜΝΗ ΠΕΛΛΑΣ	ΔΙΑΖΕΥΓΜΕΝΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
49	ΑΝΔΡΑΣ	1991	ΑΛΒΑΝΙΚΗ	ΜΑΘΗΤΗΣ	ΤΙΡΑΝΑ ΑΛΒΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
50	ΑΝΔΡΑΣ	1991	ΑΛΒΑΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΦΙΕΡ ΑΛΒΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
51	ΑΝΔΡΑΣ	1976	ΕΛΛΗΝΙΚΗ	ΜΑΓΕΙΡΑΣ	ΑΛΕΞΑΝΔΡΟΥΠΟΛΗ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
52	ΑΝΔΡΑΣ	1985	ΕΛΛΗΝΙΚΗ ΓΕΩΡΓΙΑΝΗ	ΑΝΕΡΓΟΣ	ΤΥΦΛΙΔΑ ΓΕΩΡΓΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
53	ΑΝΔΡΑΣ	1987	ΓΕΩΡΓΙΑΝΗ	ΟΙΚΟΔΟΜΟΣ	ΓΕΩΡΓΙΑ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
54	ΑΝΔΡΑΣ	1949	ΓΕΩΡΓΙΑΝΗ	ΟΙΚΟΔΟΜΟΣ	ΓΕΩΡΓΙΑ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
55	ΑΝΔΡΑΣ	1972	ΓΕΩΡΓΙΑΝΗ	ΗΛΕΚΤΡΟΣΥΓΚΟΛΛΗΤΗΣ	ΤΥΦΛΙΔΑ ΓΕΩΡΓΙΑΣ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
56	ΑΝΔΡΑΣ	1982	ΕΛΛΗΝΙΚΗ	ΑΝΕΡΓΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
57	ΑΝΔΡΑΣ	1977	ΑΡΜΕΝΙΚΗ	ΑΘΛΗΤΗΣ	ΑΡΜΕΝΙΑ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2011
58	ΑΝΔΡΑΣ	1963	ΕΛΛΗΝΙΚΗ	ΤΥΠΟΓΡΑΦΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2010
59	ΘΗΛΥ	1989	ΑΛΒΑΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΑΛΒΑΝΙΑ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2010
60	ΑΝΔΡΑΣ	1966	ΕΛΛΗΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΔΙΑΖΕΥΓΜΕΝΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2010
61	ΘΗΛΥ	1969	ΕΛΛΗΝΙΚΗ	ΟΙΚΙΑΚΑ	ΣΟΧΟΥΜΙ ΓΕΩΡΓΙΑΣ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2010
62	ΑΝΔΡΑΣ	1969	ΕΛΛΗΝΙΚΗ ΓΕΡΜΑΝΙΚΗ	ΞΕΝΟΔΟΧΟΥΠΑΛΛΗΛΟΣ	ΓΕΡΜΑΝΙΑ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2010
63	ΑΝΔΡΑΣ	1931	ΕΛΛΗΝΙΚΗ	ΟΔΗΓΟΣ ΣΥΝΤΑΞΙΟΥΧΟΣ	ΟΡΜΥΛΙΑ ΧΑΛΚΙΔΙΚΗΣ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2010
64	ΑΝΔΡΑΣ	1978	ΕΛΛΗΝΙΚΗ	ΓΕΩΡΓΟΣ	ΣΕΡΡΕΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2010
65	ΘΗΛΥ	1966	ΕΛΛΗΝΙΚΗ	ΟΙΚΙΑΚΑ	ΘΕΣΣΑΛΟΝΙΚΗ	ΔΙΑΖΕΥΓΜΕΝΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2009

66	ΑΝΔΡΑΣ	1970	ΕΛΛΗΝΙΚΗ	ΕΛΕΥΘΕΡΟΣ ΕΠΑΓΓΕΛΜΑΤΙΑΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΕΓΓΑΜΟΣ	0	ΑΘΩΩΤΙΚΗ	2009
67	ΑΝΔΡΑΣ	1932	ΕΛΛΗΝΙΚΗ	ΣΥΝΤΑΞΙΟΥΧΟΣ	ΣΕΡΡΕΣ	ΕΓΓΑΜΟΣ		ΑΘΩΩΤΙΚΗ	2009
68	ΑΝΔΡΑΣ	1979	ΕΛΛΗΝΙΚΗ	ΠΑΛΙΑΤΖΗΣ	ΣΕΡΡΕΣ	ΔΙΑΖΕΥΓΜΕΝΟΣ	5	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2009
69	ΑΝΔΡΑΣ	1974	ΕΛΛΗΝΙΚΗ	ΘΕΡΜΟΥΔΡΑΥΛΙΚΟΣ	ΓΕΡΜΑΝΙΑ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2009
70	ΘΗΛΥ	1978	ΕΛΛΗΝΙΚΗ	ΟΙΚΙΑΚΑ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2009
71	ΘΗΛΥ	1942	ΕΛΛΗΝΙΚΗ	ΟΙΚΙΑΚΑ	ΑΝΤΙΦΙΛΙΠΠΟΙ ΚΑΒΑΛΑΣ	ΧΗΡΑ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2009
72	ΑΝΔΡΑΣ	1942	ΕΛΛΗΝΙΚΗ	ΟΙΚΟΔΟΜΟΣ ΣΥΝΤΑΞΙΟΥΧΟΣ	ΜΕΓΑΛΑ ΛΙΒΑΔΙΑ ΚΙΛΚΙΣ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2009
73	ΑΝΔΡΑΣ	1960	ΕΛΛΗΝΙΚΗ	ΥΠΟΔΗΜΑΤΟΠΟΙΟΣ	ΛΕΒΕΝΤΟΧΩΡΙ ΚΙΛΚΙΣ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2009
74	ΑΝΔΡΑΣ	1987	ΑΛΒΑΝΙΚΗ	ΕΡΓΑΤΗΣ	ΚΟΡΥΤΣΑ ΑΛΒΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΑΘΩΩΤΙΚΗ	2008
75	ΑΝΔΡΑΣ	1972	ΙΟΡΔΑΝΙΑ	ΦΟΙΤΗΤΗΣ	ΙΟΡΔΑΝΙΑ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2008
76	ΑΝΔΡΑΣ	1982	ΑΛΒΑΝΙΚΗ	ΟΙΚΟΔΟΜΟΣ	ΑΛΒΑΝΙΑ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2008
77	ΑΝΔΡΑΣ	1981	ΕΛΛΗΝΙΚΗ	ΟΔΗΓΟΣ ΒΑΡΕΩΝ ΟΧΗΜΑΤΩΝ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΑΘΩΩΤΙΚΗ	2008
78	ΑΝΔΡΑΣ	1986	ΑΛΒΑΝΙΚΗ	ΠΛΑΚΑΤΖΗΣ	ΤΙΡΑΝΑ ΑΛΒΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2008
79	ΑΝΔΡΑΣ	1988	ΕΛΛΗΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΤΣΑΛΚΑ ΓΕΩΡΓΙΑΣ	ΑΓΑΜΟΣ	0	ΑΘΩΩΤΙΚΗ	2008
80	ΑΝΔΡΑΣ	1974	ΕΛΛΗΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΚΑΒΑΛΑ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2008
81	ΑΝΔΡΑΣ	1987	ΕΛΛΗΝΙΚΗ	ΑΝΕΡΓΟΣ	ΠΤΟΛΕΜΑΙΔΑ	ΑΓΑΜΟΣ	0	ΕΓΚΛΕΙΣΜΟΣ ΣΕ ΨΥΧΙΑΤΡΕΙΟ	2008
82	ΑΝΔΡΑΣ	1961	ΓΕΡΜΑΝΙΚΗ	ΑΝΕΡΓΟΣ	ΝΤΙΣΕΛΝΤΟΦ ΓΕΡΜΑΝΙΑΣ	ΔΙΑΖΕΥΓΜΕΝΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2008
83	ΑΝΔΡΑΣ	1938	ΕΛΛΗΝΙΚΗ	ΣΥΝΤΑΞΙΟΥΧΟΣ	ΧΑΡΩΠΟ ΣΕΡΡΩΝ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2008
84	ΑΝΔΡΑΣ	1975	ΕΛΛΗΝΙΚΗ	ΠΑΛΙΟΣΙΔΕΡΑΣ	ΣΕΡΡΕΣ	ΕΓΓΑΜΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2008
85	ΑΝΔΡΑΣ	1969	ΕΛΛΗΝΙΚΗ	ΑΝΕΡΓΟΣ	ΙΛΙΟΝ ΑΤΤΙΚΗ	ΔΙΑΖΕΥΓΜΕΝΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
86	ΑΝΔΡΑΣ	1994	ΕΛΛΗΝΙΚΗ	ΑΝΕΡΓΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
87	ΑΝΔΡΑΣ	1994	ΕΛΛΗΝΙΚΗ	ΑΝΕΡΓΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
88	ΑΝΔΡΑΣ	1986	ΕΛΛΗΝΙΚΗ	ΜΙΚΡΟΠΩΛΗΤΗΣ	ΤΣΑΛΚΑ ΓΕΩΡΓΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
89	ΑΝΔΡΑΣ	1953	ΕΛΛΗΝΙΚΗ	ΤΕΧΝΙΤΗΣ	ΠΛΑΤΥ ΗΜΑΘΙΑΣ	ΕΓΓΑΜΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
90	ΑΝΔΡΑΣ	1990	ΕΛΛΗΝΙΚΗ	ΕΠΙΠΛΟΠΟΙΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΑΘΩΩΤΙΚΗ	2015
91	ΑΝΔΡΑΣ	1981	ΕΛΛΗΝΙΚΗ	ΣΙΔΕΡΑΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΑΘΩΩΤΙΚΗ	2015
92	ΑΝΔΡΑΣ	1990	ΕΛΛΗΝΙΚΗ	ΑΝΕΡΓΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΑΘΩΩΤΙΚΗ	2015
93	ΑΝΔΡΑΣ	1986	ΕΛΛΗΝΙΚΗ	ΑΝΕΡΓΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΑΓΑΜΟΣ	0	ΑΘΩΩΤΙΚΗ	2015
94	ΑΝΔΡΑΣ	1985	ΑΛΒΑΝΙΚΗ	ΟΙΚΟΔΟΜΟΣ	ΚΟΡΥΤΣΑ ΑΛΒΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
95	ΑΝΔΡΑΣ	1990	ΑΛΒΑΝΙΚΗ	ΚΑΤΑΣΤΗΜΑΤΑΡΧΗΣ	ΚΟΡΥΤΣΑ ΑΛΒΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΑΘΩΩΤΙΚΗ	2015
96	ΑΝΔΡΑΣ	1991	ΑΛΒΑΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΚΟΡΥΤΣΑ ΑΛΒΑΝΙΑΣ	ΑΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
97	ΑΝΔΡΑΣ	1975	ΑΛΒΑΝΙΚΗ	ΤΑΠΕΤΣΕΡΗΣ	ΚΟΡΥΤΣΑ ΑΛΒΑΝΙΑΣ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
98	ΘΗΛΥ	1977	ΑΛΒΑΝΙΚΗ	ΓΑΖΩΤΡΙΑ	ΚΟΡΥΤΣΑ ΑΛΒΑΝΙΑΣ	ΕΓΓΑΜΟΣ	2	ΑΘΩΩΤΙΚΗ	2015
99	ΑΝΔΡΑΣ	1974	ΕΛΛΗΝΙΚΗ	ΑΓΡΟΤΗΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΔΙΑΖΕΥΓΜΕΝΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
100	ΘΗΛΥ	1978	ΕΛΛΗΝΙΚΗ	ΜΕΣΙΤΡΙΑ	ΓΕΡΜΑΝΙΑ	ΑΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
101	ΑΝΔΡΑΣ	1963	ΕΛΛΗΝΙΚΗ	ΕΛΕΥΘΕΡΟΣ ΕΠΑΓΓΕΛΜΑΤΙΑΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΕΓΓΑΜΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
102	ΑΝΔΡΑΣ	1979	ΓΕΩΡΓΙΑΝΗ	ΠΛΑΚΑΤΖΗΣ	ΓΚΟΡΙ ΓΕΩΡΓΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
103	ΑΝΔΡΑΣ	1971	ΓΕΩΡΓΙΑΝΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΚΑΡΕΛΙ ΓΕΩΡΓΙΑΣ	ΕΓΓΑΜΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
104	ΑΝΔΡΑΣ	1974	ΓΕΩΡΓΙΑΝΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	Mtskheta Γεωργίας	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
105	ΘΗΛΥ	1978	ΓΕΩΡΓΙΑΝΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	Mtskheta Γεωργίας	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
106	ΑΝΔΡΑΣ	1983	ΕΛΛΗΝΙΚΗ ΤΣΙΓΓΑΝΙΚΗ	ΠΑΛΙΑΤΖΗΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΕΓΓΑΜΟΣ	4	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
107	ΑΝΔΡΑΣ	1987	ΚΙΝΕΖΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΦΟΥΤΣΙΕΝ ΚΙΝΑΣ	ΑΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
108	ΑΝΔΡΑΣ	1966	ΕΛΛΗΝΙΚΗ-ΚΑΖΑΚΣΤΑΝ	ΕΡΓΑΤΗΣ	ΝΤΖΑΜΠΟΥΛ ΚΑΖΑΚΣΤΑΝ	ΔΙΑΖΕΥΓΜΕΝΟΣ	3	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
109	ΑΝΔΡΑΣ	1978	ΕΛΛΗΝΙΚΗ-ΚΑΖΑΚΣΤΑΝ	ΕΡΓΑΤΗΣ	ΝΤΖΑΜΠΟΥΛ ΚΑΖΑΚΣΤΑΝ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
110	ΑΝΔΡΑΣ	1960	ΕΛΛΗΝΙΚΗ-ΓΕΩΡΓΙΑΝΗ	ΕΡΓΑΤΗΣ	ΚΒΕΜΟ-ΚΕΝΤΙ ΓΕΩΡΓΙΑΣ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
111	ΘΗΛΥ	1961	ΕΛΛΗΝΙΚΗ-ΓΕΩΡΓΙΑΝΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΚΒΕΜΟ-ΚΕΝΤΙ ΓΕΩΡΓΙΑΣ	ΕΓΓΑΜΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
112	ΑΝΔΡΑΣ	1966	ΕΛΛΗΝΙΚΗ-ΓΕΩΡΓΙΑΝΗ	ΜΗΧΑΝΙΚΟΣ ΑΥΤΟΚΙΝΗΤΩΝ	ΚΒΕΜΟ-ΚΕΝΤΙ ΓΕΩΡΓΙΑΣ	ΕΓΓΑΜΟΣ	3	ΑΘΩΩΤΙΚΗ	2015
113	ΑΝΔΡΑΣ	1941	ΕΛΛΗΝΙΚΗ	ΜΗΧΑΝΟΥΡΓΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΔΙΑΖΕΥΓΜΕΝΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
114	ΑΝΔΡΑΣ	1968	ΕΛΛΗΝΙΚΗ-ΚΑΖΑΚΣΤΑΝ	ΥΔΡΑΥΛΙΚΟΣ	ΚΑΖΑΚΣΤΑΝ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
115	ΑΝΔΡΑΣ	1957	ΕΛΛΗΝΙΚΗ	ΜΗΧΑΝΙΚΟΣ ΑΥΤΟΚΙΝΗΤΩΝ	ΒΡΟΝΤΕΡΟ ΤΡΙΚΑΛΩΝ	ΧΗΡΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
116	ΑΝΔΡΑΣ	1992	ΑΛΒΑΝΙΚΗ	ΑΝΕΡΓΟΣ	ΛΟΥΤΣΙΑ ΑΛΒΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
117	ΑΝΔΡΑΣ	1982	ΡΟΥΜΑΝΙΚΗ	ΑΝΕΡΓΟΣ	ΚΟΥΝΤΕΡΑ ΡΟΥΜΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
118	ΘΗΛΥ	1979	ΡΟΥΜΑΝΙΚΗ	ΑΝΕΡΓΟΣ	ΛΟΜΝΕΣ ΡΟΥΜΑΝΙΑ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
119	ΑΝΔΡΑΣ	1982	ΕΛΛΗΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΘΕΣΣΑΛΟΝΙΚΗ	ΕΓΓΑΜΟΣ	1	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
120	ΑΝΔΡΑΣ	1965	ΕΛΛΗΝΙΚΗ	ΣΥΝΤΑΞΙΟΥΧΟΣ	ΜΥΡΙΝΑ ΛΗΜΝΟΥ	ΔΙΑΖΕΥΓΜΕΝΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
121	ΑΝΔΡΑΣ	1986	ΟΥΚΡΑΝΙΚΗ	ΣΕΡΒΙΤΟΡΟΣ	ΝΤΟΝΑΤΣΚ ΟΥΚΡΑΝΙΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
122	ΑΝΔΡΑΣ	1966	ΕΛΛΗΝΙΚΗ	ΙΔΙΟΚΤΗΤΗΣ ΕΝΟΙΚΙΑΖΟΜΕΝΩΝ ΔΣ	ΠΟΛΥΓΥΡΟΣ ΧΑΛΚΙΔΙΚΗΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
123	ΑΝΔΡΑΣ	1963	ΒΟΥΛΓΑΡΙΚΗ	ΑΝΕΡΓΟΣ	ΒΟΥΛΓΑΡΙΑ	ΔΙΑΖΕΥΓΜΕΝΟΣ	2	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015
124	ΑΝΔΡΑΣ	1988	ΕΛΛΗΝΙΚΗ	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	ΓΙΑΝΝΙΤΣΑ ΠΕΛΛΑΣ	ΑΓΑΜΟΣ	0	ΚΑΤΑΔΙΚΑΣΤΙΚΗ	2015

Training data of GCDT classifier (sample)

text	hapax	dis	num_of_words	content words	lexical_density	function words	function words	functional_density	Avg_word_leng	Word_length_	Avg_sentence_	Sentence_length_	guilty
d1	0.0594	0.0536	0.0465	0.0685	0.44	0.035	0.56	1.29	4.58	2.41	11.79	11.31	1
d2	0.0480	0.0324	0.0226	0.0363	0.48	0.016	0.52	1.1	4.77	2.62	9.91	6.5	1
d3	0.0143	0.0128	0.0062	0.0089	0.42	0.005	0.58	1.37	4.66	2.39	5.44	3.74	1
d4	0.0393	0.0340	0.0169	0.0260	0.46	0.012	0.54	1.19	4.8	2.62	8.54	5.55	1
d5	0.0378	0.0335	0.0297	0.0436	0.44	0.023	0.56	1.3	4.4	2.18	9.85	7.08	1
d6	0.0387	0.0385	0.0260	0.0384	0.44	0.02	0.56	1.28	4.54	2.35	7.7	5.3	1
d7	0.0395	0.0368	0.0286	0.0421	0.44	0.022	0.56	1.29	4.55	2.39	8.46	5.74	1
d8	0.0211	0.0206	0.0095	0.0146	0.46	0.007	0.54	1.19	4.6	2.37	6.91	4.4	1
d9	0.0209	0.0279	0.0163	0.0228	0.41	0.013	0.59	1.41	4.38	2.09	8.08	5.88	1
d10	0.0194	0.0648	0.0379	0.0571	0.45	0.028	0.55	1.23	4.59	2.4	10.31	11.52	1
d11	0.0174	0.0100	0.0056	0.0095	0.5	0.004	0.5	1.01	4.72	2.31	7.45	3.8	1
d12	0.0157	0.0279	0.0094	0.0145	0.46	0.007	0.54	1.2	4.65	2.32	6.97	6.43	1
d13	0.0068	0.0050	0.0021	0.0031	0.44	0.002	0.56	1.26	5.03	2.82	6.24	3.89	1
d14	0.0098	0.0033	0.0020	0.0032	0.48	0.001	0.52	1.08	4.69	2.31	5.56	4.31	1
d15	0.0134	0.0106	0.0058	0.0077	0.39	0.005	0.61	1.54	4.53	2.6	11.23	6.45	1
d16	0.0101	0.0128	0.0038	0.0056	0.44	0.003	0.56	1.29	4.9	2.69	7.19	4.84	1
d17	0.0234	0.0296	0.0161	0.0245	0.45	0.012	0.55	1.22	4.41	2.13	10.92	7.72	1
d18	0.0114	0.0128	0.0053	0.0077	0.43	0.004	0.57	1.35	4.6	2.28	4.88	3.24	1
d19	0.0310	0.0285	0.0174	0.0252	0.43	0.013	0.57	1.32	4.57	2.42	8.51	6.14	1
d20	0.0245	0.0262	0.0131	0.0191	0.43	0.01	0.57	1.32	4.52	2.38	7.95	5.08	1