AUTHORSHIP ATTRIBUTION IN TURKISH TEXTS

DR. HÜLYA KOCAGÜL YÜZER

ARTSÜREM

ISBN: 978-605-72285-0-5

Yayıncı / Publisher:
Artsürem Bilim Sanat Danışmanlık Bilişim AŞ.

Genel Yayın Yönetmeni / Editor - Publishing Director:
Fırat BAŞBUĞ

T.C. Kültür ve Turizm Bakanlığı Yayıncı Sertifika Numarası / Publisher Certificate Number:
46183

ISBN:
978-605-72285-0-5

Açık Erişim Tarihi / Open Access Date:
2022

Bu çalışma yazarın doktora tezinden üretilmiştir. Bu çalışmaya dünyanın birçok üniversite kütüphanesinden erişilebilir.

artsurem@artsurem.com
www.artsurem.com

SUMMARY OF BOOK

Authorship Attribution in Turkish Texts

Dr. Hülya Kocagül Yüzer

The latest developments in the field of computer technology have created new ways to share information without time and space limits. Computer technologies have not only made life easier and more accessible for users, but they have also opened up a new arena for illegal activities. These illegal actions have found an opportunity to spread via e-mails, websites, Internet chat rooms, forum pages, and social networking websites (like Facebook, Twitter, Instagram). Online contributors do not need to provide information such as their real names, the city where they live, age or gender in order to share their opinions, and such feelings of anonymity encourage criminal activities. Thus, disputed authorship cases have become one of the main challenges of the technological era.

This research is a corpus-based simulated authorship casework application in Turkish. Texts for the corpora were collected from a collaborative online encyclopaedia – Eksi Sozluk (Sour Times) and Twitter. The corpus consists of 900 texts from 52 authors in total. However, 105 texts belong to seven authors from Twitter.

The two methodological approaches that were applied are qualitative and statistical methods, according to Grant's (2013) approach. Ten different tests were applied, depending on the various parameters that are forensically possible in real-world cases. Accordingly, the role of feature type, size, including the candidate author size, text size and a limited number of texts per author and finally cross-genre application were tested. The analyses revealed that such a combined approach has promising results in some tests in that they attributed authorship in Turkish. The findings of the research indicated that there is the potential to attribute unknown authors in Turkish and it appears that the results have significant conclusions for the broader application of forensic authorship attribution techniques in Turkish texts.

Keywords: Authorship Attribution, Turkish, Forensic Linguistics, Authorship Analysis

Table of Content

List of Figures

List of Tables

Chapter 1: Introduction

In recent years, forensic linguists' input relating to several applications of linguistic analysis in a forensic context, for example, voice analysis, trademark cases, native language identification, dialect identification, authorship profiling, discourse analysis, and detecting authorship in debatable cases, has been increasingly used in courts (McMenamin, 2002). In this context, forensic linguistics applies linguistic methods to forensic text research, including authorship problems (e.g. Grant, 2013), explaining the textual meaning of legal documents (e.g. Tiersma, 2010), and police interviews (e.g. Macleod, 2010; Oxburgh et al., 2015). Even a single word can play a major role in the court process, such as in the case of the misattributed and posthumously pardoned Derek Bentley in the 1950s (Coulthard, 2005). Despite the number of studies on other applications, authorship analysis is one of the widely known problems in forensic linguistics. Because of its popularity, many scholars have proposed various approaches and outlined different scenarios and problems in authorship analysis.

According to Hänlein (1999), 'authorship analysis', relating to the situation where large-scale materials are compared with examined texts, is 'a process of examining the characteristics of a piece of writing to draw conclusions on its authorship' (Zheng et al., 2006, p.379). In general, the authorship problem is formulated as assigning a text of unknown authorship to one of the candidate authors from a given set of authors and samples of written texts that belong to them (Stamatatos, 2009).

McMenamin (2002) identified three main problems in authorship studies as:

> (i) one may want to determine if one author wrote all the writings in a questioned set (ii) one may be asked to compare a questioned writing with the writings of a large number of possible authors (iii) the most common type of forensic problem is to assess the resemblance of a questioned writing to that of one author or a small number of candidate authors. (McMenamin, 2002, p.76)

The analysis of each problem is outlined in three models: the resemblance, consistency and population models. The resemblance model tries to identify the suspect author from among a small number of authors, the consistency model is 'used to determine whether two or more writings were written from the same author', and the population model is 'occasionally used in forensic contexts that do not provide external (non-linguistic) evidence suggesting just one or two candidate writers' (McMenamin, 2002, pp.117-118).

Based on these problems, authorship analysis can be evaluated in different ways regarding the identification of an author: authorship identification, authorship characterisation or authorship profiling, similarity detection, and authorship attribution. Although each of these processes requires different methodological approaches, all of them broadly seek to deduce the owner of a piece of written data. Zheng et al. (2006) categorised the different identification methods in authorship analysis.

First, authorship identification looks at the likelihood of whether one particular author produced the writing. This is known as authorship attribution in some studies. Authorship attribution is the most common authorship task with the aim of finding the author of unknown texts among known authors. It is used by Koppel et al. (2013) as an umbrella term for all the problems when a disputed text is possibly attributable to a small set of suspects. In this case, there is one possible author for the disputed texts (Juola, 2008). Solan and Tiersma (2005) defined this task as where there is a limited number of possible closed set authors, the author of the anonymous text or texts should be identified.

Second, authorship characterisation produces the sociological background based on the writing, such as native language, cultural background, gender, and educational background (Zheng et al., 2006). Koppel et al. (2009, p.3) described the profiling problem when 'there is no candidate set all' it is required 'to provide as much demographic or psychological information as possible about the author' this case is referred to as authorship categorisation by Abbasi and Chen (2005) and Zheng et al. (2006). It aims to summarize the characteristics of the author of a text and find out the author profile based on the anonymous document. Stamatatos (2009) classified authorship characterisation and profiling as the same problems since both of them focus on the same method to identify the author of a given text. This is also known as authorship profiling in some forensic linguistics research.

Authorship profiling has a more straightforward form when compared to the problems of authorship attribution which focuses on the question, 'What kind of a person wrote this text?' (Grant, 2008, p.222) while aiming to deduce the author's characteristics without depending on any psychological observations of him/her. It does not start with known texts from known authors; instead, it is grounded in sociolinguistic research based on demographics or language information about the author. For example, Nini (2015) carried out authorship profiling research that considered profile dimensions such as gender, age, level of education, and social class.

Olsson (2004, p.98) used authorship profiling and stated that 'this is not a psychological type of study'; instead, it aims to describe 'the profile of an individual as an author, rather than the author as an individual, in other words, the sum of authorship characteristics which describe the author *qua* author' (ibid.). According to Coulthard (2011) authorship profiling 'involves taking a single example and matching it to a well-founded generalisation, drawing a conclusion about that instance' (p.538).

Finally, Zheng et al. (2006, p.379) stated that similarity detection 'compares multiple pieces of writing and determines whether they were produced by a single author without actually identifying the author'. This is related to plagiarism detection in forensic linguistics.

Furthermore, some other terms are used in the authorship analysis task. For instance, Abbasi and Chen (2005) used the term 'authorship identification', which focuses on analysing the disputed texts based on the similarities between an author's known texts, which is called authorship attribution in other studies. Similarly, Luyckx (2010) suggested that authorship attribution is the same concept as authorship identification or authorship recognition and described it as a task of discriminating the author of a disputed text from a set of candidate authors. However, authorship attribution is different from other authorship categorisation tasks as it focuses on stylistic information which is based on frequent patterns (Stamatatos, 2013).

Juola (2008, pp. 238-239) categorised the authorship attribution scenarios based on the candidate author numbers as (i) given a particular sample of text known to be by one of a set of authors, determine which one (closed-open candidate set); and (ii) given a particular sample of text believed to be by one of a set of authors, determine which one or who wrote it (open-class). This is described as the verification problem by Koppel et al. (2006). Similarly, Koppel et al. (2011) approached the problem in terms of candidate author numbers and unknown texts as: (i) there is an unknown text and the author is not among the suspects; (ii) there may be thousands of known candidate authors, previously described as the 'needle-in-a-haystack problem' (Koppel, 2009, p.3); (iii) the known texts from each candidate or the anonymous text might be very limited (Koppel et al., 2011, p.84).

Nevertheless, Coulthard (2004: 432) indicated that 'the task of the linguistic detective is never one of identifying an author from millions of candidates on the basis of the linguistic evidence alone, but rather of selecting (and, of course, sometimes deselecting) one author from a very small number of candidates'.

As stated above, authorship studies are categorised differently in previous works, since in a criminal case, forensic linguists apply different methods to analyse questioned texts to obtain the most reliable results for the particular case under investigation.

However, the current study adopts the authorship attribution work definition by Solan and Tiersma (2005) and McMenamin's (2002) resemblance model because the number of candidate author is relatively small when the cases with thousands of candidate authors are considered as it is defined as '*needle-in-a-haystack'* by Koppel et al. (2009, p. 3). This view is also supporting the view of Coulthard (2004) in selecting an author from a very small number of candidates.

In the following sections, recent trends in online crime (Section 1.1.), the reliability of forensic linguistics evidence (Section 1.2.) in the legal context, and the rationale of the study (Section 1.3) are outlined. The research questions of the study (Section 1.4.) and the book overview (Section 1.5.) are presented in the last part of this chapter.

1.1. Recent Trends in Online Crime

The internet has played a vital role in the life of people in the modern world for the past 20 years. On the one hand, having access to large sources and its speed, being free of charge and availability have changed many things. On the other hand, the convenience of creating 'fake online personas' easily and operating them to give a false impression or participating in criminal activities has led to many cases in an online setting. The internet provides the facilities to commit a crime with little risk of revealing the real identity of the users, who tend to 'talk more and in different ways from their real-world linguistic repertoire' (Crystal, 2011, p.14).

In December 2014, a pseudonymous Twitter account, Fuat Avni, started to leak confidential information about the Turkish president and members of parliament. The account holder was portraying a government insider profile and had almost three million followers. As time passed, most of his/her tweets turned out to be true. Therefore, the account gained the public's trust and more popularity. In one of his/her interviews on Twitter, the account holder claimed that s/he had been working in a sensitive position within the ruling party.

Therefore, the information that s/he was leaking was reflecting the truth. The account has gained a lot of interest and popularity among the Turkish internet users as it was leaking much information about the government.

Although the Turkish National Intelligence Organization has repeatedly attempted to find out the identity of the account holder, up to now, there has been no exact answer to this question.

There are many disputable cases like this pseudonymous Twitter account which are sued on the linguistic data in computer-mediated communication mediums such as terrorist group blogs, online messages, and Twitter or Facebook messages.

Although there is a debate on the *right to privacy* (Coleman, 2006), it is evident that online anonymous personas have victimised many people or even governments in various ways. Such cases are investigated under the title of forensic authorship analysis/attribution. Moreover, the 'fake persona' and 'anonymity' issues, the size of texts and the number of candidate authors may cause problems in authorship analysis, since online texts are 'shorter, noisier, and they have a greater number of candidate authors' (Abbasi and Chen, 2005, p.67). The issues of text size and the number of candidate authors are discussed in the following chapters. Prior to that, some standards for analysing forensic evidence to establish reliability are discussed. The standards of the reliability in forensic linguistics are a controversial issue which has arisen in recent years. While some countries set up authorship attribution reliability criteria, others focus on the qualification of experts or both. In the next section, the reliability of forensic linguistics in courts is examined by describing the current situation in different countries, including Turkey.

1.2. Reliability of Forensic Linguistics in the Legal Context

Research on forensic linguistics is beneficial to government authorities, universities for plagiarism, companies for trademark issues, and individuals for private cases. When the cases are brought to trial, some countries require specific assessments of the expert's testimony and the method used. In criminal cases, any evidence provides reliability and validity criteria to submit to the courtroom as proof of a crime. This section of the study summarises the issue of reliability in different countries, including Turkey.

First, some legal concepts should be clarified to understand the context. Regarding the term 'expert', in the American context, Shuy stated that an expert 'is understood to be a person who has special knowledge or skill in a subject' (Shuy, 2009, p.220).

Due to the latest developments in science, such as voice recognition technologies or DNA sample investigations, experts now play a vital role in the solving of crimes.

An expert can be skilful in languages, computers, engineering, etc. in forensic science; the scientist needs 'to champion their expert opinions based on accepted, properly performed scientific inquiry' (Fisher and Fisher, 2012, p.15) and to assist the court in determining disputable cases with objective and accurate evidence.

The experts in the forensic linguistics field are mostly invited to help a court in almost any area of linguistics, including the phonetics, discourse analysis, syntax, semantics or pragmatics of legal documents under investigation (Tiersma and Solan, 2002). The forensic linguist mainly helps the court by answering such questions as 'What does the text say?' and 'Who is the author?' In this matter, forensic linguists 'draw on knowledge and techniques derived from one or more of the sub-areas of descriptive linguistics: phonetics and phonology, lexis, syntax, semantics, pragmatics, discourse and text analysis' (Coulthard, 2005, p.10).

Furthermore, Grant (2008, p.224) stated that 'understanding that different sorts of linguistic evidence may play different roles within the investigative and judicial process can be key in pursuing the forensic practice. A sociolinguistic profile might assist a police investigation but have no evidential value'. Sometimes, it is not possible to present substantial evidence from linguistic data. The expert should provide a detailed report that includes a detailed description of the analysis. In some countries, the linguist expert is supposed to satisfy specific criteria, such as the Daubert criteria in the United States of America (USA).

The Daubert criteria include the following rules:

> (a) the expert's scientific, technical, or other specialised knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
>
> (b) the testimony is based on sufficient facts or data;
>
> (c) the testimony is the product of reliable principles and methods, and
>
> (d) the expert has reliably applied the principles and methods to the facts of the case.
>
> (Federal Committee, Rule 702, 2017)

Moreover, the Federal Rules of Committee set a non-exclusive checklist for expert testimony to gauge its reliability and acceptability in US courts regarding (1) whether the expert's technique or theory can be or has been tested; (2) whether the technique or theory has been subject to peer review and publication; (3) the known or potential rate or error of the technique or theory when applied; (4) the existence and maintenance of standards and controls; and (5) whether the technique or theory has generally been accepted in the scientific community.

The said expert should apply his/her expertise in his/her field on sufficiently representative data by using the reliable principles and methods. Coulthard (2004) propounds the view that the methods of authorship attribution meet the Daubert criteria.

The first criterion grounded on the assumption of the occurrence of 'shared identical items is conclusive evidence that two texts have not been independently created' (Coulthard, 2004, p.445). The second one can be met simply when the current published forensic linguistics methods considered have complied with the first and the second criteria, as Tiersma and Solan (2002, p.225) stated that linguistics 'is a robust field that relies heavily on peer-reviewed journals for dissemination of new work'. Regarding the rate of error of the technique, Kredens (2006, p.28) claimed that 'there is indeed no known rate of error in forensic linguistic analysis' and 'idiolectal variation and the uniqueness of utterance is accepted across the whole linguistic community' (Coulthard, 2004, p.445).

In the United Kingdom (UK), unlike in the USA, the criteria do not relate to the evidence; instead, they are linked with the expert. In other words, 'they do not impose any special requirements for the admissibility of the expert evidence other than evidence and reliability' (Kredens, 2006, p.25).

The Practice Direction 35 – Experts and Assessors, requirements for an expert witness, includes the following criteria:

> (2) The opinion evidence of an expert witness is sufficiently reliable to be admitted if:
>
> (a) The evidence is predicated on sound principles, techniques and assumptions;
>
> (b) Those principles, techniques and assumptions have been properly applied to the facts of the case; and
>
> (c) The evidence is supported by (that is, logically in keeping with) those principles, techniques and assumptions as applied to the facts of the case.
>
> (The Law Commission, 2011)

The Daubert criteria are looking for sufficient facts and data which are based on reliable techniques and methods as opposed to the criteria implemented in the UK, where the expert should provide an objective, unbiased opinion to the court. The expert should never assume a

role of advocate, should state the facts or assumptions, and should indicate the borders of his/her knowledge and expertise.

In Poland, 'all expert evidence commissioned by the court is admitted automatically' (Kredens, 2006, p.29); however, the expert is an accepted person who works in the national forensic science laboratory rather than one of the independent ones.

In Australia, the requirements of the expert witness were established in 1995. While they include reliability and relevance, the experts are not required to provide any scientific evidence – they must only give reasons for their opinions (Olsson, 2004). Similarly, in the Philippines, only the expert's opinion may be accepted as evidence (Olsson, 2004). In France and Germany, experts are appointed by the court, as in Poland (ibid).

In Turkey, up to now, apart from the abovementioned Fuat Avni case, few forensic linguistics evidence-related cases have been highlighted by the Turkish media. For this reason, there are no specific regulations with regard to the admissibility of forensic linguistic expert evidence under the Turkish law. However, parallel to the UK, Turkey focuses on the expert rather than the evaluation of the evidence. As per Article 65 of the Turkish Criminal Procedure Code (2009), individuals who are eligible to take the expert stand are:

> (b) Those who are professionals in the field of science and arts that is necessary for conducting the inspection.
>
> (c) Those who are officially empowered to work as professionals in that field of inspection.

The expert should give an oath repeating the following words in front of the judicial commission at the courts of ordinary jurisdiction in compliance with Article 64 part (d):

> *'I swear on my honour and conscience, that I shall fulfil my duty pursuing the justice and by sciences and technology, in an impartial manner'.*

The expert need not take a repeated oath for every single expert testimony they give in the future (Turkish Code of Criminal Procedure, 2009).

Finally, in all the countries mentioned above, there is no agreement on assessment of the evidence, as while some of them emphasise 'scientifically valid' evidence, others give weight to the reliability of the 'expert's testimony'.

In addition to the criminal law which they applied, each language has a different nature, which is why applying an accepted linguistic method may not work in all languages regarding reliability and validation in the courts. Kniffka (2007, p.45) stated that 'the overall success of an expert opinion depends on the scientific and professional stringency of the analysis and the translation of the results into accessible language'. However, regardless of the language and the methods applied, any researcher would want to use scientific methods and obtain valid results. Therefore, Daubert accepted that in the USA, evidence rules as a 'good starting point' (Juola, 2015, p.103).

To the best of the researcher's knowledge, there have been no case that used independent expert testimony in authorship attribution in Turkey. Along with criminal evidence investigation, the Turkish Police Department has a branch that deals with voice and authorship recognition. However, none of the cases has been presented to the public due to the confidentiality. In the same way as the Daubert criteria, Turkish courts need a standard to admit forensic linguistic expert evidence, as the court does not have any criteria in place to assess the reliability of the expert evidence in Turkey. The proper criteria will help the judge to assess trustable decisions by depending on scientifically trusted methods rather than doing so intuitively. In this case, Turkish courts need a standard guideline to admit the expert evidence rather than expert qualifications. This guideline can help the court to determine the reliability and validity of the evidence presented by the expert in many fields, particularly in forensic linguistics. Moreover, having a detailed guideline for expert evidence will ensure that justice will be served without overshadowing the expert's intuition.

After starting with such initial concepts regarding the reliability of forensic linguistics in legal contexts on the mind, the suitability of the authorship attribution approaches is discussed in the literature review chapter after the methods are presented. In the following section, the motivation of the study is outlined.

1.3. The Motivation of the Study

The information age is a remarkable period in human history that has shifted human life onto a new path. It has created convenience and the sharing and receiving of information between users. This significant development has not only facilitated this immense exchange of data online but also opened a new area for illegal activities, as mentioned earlier in Section 1.1.

There is an opportunity and an environment for these activities to spread via emails, websites, internet chat rooms, forum pages, short text messages or social networking websites.

In 2009, Professor Tim Grant was consulted as an external expert on a domestic murder case in which Amanda Birks had died due to a house fire. The issue related to whether her husband had sent several text messages from Amanda's mobile phone before the fire started. After forensic analysis of the text messages, Professor Grant concluded that Amanda's normal messages were different from those on the night in question. Although Amanda's style was significantly different from that of the disputed texts, her husband's was not. Based on this evidence, on 2 November 2009, Amanda's husband was jailed for murdering her. This case is just one example of computer-mediated communication data playing a vital role in forensic linguistics in real casework.

In light of these new overwhelming developments in the information age and the increased risk of crime and fraud that web users are likely to face, there is a need for a sophisticated forensic linguistics framework. This framework should set out rules to bring about a sufficient, valid, reliable response to ensure that justice is done where necessary.

There have been many attempts to create a framework for authorship in linguistic research (e.g. Chaski, 2001; McMenamin, 2002; Grant, 2008; MacLeod and Grant, 2012; Grant, 2013; Wright, 2013) and computer science disciplines (e.g. Argamon, 2005; Koppel et al., 2009). However, many of these studies are dominated by English language problems in the field, although others were conducted in Spanish (Turell, 2010) and Portuguese (Sousa Silva et al., 2011). To the best of the researcher's knowledge, there is no study in the forensic linguistics context in the Turkish language apart from computer-based studies (e.g. Diri and Amasyali, 2003; Bozkurt et al., 2007; Türkoğlu et al., 2007; Kucukyilmaz et al., 2008). Therefore, there is a research gap in the Turkish language analysis which focuses on authorship attribution from the forensic linguistics point of view. However, such a study should depend on the reliable and valid methods to meet Daubert-like criteria in the courtrooms. For that reason, this study aims to apply an authorship attribution approach which is reliable and valid in interpreting the results.

Although the motivation of the study is briefly discussed in this section, the main gaps in the literature regarding Turkish authorship studies are presented in Chapter 2. Prior to that, there are some essential questions to ask in framing the research. In the following section, the research aims, and the research questions are presented.

1.4. Research Aims and Research Questions

As mentioned above, this research is motivated by the lack of authorship studies in the Turkish language in the forensic linguistics context and the current trend in online crime. The present study aims to conduct reliable and valid research in the field of authorship attribution in Turkish online texts.

First, the approach should depend on the reliable principles and methods which are applied by a specialised, knowledgeable person to understand the facts or data. Previously, some scholars from both forensic linguistics and computational studies worked on authorship attribution methods. However, it was found that some of them were not reliable or valid enough when considering the Daubert criteria (see Discussion in Chapter 2). However, Grant (2010; 2013) proposed a method which combines both approaches that had been successfully applied to short text messages. The aim here is to expand these techniques to improve the accuracy, reliability and validity of the authorship attribution problems in Turkish. With this aim, first, distinctive linguistic features are elicited, and then an attribution test is applied to the process. The result of this study has attributed the authorship based on their language production. Nevertheless, the characteristics of the collaborative online Turkish language encyclopaedia *Eksi Sozluk* data collected for the previous studies are different from short text messages regarding its nature. Testing this method in a different genre is the second aim of the study. Since any content may potentially be the forensic text for a forensic linguist if it is implicated in a criminal context, such as instant text messages, letters, tweets or emails, more research is needed on different genres. It will be decided whether such an approach can be extended to most online communication mediums because of the general nature of the features. Furthermore, as stated above, there are some other problems in authorship attribution studies, such as size and candidate authors.

Therefore, attributing authorship by looking at text sizes, analysing the performance of particular feature sets, performing tests by looking at the candidate author size and cross-genre analysis is added to widen the understanding of the field for Turkish. According to Luyckx and Daelemans (2008, p.2), current approaches focus on solving ‘an authorship attribution task given a small set of candidate authors and a large set of training data’. However, a genuine forensic linguistics case is unlikely to create this scenario. Similarly, Stamatatos (2009, p.22)

stated that authorship attribution methods need 'the examination of its performance under various conditions'.

Consequently, this study investigates whether the approach taken is applicable in the Turkish language in various situations by dividing datasets for different purposes in terms of the size that includes sub-parameters such as candidate author set size, text size and disputed text size, feature types and cross-genre text analysis to gain a further insight into the advantages and disadvantages of the approach which is applied in this study. However, there are hundreds of possible scenarios for authorship cases; these three parameters are just an initial framework for authorship attribution in Turkish. It is aimed to contribute to the exploration and improvement of authorship methods in Turkish to achieve those goals. Therefore, this research is focused on the following research questions:

1) What is the role of feature type in authorship attribution research in Turkish texts?
    i) Which feature set achieves the accurate attribution of authorship? Is it possible to increase authorship attribution performance by selecting the appropriate feature set or combinations in the current data set?
2) Does the text size and candidate author size affect the attribution of authorship correctly in Turkish texts?
    i) What is the effect of text size in authorship attribution?
    ii) What is the effect of candidate author set size in authorship attribution?
    iii) How many texts are needed in assigning the disputed texts to the correct author?
3) To what extent can the authorship attribution method be applied in cross-genre comparison?

Each question is explored in two approaches as treating each text individually and cumulatively. First, text vs. text comparison is identified the similarities between individual texts. Then, all available texts per author combined into a single document and extracted a cumulative representation of that author's style from combined text in comparing the disputed author with the other authors. This is called as author vs. author comparison. Author vs. author comparison produce one cumulative representation for all texts per author while the text vs. text comparison produce individual representation for each text. These two approaches are complementary to each other in establishing reliable authorship attribution results. The research questions are addressed in a separate section in Chapter 6. In the following section, an overview of the book is presented.

1.5. Book Overview

This book contains seven chapters. The first chapter has introduced the critical issues regarding forensic linguistics, recent trends in online crime, the reliability of forensic linguistics in courts along with the situation in Turkey, the motivation of the study, the research aims, and the research questions.

Chapter 2 of this study provides an overview of the previous approaches in forensic authorship attribution and computer-based methods, dividing them according to the three parameters which are mentioned in the previous section, namely size, feature type, and cross-genre comparison. Later, combined methods which are based on stylistic and stylometric approaches and the advantages and disadvantages of the methods are discussed along with the theory of idiolect. Following this, in Section 2, the language of the corpus, the features of the Turkish language, is presented.

Chapter 3 is dedicated to providing a brief description of computer-mediated communication and the internet language and introduces the corpus Eksi Sozluk and Twitter.

Chapter 4 details the aims of the research and the theoretical background of the adopted methodology. Later, the data collection, the corpus design, the criteria that are applied during the data collection, and the ethical considerations are outlined. Moreover, the statistical procedures and visualisation methods that are used in the analysis of Eksi Sozluk and Twitter are presented in this chapter.

Chapter 5 explains the feature selection approach in detail. A list of features which have been used in this authorship attribution study is created, and its reliability is tested among other coders.

Following this, Chapter 6 includes the ten tests which are based on the role of feature type, text size, candidate author set size, disputed text size, and cross-genre comparison. By using four different corpora with different text sizes, the analysis focuses on the accuracy in attributing Turkish texts to their correct authors. Extracts from the analyses and figures are presented to illustrate the results that correspond to each of the research aims.

Finally, Chapter 7 concludes the book by summarising the key findings and assessing the significant contributions it makes, primarily to studies in Turkish and other languages, theoretically and methodologically in the field of authorship attribution.

Furthermore, the research questions are revisited, and the limitations of the study and suggestions for future studies are outlined.

A reference list and appendices follow, which include the Nvivo projects of coding and a copy of the Ethical Approval form from Aston University.

Chapter 2: Literature Review—Forensic Authorship Attribution and the Turkish Language

The following sections present a discussion of the notion of the *idiolect* (Section 2.1), followed by a review of authorship studies based on stylistics (Section 2.2.1), and a description of previous studies using statistical approaches (Section 2.2.2). Next, a comparison of studies using stylistic and statistical methods is presented (Section 2.2.3), as well as studies that combine both methods (Section 2.2.4). Following this, sections are divided according to some of the parameters used in authorship studies, such as the role of feature types (Section 2.3) and the role of size (Section 2.4). Section 2.4 has three subsections, relating to text size (Section 2.4.1), candidate author size (Section 2.4.2), and limited numbers of texts per author (Section 2.4.3). Section 2.5 describes cross-genre comparisons in authorship attribution. In the final two sections, 2.6. and 2.7, previous authorship attribution studies in Turkish are introduced, and the features of the Turkish language and some general information about it and the summary of the addressing issues are given.

2.1. The Notion of the Idiolect

Bernard Bloch initially proposed the term *idiolect* in 1948 in his attempts to explain the relationship between speech and personality as discussed in Bloomfield (1933). Although Bloomfield (1933, p.45) did not use the term himself, he emphasised the following:

> The difficulty or impossibility of determining in each case exactly what people belong to the same speech community is not accidental but arises from the very nature of speech communities. If we observed closely enough, we should find that no two persons—or rather, perhaps, no one person at different times spoke exactly alike.

According to Bernard Bloch (1948, p.7), an idiolect is 'the totality of possible utterances of one speaker at one time in using a language to interact with one other speaker'. In the same vein, Barber (2004) describes a person's idiolect as 'a language that can be characterised exhaustively in terms of intrinsic properties of some single person at a time'. In contrast to Bloch (1948), some linguists have denied the existence of the idiolect; for instance, Barthes (1977, p.21) states that it 'would appear to be largely an illusion'. Similarly, Jakobson (1971) claims that it 'proves to be a somewhat perverse fiction' (cf. Barlow, 2010, p.1).

Since the notion of the idiolect is vital to explaining linguistic variation, some scholars have examined the concept and whether it is empirically possible to demonstrate its efficacy in forensic linguistics. For instance, Kredens (2002, p.2) suggests that forensic linguistics needs 'an empirically verified theory of idiolect, which could be used for both investigative and demonstrative purposes in cases where infringements of the law have been effectuated with or accompanied by, the use of language'. In a study carried out with 'the intention of establishing the degree of idiolectal variation' (ibid, p.2), he compared pairs of writers for whom factors such as age, gender, language, and topic were strictly controlled. He suggested that if there are people from the same demographical and sociocultural background, it allows us to find idiolectal clues thus it can also be stated that there are similarities between distant characteristics. Kredens concluded from his analysis that 'any method of authorship attribution is worthless unless it is capable of demonstrating that the similarities are idiolectal' (2002, p.25). Making a similar point to Kredens, Howald (2008, p.232) states the importance of 'demonstrating the existence of an idiolect and characterising what form it will ultimately take are the first threshold in determining the linguistic and scientific sufficiency of AuA [Authorship Attribution] techniques.'

Coulthard (2004) asserts that every speaker of a language has developed a considerable active vocabulary that differs from the vocabularies held by other speakers, and states the position of idiolect in forensic linguistics as follows:

> The linguist approaches the problem of questioned authorship from the theoretical position that every native speaker has their own distinct and individual version of the language they speak and write, their own idiolect, and […] this idiolect will manifest itself through distinctive and idiosyncratic choices in texts. (Coulthard, 2004, p.432).

In this context, the choice is described as a conscious process while habits are unconscious patterns (Tomblin, 2013). Following on from this, McMenamin notes that the role of style markers in the theory of idiolect is 'the observable result of the habitual and usually unconscious choices an author makes in the process of writing' (2010, p.488). Since writers differ in terms of their backgrounds, factors such as the places they have lived, the schools they have attended, their income, and many other variables affect their linguistic choices. These choices will be reflected in their use of style markers.

However, Olsson (2004, p.29) has argued that in order 'to measure unconscious style markers meaningfully, you need a great deal of text—such as a full-length novel, or hundreds of short

texts.' Similarly, Turell (2010, p.217) claims that 'rightly enough, idiolects can only be determined with countless amounts of data from each, something which never happens when dealing with real forensic linguistic data'.

Due to these methodological difficulties with the notion of the idiolect, Turell (2010, p.214) proposes the alternative term *idiolectal style* in order 'to extend its application to the study of less substantive, less overt and less discrete variables such as sequences of linguistic categories' in the context of authorship attribution. This is explained as follows:

> The concept 'idiolectal style,' following the use of the term 'style' in pragmatics, is proposed as a notion which could be more relevant to forensic authorship contexts. 'Idiolectal style' would have to do primarily, not with what system of language/dialect an individual has, but with a) how this system, shared by lots of people, is used in a distinctive way by a particular individual; b) the speaker/writer's production, which appears to be 'individual' and 'unique' (Coulthard, 2004); and also c) Halliday's (1989) proposal of 'options' and 'selections from these options'. (Turell, 2010, p.214).

Turell (2010) also suggests that idiolectal evidence may be useful for forensic text analysis when the task can be related to base rate knowledge. Base rate knowledge is 'a collection of data regarding the general usage of the linguistic parameters being considered by a relevant population, or group of language users from the same linguistic community, with which the specific behaviour of the speakers or writers under comparison can be compared' (Turell and Gavalda, 2013, p.499).

Accordingly, Grant (2010, p.509) has questioned the validity of the notion of the idiolect in authorship analysis and hypothesises that even though an idiolect is a distinct and individual version of a language, it only becomes useful when the idiolectal features are repeated either within one text or across several texts produced by the same author. Although he does not ultimately reject the notion of the idiolect, he makes claim that 'the idea that comparative authorship analysis rests upon a strong theoretical assertion of an idiolect is false' (Grant, 2013, p.473). Instead, he proposes the notions of 'consistency' and 'distinctiveness' which he believes are more helpful than idiolect.

He also claims that comparative authorship studies are based on two assumptions: the first is that 'there is a sufficient degree of consistency of style' (ibid) in the texts written by a given author; the second is that 'this consistency of style inherent in an author's writings is sufficiently distinctive to discriminate one author from other relevant authors' (ibid). In a similar manner, McMenamin (2010, p.490) suggests that the consistency model 'is used to

determine if various writings were written by the same author', and the distinctiveness model is referred as to 'looking for resemblance'. Olsson (2008, p.33) uses the terms *intra-author* and *inter-author* variation to describe how and why authors vary from each other.

Grant (2013) presents a method based on descriptive and statistical approaches that depend on linguistic features such as abbreviations, punctuation, and lexical choices. This has produced accurate results which demonstrate that different authors can be consistent and distinctive in their use of language; nevertheless, there is no requirement for absolute consistency, since language itself is not a stable entity. Such a 'method allows for and detects the fact that one author may be consistent in, for example, a form of abbreviation, whilst another author may tend to punctuate in an idiosyncratic manner' (Grant, 2010, p.521). Moreover, texts from the same genre may have a style that is consistent across the individual author and the disputed text. However, this consistency may not be observed entirely from the texts; in this case, the important thing is to 'take note of relative frequencies of alternative variants' (Coulthard et al., 2017 p.157).

Distinctiveness can be taken as a measure of the similarities and differences between two authors and can be divided into two types: (i) pairwise distinctiveness; (ii) population-level distinctiveness. These are defined as follows:

> If it can be demonstrated that the suspect exhibits a consistent style in text messaging and also that the victim has a consistent but different style then the first level of distinctiveness will have been proved. I shall refer to this as pair-wise distinctiveness. The second possible level of distinctiveness, however, may have more profound implications for theoretical discussions of idiolect. This would occur if one person's text messaging style can be said to be distinctive, unusual or even unique against a reference population of text messages. This I shall refer to as population-level distinctiveness. (Grant, 2010, p.515).

Grant (2013) also stresses the importance of selecting the correct population for use as the 'relevant comparison corpus' which is required to 'identify […] consistency within relevant texts' (p. 473). Population-level distinctiveness is related to Turell's (2010) concept of base rate knowledge. Leonard (2005) states that in order to find idiosyncratic patterns related to idiolect, it is necessary to make a base rate compared with a reference dataset.

Coulthard et al. (2017) ask 'how does one establish what is a relevant comparison population of speakers of texts'? They continue that for linguistic studies 'there is much [less] reference data, although specialist corpora are being created'.

Wright (2014; 2017) has attempted to compile relevant reference data for authorship attribution by analysing the Enron Corpus. This dataset consists of 63,369 emails written between 1998 and 2002 by a group of authors from the same linguistic community, and it provides base rate knowledge against which population-level distinctiveness can be assessed. Wright analysed it with the purpose of exploring idiolect by focusing on n-grams in a manner that was different from the methods employed in previous studies on the same corpus (e.g. Iqbal, et al., 2010; Chen et al., 2011). Firstly, distinctive ways of using the first-person pronoun *I* were examined in order to discover 'author-distinctive variation in the use of an extremely common word' (ibid: p. 154); secondly, a number of distinctive collocations were identified for the word *deal*; and finally, please-mitigated directives were examined in order to find idiolectal variation. The conclusion was that in the mails of Enron Corpus, 'production[…] of these collocations and lexical sequences are identifiable realisations of [the authors'] idiolects' (Wright, 2014, p.156).

Although idiolect is a useful theory in authorship analysis it is lack of em

In this section, various views on the existence of the idiolect have been discussed, and it has been stated that there are two main assumptions underlying the concept. Alternative terms used to report variation between authors have also been presented. In conclusion, there is still very little empirical study into the existence of the idiolect, but Wright (2014) is one of the first to present evidence that particular linguistic features may be idiolectal. Next, approaches to authorship attribution will be presented.

2.2. Authorship Attribution Approaches

For many years, authorship attribution has attracted the interest of scientists wishing to create a standard method for identifying authors. First, statisticians tried to solve the problem of authorship in written discourse. Augustus de Morgan was one of the first to express an interest in the problem in 1851, comparing texts by using the average number of letters per word (Olsson, 2004). Later, Mendenhall attempted to classify authorship according to the frequency distributions of the words in the plays of Shakespeare (Mendenhall, 1887). Since then, numerous studies have applied different approaches to identifying authorship, from computational methods to using descriptive linguistic analyses to qualify the disputed author's style. Although authorship attribution studies date back to the 19th century, the first known forensic linguistics analysis was not provided until 1968, with Jan Svartvik's research into 'The Evans Statements'.

Nowadays, there are many statistical techniques employed with the aim of analysing authorship. These include a combination of linguistic, descriptive, and statistical methods (MacLeod and Grant, 2012), which can be divided into two main approaches: those which are based purely on stylistic analysis, and *stylometric* methods which depend on reliable statistical procedures. In the stylistic approach, 'the analyst uses his or her expertise to reach a conclusion' (Rico-Sulayes, 2011, p.54), while the stylometric analyst 'appl[ies] classification techniques common in scientific research to all instances of the features'. This produces a measurement of the similarity and dissimilarity between texts and authors with the aim of providing statistically significant results (ibid).

Previous authorship attribution studies have focused on some of the following questions:

- How many and what features are used for attributing authorship?
- Which methods are employed to process those features?
- Which methods are used for authorship attribution?
- Is the unknown author part of a given set or not?
- Does the text size have an effect on the analysis?
- Does the genre have an effect on authorship attribution?
- Does the topic have an effect on authorship attribution?

That is to say, the field of authorship attribution generally stresses the roles of feature types, methods, candidate authors, text size, genre, and topic. In the following sections, the two main approaches will be discussed.

2.2.1. Stylistic Approaches

Style can be defined as 'features of the text that indicate the author's choice of one mode of expression from among a set of equivalents modes for a given content' (Argamon and Koppel, 2013, p.299). The term *stylistics* is defined by McMenamin (2002) as the 'application of the science of linguistic stylistics in forensic contexts'. He continues:

> Forensic stylistics makes use of stylistic analysis to reach a conclusion, and opinion related to the authorship of a questioned writing within the context of litigation. Stylistics is the scientific study of patterns of variation in written language. The object of study is the language of a single individual, resulting in a description of his or her identifying linguistic characteristics. (p.163).

When doing stylistic analysis, it is important to bear in mind that: (i) qualitative description is a first-order problem—measurement depends on meaningful discovery, description, and categorisation of relevant linguistic features within a text; (ii) qualitative evidence is more demonstrable in the courtroom than quantitative evidence; and (iii) qualitative results appeal to the nonmathematical but structured sense of probability held by judges and juries (McMenamin, 2002, p.129). McMenamin (2010, p.163) mentions that stylistics takes advantage of two main ideas: 'no two writers of a language write in the same way, and no individual writer writes the same way all the time'.

Nini and Grant (2013) have classified previous stylistic methodologies into four types: (i) studies which claim that 'the theory of idiolect is important but not necessary'; (ii) those based on the idea that style results in identifiable individual uniqueness; (iii) those taking the position that 'the style of an individual can be found in language through the analysis of the array of style markers'; and (iv) those for which 'the repertoire of style markers [is] generated by the variability inherent in […] historical-sociological difference[s]' (p. 174). That is to say, the methodology of forensic stylistics mainly deals with *style markers* which indicate that 'each human being has a different repertoire of linguistic variables and that these variables manifest themselves in their writings' (Nini and Grant, 2013, pp.174-175).

Style markers can be observed, described, and analysed in the language of groups and individuals, and they are the result of the habitual and unconscious linguistic choices of an author (McMenamin, 2010, p.488). According to this view, there are different ways to organise a sentence, either from a choice of optional forms or as a deviation from the norm. Deviations from the norm are common among careless and under-educated writers, for instance, mixing homonyms such as *its/it's* (ibid, p. 489), and choosing alternative forms or usage variations, such as *I give you my heart/I give my heart to you/I give to you my heart.* While some authors tend to use standard structures, others may have certain preferences that make them distinctive.

In order to compare known texts with unknowns, McMenamin (2002) has produced a summary of style markers drawn from 80 cases which have played a role in identification. These include text format, the use of numbers and symbols, abbreviation, punctuation, capitalisation, spelling, word formation, syntax, discourse, errors, and high-occurrence words. Alison Johnson (1997) examined three student essays which were suspected of plagiarism; she focused on shared lexical vocabulary as a style marker and found that the suspect texts had a vast degree of similarity in lexical vocabulary but a very low incidence of novelty.

Former FBI detective James R. Fitzgerald used lexical features to analyse the Unabomber Manifesto. This 35000-word manifest was published in the US national press in 1995 as part of a deal made with the police that the writer would halt a bombing campaign against university and airline workers that he had been carrying out since 1978. After its publication, the police were contacted by somebody who claimed that he was the brother of the writer, saying that his suspicions had been aroused by one lexical string in the manifest – *cool headed-logicians*. From this starting point, a 300-word mail was found that originated from the suspect, and when the two texts were compared, a series of common lexical and grammatical words and fixed phrases were found. These included items such as *at any rate, clearly, gotten, in practice, moreover, more or less, on the other hand, presumably, propaganda, thereabouts* (Coulthard et al., 2017).

In his work on the Jenny Nicholl case, Coulthard created a set of small corpora from text messages sent by Jenny Nicholls and her ex-boyfriend with the aim of finding similar features among them. According to the results, certain abbreviations were used in Jenny's messages, e.g. *'im', 'm not', 'ive' 'cu'* and *'fone',* that were not found in the other corpora, suggesting that these were intra author differences (Kredens and Coulthard, 2012).

Coulthard (2013) has also recently worked on another case relating to a disputed email which was allegedly sent by Mr Stephen Goggin. There were four people who could possibly have been the author: Mr Goggin himself; Mr Tim Widdowson, the CEO; Mr John Shuy, the Finance Director of MaxiSoft; and possibly their PA, Ms Janet Gavalda. Coulthard accessed 19000 texts, including company emails and meeting reports, and found that a number of collocations from the disputed email also occurred in other emails, such as the combination of *employee* with *disgruntled* and with *former.* He then used the Google search engine to find the frequency of these collocations and lexical strings in a more extensive database and concluded that 'significant lexical choices in the questioned email are consistent with choices Widdowson makes elsewhere', and also that 'these selections do not occur in emails sent by anyone else and so are distinctive' (Coulthard 2013 p. 458).

As demonstrated in the cases presented in here, the stylistic approach is based on identifying distinctive style markers and depends upon the linguistic expertise of the analyst. However, some researchers consider the stylistic methodology to be non-empirical since it 'rests on erroneous assumptions about individuality in linguistic performance and violates theoretical principles of modern linguistics' (Chaski, 2001, p.3). In the following section, alternative stylometric approaches are presented.

2.2.2 Stylometric Approaches

Stylometric approaches are ‘exemplified by scholars across the field seek[ing] to find or describe quantifiable markers of authorship, which in the general sense vary more between authors than within authors’ (Grant, 2013, p.470). Stylometry is defined as a ‘wide range of methodological approaches to authorship analysis in which the similarit[ies] and difference[s] between […] styles are statistically measured by their use of a particular set of linguistic features’ (Coulthard et al., 2017, p.153). Early work on stylometric authorship attribution focused on the average length of words, sentences, or syllables. Yule (1938) used a method based on sentence length and noun frequencies. At the same time, Zipf (1932) used word frequencies. Almost forty years later, Mosteller and Wallace (1964) produced the first ever large-scale study attempting to identify the linguistic style, which is now accepted as the ‘the most influential work in authorship attribution’ (Stamatatos, 2008, p.1). This study on the authorship of ‘The Federalist Papers’, used the frequency of function words with great success. From then on, the use of measurements of linguistic style in authorship attribution became known as stylometry. Most stylometric approaches involve the use of statistics, such as the distribution of sentence length, word frequencies, vocabulary richness, paragraph length, function words, or character n-gram features, and they achieve good performance in accuracy. It has been claimed that the distribution of function words and syntactic features are not under control of the author (Forsyth and Holmes, 1995). In general terms, stylometric authorship studies involve two steps: identification of discriminative style markers (mostly frequency-based features), followed by the application of a suitable statistical algorithm to determine the most likely authorship.

Following this brief introduction to stylistic and stylometric approaches, the following section will now review studies which have compared and combined these two approaches.

Following section, the structure of Turkish language will be explained.

2.2.3. Stylometric Approaches versus Stylistic Approaches

Authorship attribution in the forensic setting must satisfy four criteria in order for it to be admitted as scientific evidence in the courtroom (Chaski, 2005). First, 'the method must be linguistically defensible', second, 'the method must be forensically feasible' in other words, a forensically feasible method is one that is sensitive to the actual limitations of real data, third 'the method must be statistically testable' and fourth, 'the method must be reliable, [and] based on statistical testing' (Chaski, 2005 p.46–47).

When an authorship method corresponds to these criteria, it will also be considered as meeting the Daubert criteria. Cheng reported that (2013, p. 550) 'unlike most forensic fields, which arose long before the invention of DNA typing and the decision in Daubert, forensic linguistics will blossom within a modern scientific evidence framework'. Stylistic approaches meet the linguistically defensible and forensically feasible criterion easily, while stylometric approaches disregard linguistic theories and the limitations of real-world data. For instance, stylometric approaches require hundreds of features and texts, which is not easily found in real cases. However, stylistic approaches fail in the statistical testing criterion, while stylometric approaches only depend on statistics.

It is claimed that forensic linguistics, whose methods mostly rely on stylistic markers, 'conduct little or no laboratory work', while computational linguists use different algorithms with successful results in predicting authorship; however, this is not accepted in the forensic application (Solan, 2013, p.557).

Similarly, Chaski (2005, p.2) stated that ' without the databases to ground the significance of stylistic features, the examiner's intuition about the significance of a stylistic feature can lead to methodological subjectivity and bias'. Furthermore, Nini and Grant (2013, p.176) highlighted the bias in stylistic approaches in that it is difficult to 'replicate the analysis and therefore to claim objectivity and universality'. Even though using computational methods are necessary for a reliable authorship analysis, without a linguistic theory, they are open to criticism in terms of reliability and objectivity.

However, 'in many cases stylistic reasoning can be perfectly persuasive without any resort to statistical processing, while when viewing the matter, the other way around, numbers and ratios can never be fully persuasive when we have no understanding of what elements in the language of the text have given rise to them' (Love, 2002 p.101).

Ian Stewart makes the following comparison between qualitative and quantitative evidence:

> There are many circumstances in which qualitative information is what really counts, and quantitative measurements are a rather poor route toward finding that information. For example, the important question about a bridge is, 'will it fall down?' and calculating its precise numerical breaking strain on a supercomputer is just an exceedingly complicated way to get a yes/no answer – and perhaps an unnecessarily complicated one. (cf. Love, 2002, p.211).

Rather than applying complicated methods, McMenamin followed a stylistic methodology (2002, p.129) and stated that 'in the courtroom, qualitative evidence is more demonstrable than quantitative evidence because it is the language that is presented' and he went on to claim that 'qualitative results appeal to the non-mathematical but structured sense of probability held by judges and juries'. This is in contrast to Chaski's (2001, p.3) claim that a stylistic approach 'rests on erroneous assumptions about individuality in linguistic performance and violates theoretical principles of modern linguistics'. The 'forensic practitioners working on shorter and sometimes fragmentary texts have tended to use more stylistic approaches' (Grant, 2013, p.471) and focus on individual variation, which is 'understood as being created by habitual choice across a wide and unpredictable range of features' (ibid).

Furthermore, Coulthard (2013, p.42) commented on stylistics in that 'in cases where conclusions depend on observations about the frequency or rarity of particular linguistic features in the texts under examination; many linguistics would have considerable difficulty in stating a known rate of error for their results'. According to Coulthard et al. (2017), three things need to be applied in order to meet the Daubert criteria: (i) increasingly more extensive databases are needed in order to derive a population statistics; (ii) there should be an expert proficiency test; and (iii) the linguist should express opinions statistically, as reported by Grant (2010; 2013).

The main weakness of the statistical approach is that it presents no linguistic explanations for the analysis results. Stylometric studies focus on thousands of features without considering the relevance of them to the particular case (Solan, 2013). Furthermore, short texts are not suitable for stylometric feature selection, such as the bag-of-words technique. Thus, stylometric techniques are used in many types of research, but they are 'rarely applied in forensic casework' (Coulthard et al., 2017, p.154).

Rudman (1998, p.355) listed the generic problems in authorship attribution studies. The validity of statistics-based studies and research papers that place too much importance on

statistical techniques are criticised as 'they try to create an aura of scientific invincibility without scientific rigour'.

In a similar manner, Juola (2008) discussed the position of computer programmes in authorship attribution:

> Even when the techniques themselves are accurate, they still may not be adequate for a research study. 'Because the computer said so' is not really a satisfactory explanation for many purposes (this is also one of the major weaknesses with artificial neural networks as decision makers, and one of the reasons that expert systems are usually required to provide explanations of the reasons underlying their decisions). A jury may be reluctant to decide based only on 'the computer,' while a researcher interested in gender differences in writing may be more interested in the reasons for the differences instead of the differences themselves. (p.247)

There are many computer-based authorship attribution studies. However, none of them has yet made a connection with the forensic application, although they have got successful results in attributing authorship (Solan, 2013). Cheng (2013) argued that the reliability of computational procedures in forensic linguistics favours theory-based linguistics studies as 'the linguists, although perhaps not the computational ones, would feel more comfortable if the methods and results were better rooted in linguistic theory' (p. 545). Grant (2008, p.226) stated the difference between stylometric researchers and forensic linguistics as follows:

> It might be reasonable to sacrifice the linguistic validity in a rush to discover an authorship algorithm, but in the forensic field the analyst must be able to say why the features they describe might distinguish between two authors in general, and why they distinguish between the particular authors of the case.

Stylometric approaches aim to achieve higher success in results rather than finding the linguistically important features and style markers. No matter how accurate the results are, the researcher is not able to explain why the pair of texts show significantly different characteristics (Kotzé, 2010).

Unlike previous stylometric studies, Argamon and Koppel (2013) considered the theoretical explanation behind stylistic features along with their accurate classification using automated methods. Furthermore, Argamon and Koppel (2013, p.300) stated that:

> …without a firm basis in a linguistic theory of meaning (not just of syntax), we are unlikely to gain any true insight into the nature of any stylistic distinction being studied. Such understanding is key to both establishing and explaining evidence for a proposed attribution. Otherwise an attribution method is merely a black box that may appear to work for extrinsic or accidental reasons but not actually give reliable

> results. Furthermore, an attribution method that produces insight into the relevant language variation is more likely to be useful and accepted in a forensic context, all else being equal, as the judge and jury will be better able to understand the results.

Such an interpretation is crucial in the case where authorship attribution technology is used as evidence in a judicial process. Furthermore, Stamatatos (2013) indicated that with regard to authorship attribution studies with fewer statistics 'that [it] is necessary to use this technology as evidence in court is the ability to explain the automatically derived decisions' and the high level of automated approaches 'is needed a way to associate… dimensional information to some human interpretable high-level features' (p.439).

In the same way, Luyckx (2010, p.7) also used a qualitative feature analysis in contrast to previous studies in stylometrics and stated the importance of the qualitative method in that 'by providing a qualitative analysis of features, we can evaluate whether the individual features are scalable or not', and she considered it 'crucial to gain insight into the attribution model in order to increase our understanding of the effect of experimental design'.

However, there is no answer to explain which approach is the best in the field of authorship attribution (Solan, 2013). Moreover, in order to meet the Daubert criteria, the authorship attribution method may be a combination of several features and techniques from several different proposed methods (Juola, 2007). The field is going to get the advantage of both approaches and corpus linguistics (Coulthard et al., 2017). In the following section, combined stylometric and stylistics methods are presented.

2.2.4. Combining Stylometric and Stylistic Approaches

An ideal forensic authorship analysis method is described as 'a method in which the measurement of similarity or difference between authors' styles or texts is represented statistically, but explainable in linguistic terms and also, importantly, has a known error rate' (Coulthard et al., 2017, p.207). Since stylistic analysis results are based on verbal expressions instead of calculations, although there are disadvantages in estimating the authorship with such methods, expert reports provide detailed and transparent explanations in the courts (e.g. McMenamin, 2002). However, in some countries, the legal system determines the admissibility of evidence in the courts (e.g. USA – Daubert criteria) in order to satisfy the scientific validity of evidence. As a result, in combined authorship studies, the style markers that are linguistically

meaningful are uncovered, and a suitable statistical method is applied to distinguish identical authors. In other words, stylistic approaches 'do justify why their markers distinguish authors' (Nini and Grant, 2013, p.176), but it is far from the notions of 'objectivity and universality'. Juola (2008) stated that it is possible for people who are linguistics and statistics should do productive research by combining their expertise. That is to say, an ideal authorship attribution system should be theoretically well supported, and the features should have an explanation in the field of linguistics, psychology and neurology in terms of their efficiency (Juola, 2008, p.251). The combined approaches in authorship attribution are needed to produce 'reliable quantitative results with clear linguistic underpinning' (Wright, 2014, p.23).

Setting up a reliable method for forensic linguistics studies has crucial importance as is mentioned by Solan (2013):

> For those who rely upon judgements of authorship based on their knowledge of linguistic features and upon a sense that a large cluster of difference or commonalities in a particular case cannot be a matter of accident, research into methodology should be a top priority. (p.555).

That being said, one-case only personal judgement or skill dependent methods are not the best for the field. At the same time, intuitive expertise is not stated as undoubtedly unreliable in some settings, on the contrary, 'stylistic analysis is not provably less reliable than the quantitative methods' (ibid: 572). Solan (2013) indicated the importance of combined studies among scholars from different areas of expertise in forensic linguistics. In a similar manner, Juola (2008) remarked that the authorship attribution analysis report should be understandable by the jury and the general public. Instead of using entirely computer depending methods, a satisfactory explanation is needed in order to present the results using an appropriate graphics and visualisation method (ibid).

In light of this information, Grant (2013) developed a 'general statistical method for forensic analysis of text messages', which combines both descriptive and statistical approaches in attributing authorship (p. 470). The murder of Amanda Birks in Stroke-on-Trent, UK was the case for his analysis (see Chapter 1).

Grant used a corpus of 204 texts messages from Amanda Birks' mobile phone and a set of 203 texts from Christopher Birks' (victim's husband) phone. First of all, WordSmith Tools (Scott, 2012) listed all the variants including non-standard spellings, abbreviations etc. from 407 messages. 154 linguistic features created for the analysis, next the features that occurred less

than ten times were extracted from the lists. Subsequently, a final list of 18 features remained for the statistical part of the analysis. For the statistical analysis, every feature was coded as either present or absent in every text message by both authors. A binary similarity test, Jaccard's similarity coefficient, was used to calculate the similarities between authors. It is used to establish the consistency and distinctiveness of the features by looking at the presence or existence of them, rather than the frequency rate in short text messages. In previous cases, Jaccard has been accepted as a reliable analysis method by the UK (e.g. Grant 2013) and US courts (e.g. Juola, 2013).

Furthermore, another statistical test called the Mann Whitney U test was used, and it found that Amanda's last messages were inconsistent with her previous messages. However, Christopher's style was similar to other messages she had sent. Based on the results, Christopher Birks was given a life sentence in November 2009. Moreover, Grant (2013) proposed a reliable and valid protocol for the authorship analysis framework, which is a 'methodologically rigorous approach to stylistic authorship analysis that can result in statistically described results' (ibid p.472).

Turell (2010) took a combination of approaches and methods for the study, and with this objective, a discriminant function analysis was conducted to compare the four non-disputed faxes and disputed emails while using qualitative textual analysis, and it was concluded that qualitative analysis is necessary for eliciting linguistic markers that cannot be found by automatic procedures. Subsequently, Turell and Rosso (2013) used both computation and forensic linguistics in forensic plagiarism detection and authorship attribution and concluded that both approaches are advantageous. Furthermore, Queralt Estevez and Turell (2013) presented a semi-automated method for attributing authorship using both qualitative and quantitative approaches in two cases. The most frequent tag sequences are morpho-syntactically tagged bigrams and trigrams. The study used only two candidate authors and three control authors; in total, five authors with texts of a similar length in both cases since they were real forensic cases. The results of the bigrams and trigrams showed high probability in attributing the correct author in both cases.

Later, Kotzé, (2010) reported two-real cases from South Africa that were analysed by combining stylometric and stylistic approaches. Firstly, in the 'The Father Punch' case, a set of memoranda including 5482 words from an accused person and a set of eleven chronicles including 25431 words that were suspected to have been written by the same author were collected. For the analysis, wordlists, concordances and keywords were found by WordSmith

Tools. It was found that the keyness value of grammatical keywords is more than 15 values when comparing the core chronicles and memoranda. As a result of the stylistic analysis, a total of six authorship markers were identified. The results of the analysis were presented to the Supreme Court including the qualitative analysis and graphical representations of the statistical findings. Eventually, the court accepted the evidence and sentenced the accused (Kotzé, 2010, p.191). Secondly, in the 'The Angry Academic' case, three sets of anonymous documents, which were circulated on the campus at a South African university, were investigated. For the analysis, 14 documents were analysed using the same method as was used in the previous case and 'an average chi-square value of 10 for the keywords [was] identified' (Kotzé, 2010, p.192). In contrast, the stylistic analysis did not provide particular patterns that showed whether the letters had been written by more than one person or not. The University Council also accepted these results. Kotzé (2010) stated that combining both approaches in authorship attribution studies 'must contribute to the principle of scientific rigour underlying the credibility and acceptance of expert evidence submitted by linguists in court' (p.194–195).

Furthermore, in 2012, MacLeod and Grant tested the applicability of Grant's (2010) stylistic and statistical approaches, which were originally applied to short text messages, on micro-blogs. The data for the research was collected from the microblogging website 'Twitter' and demonstrated a feature list that was extracted from previous studies by Smith et al. (2009) in the area of authorship attribution of SMS texts. 19 features were used along with a detailed list from the CMC literature. A set of 18500 tweets were analysed on WordSmith Tools to extract the data and analysed the different linguistic features. First, the performance of single messages was tested, and the results showed reasonable accuracy and discrimination for a single message. Later the performance for aggregated messages was used, and the results showed an improvement in performance after aggregation. Finally, the results of the performance for multiple authors were presented, and it showed a reasonable success rate in identifying authors from a set of 20 authors, who were defined as a large candidate set. Eventually, this study demonstrated positive results, although the data size was considerably small.

Nini and Grant (2013) collected data from three second-year undergraduate students from a UK university by controlling their demographical backgrounds in terms of its closeness. Each student wrote three academic papers and only the first 300 words were chosen, and the variables were coded depending on the systemic functional linguistics and multidimensional analysis.

Larner (2014) collected 100 texts written by 20 authors over a five-day period; this was a total of 65113 words with each author producing an average of 3325 words. The shortest text was

485 words, while the longest was 822 words. Unlike previous studies, six types of formulaic language categories were used as features, and a total of 604 formulaic sequence tokens were extracted. The 26 most frequent features were selected, which occurred at least five times in the corpus. When 190 author pairs were tested, 38 were the correct attribution, while 28 were incorrect. In 125 pairs, no attribution was successful, and in 107 cases there was no significant difference between the candidate authors. The results demonstrated a low success rate and no difference in authorship between individuals.

Furthermore, word n-grams have been chosen as reliable features for such studies since the absence or presence of a word string in the text is rare, thus when it is compared with other authors, it may reveal the *consistency within* the author and *distinctiveness between* authors. Johnson and Wright (2014) used n-grams of between two and six words long as features, and a software (Jangle by Woolls 2013) developed with the purpose of measuring the Jaccard coefficient. Jangle software 'provides the analyst with a list of all the word n-grams that are shared by the disputed and the known email sets, that is, those which account for the similarity statistic' (Coulthard et al., 2017, p.208). Johnson and Wright's (2014) analysis found that five-words strings have 90 per cent accuracy with an error rate of ten for texts between 753 and 1018 words in length for a specific author used as a case study. However, when the size of the messages was reduced to 55–145 words, the accuracy rate decreased by 30 per cent with an error rate of 70 per cent. These findings are explained by Coulthard et al. (2017, p.209) in terms of the error rate as 'in an experimental context [these results] satisfy the error rate criterion in Daubert', namely, if 5-word grams for the long texts [between 753 and 1018 words] have a zero-error rate, it leads to an understanding of how good or how bad the results are depending on the performance.

Although the reliable percentage of error rate according to the Daubert criteria is argued in many studies (e.g. Haug and Baird, 2011; Jabbar, 2010; Kaye et al., 2000), there is still no agreement on it.

Although there are some studies which state that a 70-percent error rate is acceptable, (Coulthard et al. 2017) from a linguistic analysis viewpoint, a known error rate is still a questionable issue and according to Castelle and Shuy (n.d.) 'the absence of a known or potential rate of error should not have been an issue' since it is presented in 'an observable fashion' and is 'verifiable by all parties'.

Overall, the notion of idiolect has changed and developed due to further studies on idiolect over the years. In recent years, some scholars (e.g. Grant, 2013) described similar linguistic patterns with consistency and distinctiveness rather than idiolect. Although some authorship attribution studies were based on idiolect, some used consistency, distinctiveness, the resemblance of 'Base Rate Knowledge' terms. These studies were divided into two approaches: stylistic and stylometric. On the one hand, stylistic approaches identified the style markers based on the linguistic theories while stylometric approaches explored the statistical reliable methods. However, in recent years, both approaches have used as complementary to each other. Therefore, this study is aimed to develop a combined method which uses linguistic features as style markers and a reliable statistical test.

Following sections introduce the role of feature types, sizes and cross-genre comparisons in previous authorship attribution studies.

2.3. The Role of Feature Types in Authorship Attribution Approaches

Stamatatos (2009) categorised five different features as lexical, character, syntactic, semantic and application-specific features in identifying unknown text documents. In this section, the studies which have reported the success of different feature types are presented. Nevertheless, it is worth noting that as there is no clear division between studies which are based on only features, such as candidate author size, text size or cross-genre comparison, most of the studies in this section employed a method that depended on various text sizes, features and candidate authors.

Since the first authorship attribution studies (e.g. Mosteller and Wallace, 1964) were conducted, choosing the suitable and most discriminative feature type has been one of the main issues. Rudman (1998, p.358) stated that almost 1,000 writing style features had been used in previous authorship analysis studies. He stated that 'style is a complex package consisting of a theoretically unique combination of thousands of individual traits - a very large but finite number'. However, there are no generally accepted good features for authorship attribution (Rudman, 1998) because of the varying conditions in different settings.

Baayen et al. (1996) used syntactic features consisting of parts of speech and n-grams when working with sections of two English crime novels with around 20,000 words each and claimed that they are reliable features to distinguish the authors.

In the same vein, Chaski (2001) suggested that syntactically classified punctuation based on its role in the sentence had better performance than simple punctuation marks and stated:

> The simple punctuation-frequency technique is able to differentiate between different writers most, 92.8 per cent, of the time, but fails to cluster the questioned document with the actual writer. The syntactically classified punctuation technique is able to differentiate between different writers most, 86 per cent, of the time and it also can cluster the documents of one writer. (Chaski, 2001, p17.)

This study has been criticised by some researchers (Grant and Baker, 2001; McMenamin, 2001) in terms of its methodological weaknesses. They questioned the reliability of the results regarding the unbalanced success rate, backgrounds of the candidate authors (socio-linguistically similar authors), linguistic theory, and whether word length is enough to use as a valid marker since it has no theoretical understanding.

Similar to Chaski (2001), Rico-Sulayes (2011) used syntactically classified punctuation marks as features. However, the method adopted was based on quantitative, statistical and machine learning techniques. For the study, six syntactically classified punctuation-related features, 16 structural media-specific features, 13 emo texts (the collection of textual conventions which is to convey emotional content in the online communication) and two structural document- level pieces of information were selected to perform the task. Of a total of 141 users with a minimum of 2,000 words, 106 had at least 40 posts, ten of which were randomly chosen for the purpose of the study. As comparison material, a 500-word text was used, and the probability that this sample belongs to the author was found to be 93 per cent by using the combination of the features (Rico-Sulayes, 2011, p. 71).

Stamatatos et al. (2000; 2001) analysed 300 Greek newspaper articles with an average of 1,100 words each using the fully automatic method to extract POS (Part of Speech) features and achieved better performance than lexical-based approaches. Likewise, syntactic features including the three most frequent POS tags, suffixes of a word is adapted to a study and had high classification performance (van Halteren, 2007). Silva et al. (2010) investigated authorship analysis in Portuguese texts at the sentence level and focused on six potential and observable markers of authorship of stylistic features. For the study, POS-based features, the frequency of punctuation, and quantitative features such as the number of words per sentence, suffixes, superlatives, diminutives, pronouns and conjunctions were listed as content agnostic features (Silva et al., 2010, p.52). In total, 915 texts from 23 columnists from one of the Portuguese daily newspapers were selected. Since POS-based features carry more information, they performed as well as the combination of all the features. Furthermore, punctuation and

length worked well alone. Therefore, the study concluded that content agnostic features are effective in authorship analysis at the sentence level.

Moreover, punctuation marks are useful features to identify the author and have been tested in many languages and with different approaches (e.g. Portugese, Silva et al., 2011). Mingzhe and Minghu (2012) focused on punctuation marks for Chinese and achieved 96–99 per cent success using Ward's hierarchical clustering method. Furthermore, punctuation has been taken as a feature type (McMenamin, 2002) in stylistic studies regardless of their frequency of in the sentence unlike in other studies (e.g. Chaski, 2001; 2005; 2007).

Instead of focusing on only one type of feature, Grant and Baker (2001) used principal component analysis, which identifies the most efficient features or the combination of them in attributing authorship. This method diagnoses the most useful markers and their combinations in discriminating authorship (Grant and Baker, 2001).

> Even though 'in PCA strange combinations of markers, such as the distribution of specific letters and average word length might prove to be effective discriminatory components' due to the lack of validity, 'both further empirical research and a theoretical explanation of how the discriminator may be working would help construct some validity for the marker'. (Grant and Baker, 2001, p.77)

Baayen et al. (2002) employed another approach which has better results than principal component analysis. In their method, they used punctuation marks, function words and content words in discriminant analysis. The analysis was carried out in Dutch. Eight students who were native speakers of Dutch were asked to write three texts with around 1,000 words in three different genres on various topics. At the end of the data collection, there were 72 texts with an average of 908 words. The experiment achieved 88.1 per cent success when the frequency of punctuation was added along with 50 function words. Furthermore, the results of the study show a connection between the authorial structure in written texts and the authors' backgrounds. Similar to Baayen et al. (2002), Zheng et al. (2003) used 122 function words in comparing English and Chinese email messages and forum posts in authorship attribution. Furthermore, Argamon and Levitan (2005) measured the usefulness of function words in authorship attribution using 20 samples of 10,000 words and found that the most frequent words in the corpus give the highest discrimination measure.

Burrows (2002) stated that computational stylistics uses multivariate statistical comparisons between some features which comprise the relative frequencies of various simple phenomena, including alphabetic characters, strings of characters, whole words or common grammatical

forms. The positive thing about working with whole words depends on 'their accessibility and their meaningfulness' (Burrows, 2002, p.268). Burrows also compared the weak and strong features in terms of their discrimination power. Weak ones 'offer more tenable results than a smaller number of strong ones', while strong features are 'easily recognised and modified by an author and just as easily adopted by disciples and imitators'. He concluded by stating that 'at all events, a distinctive 'stylistic signature' is usually made up of many tiny strokes'. Burrows (2002) used the Delta method, which is used in some other studies in statistics-based authorship attribution (e.g. Hoover, 2004) and in one of the forensic linguistics studies on microblogging texts (e.g. MacLeod and Grant, 2012). Delta is defined 'as the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text' (Lucykx, 2010, p.16).

Spassova and Turell (2007) used two-, three- and four-word grams as features such as determiner-noun and determiner-adjective-noun in the data that was collected from three authors. The methodology was based on qualitative, quantitative and semi-automatic approaches. The researchers reported that 'the use of trigrams produces more statistically significant results than the use of bigrams, especially trigrams consisting of prepositional phrases and, to a lesser extent, verbal and compound adjective phrases' (Spassova and Turell, 2007, p.192). Word grams were used in different studies and had promising results. For instance, Wright (2014) conducted a study on emails and used word n-grams between two and six, which were helpful in analysing authorship. Moreover, Nini (2018) investigated the corpus of the serial killer Jack the Ripper who was involved in the Whitechapel murders in 1888. There were 209 texts with an average length of 83 words, and two-word grams were used, which are the most basic useful features in an authorship attribution task. However, all words cannot be treated as features. Two-word sequences were used in a study conducted by Nazar and Sanchez Pol (2006). They collected 100 newspaper articles of around 400 words each from five authors and achieved 92 per cent performance when half of the data used. Although the length of the word grams did not consist of longer sequences, even a sequence of two words showed considerable success in authorship attribution.

Moreover, in stylistic studies, unusual word combinations or structures are mostly considered as better discriminative features than the highly frequent ones. Furthermore, 'once occurring markers, or infrequent markers are also paramount for identifying individual authorship as they indicate the individualised language uses of the particular author' (McMenamin, 2002, p.311).

Different from the previous studies, Koppel and Schler (2003) used error-based features, which were based on word-based and syntactic-based errors of an author in stylometric studies. However, errors are generally used by stylistic researchers.

Complexity features were also used in the previous authorship attribution studies. For instance, Grieve (2007) used 39 sets of textual measurements such as one-character and two-character word lengths, sentence lengths, vocabulary richness, grapheme frequency, word frequency, punctuation marks, and character-level n-gram frequency. The best results were attributed to the combination algorithm, which 'distinguished between five possible authors with over 90% accuracy' (Grieve, 2007, p.267). Similarly, Stamatatos (2013) analysed a Guardian newspaper corpus and found that character n-grams had more discriminative potential than the other features.

Moreover, Lopez-Escobedo et al. (2013) used several stylometric features to performed for authorship attribution when the data set varies regarding size and genre. The data set consisted of 27 long texts from professionals (10,000-word tokens on average) and 15 short texts from students (500-word tokens on average) in various genres ranging from journalism, novel, essay and play to the short story. Similarly, Garcia-Barrero et al. (2013) employed an approach in authorship attribution in Modern Standard Arabic and used several variables, including type-token ratio, word length, punctuation, conjunctions, the combination of conjunctions and punctuation, and sentence length in words. Five equally sized samples from three authors were tested using a quantitative approach, and very positive results were obtained. For instance, a combination of punctuation marks and conjunctions classified 59 out of 60 correct assignations.

As defined above, there is a great difference in the number or type of the features that are used in different studies. For instance, most of the studies performed part of speech features (e.g. Baayen, 2002) punctuation marks (e.g. Chaski 2001), function words (e.g. Argamon and Levitan, 2005), character n-grams (e.g. Grieve 2007) and word n-grams (e.g. Wright. 2014). Overall, there are studies which focused on single category (lexical, syntactic or structural), and two or three of them together. Examples of these feature subcategories are presented in this section can increase potentially in the future. There are studies that used only one category or three of them together for instance Wright (2014) focused on only word n-grams rather than separately performed various features in his study. Moreover, the number of features employed in a study can easily reach to high numbers since it is not possible to present each feature individually in this study. Some studies mentioned above used less than 100 features (e.g. Grant, 2007; Grieve 2007) at the same time some employed statistical feature reduction

techniques in order to decrease the number of features (e.g. Grant, 2007; Rico-Sulayes, 2011). However, such type of statistical feature reduction techniques is not used in stylistic approaches. There are many studies based on stylometric complexity features; however, only those that are regarded as significant are presented in this chapter, since the purpose of the book is not related to fully stylometric approaches.

2.4. The Role of Size in Authorship Attribution Approaches

In authorship attribution studies, it is widely accepted that the longer the text, the easier it is to eliminate features, and the results are more reliable than for the short texts. Furthermore, it is assumed that there are long texts but a few authors. However, 'in the real world…there is no guarantee that the true author of an anonymous text is even among candidates' and 'the amount of writing we have for each candidate might be very limited, and the anonymous text itself might be short' (Koppel et al., 2012, p.317). In an ideal case, there would be enough disputed and known texts. However, 'the forensic world is rarely ideal, and the texts are often unhelpfully short; many of the texts which the forensic linguist is asked to examine are very short indeed' (Coulthard et al., 2017, p.152). There are two main challenges in authorship attribution cases, either the text is long, but the style is unique (e.g. the Unabomber case) or the text is too short to be considered as evidence (e.g. short text messages).

Due to the popularity of social media, the role of size in authorship attribution studies has changed over the years from one of the longest documents, Federalist papers (Mosteller and Wallace, 1964), to micro-messages (e.g. Layton et al., 2010; MacLeod and Grant, 2012; Grant, 2013) with limited word lengths.

The role of size is classified into three sections: text size, candidate author size, and limited texts per author. Since some approaches overlap in terms of their methodology, some studies are presented once in one section and not repeated in the following sections.

2.4.1. Text Size

Stylometric research studies have analysed texts with more than 1,000 words (e.g. novels, essays, newspaper articles). Forsyth and Holmes (2001) suggested that 1,000-word texts could be the limit for correct attribution. However, Argamon et al. (2008) reset the minimum length

limit in their study and achieved good accuracy results in shorter texts. Coulthard (1994) discussed the sample size:

> No one really knows how small a sample one can reliably work with, at what size significant irregularities begin to emerge, whether two samples by the same person can be treated as one larger sample and to what extent regularities vary across registers. (Coulthard, 1994, p.93)

Burrows (2002) applied a statistical algorithm to 200 poems by 25 poets. He divided the poems into five bands according to the length: 500 words or fewer, 500 to 1,000, 1,500, and 2,000 or more. The study found that the texts of 1,500 words or more were effective enough to determine correct authorship. Diederich et al. (2003) experimented with short and medium-sized texts (average 720 words but some with 200 words) collected from German newspapers. The analysis was performed on 82 and 118 texts from seven authors. Word frequencies achieved 72 per cent accuracy, while bigrams and function words achieved 61 per cent.

Moreover, because of the internet, authorship attribution studies have started to focus on small texts. For instance, Abbasi and Chen (2005; 2008) used web forum messages and eBay comments and collected 20 messages from 20 authors, a total of 400 messages, in English and Arabic. The average length of the English texts was 76.6 words, while it was 580.69 words in the Arabic texts. For this study, the English language feature set was adapted from the previous authorship attribution studies. Lexical (character-based, word-based), syntactic (punctuation, function words, word roots), structural (word structure, technical) and content- specific (race/nationality, violence) feature types were used to identify the writing style in web forum messages. The results showed up to 97 per cent success, and since they have promising results, they assumed to apply this method to many authors. for the future studies which may include many authors depending on the promising results. Furthermore, Pavelec et al. (2007) collected data on different topics from ten columnists from two Brazilian newspapers. There were 15 short articles on various topics, namely economics, politics, sports, literature, wine, gossip, and miscellaneous, with an average of 600 words. For the analysis, 77 conjunctions were proposed, and the recognition rate of the method achieved 75.1 per cent accuracy. Although the average text size was similar to the average length of the English texts in Abbasi and Chen's (2005;2008) research, there was a 20 per cent difference in the success rate; this may be due to the usage of a single feature type rather than focusing on different feature types.

Eder (2010) tested the correlation between test size and authorship attribution success on different samples. The results for a corpus of 63 English novels showed that samples shorter

than 5,000 words provide poor 'guessing'. Furthermore, texts with fewer than 3,000 words caused more than 60 per cent false attribution. However, Argamon and Koppel (2013) suggested that 1,000 words are enough for limiting the number of features. In their study, functional lexical features, conjunctions, prepositions, modal verbs, adverbial adjuncts, projective clauses and content features were used. However, the use of content words is described as problematic, as it may present highly optimistic accuracy results, but these results do not correspond to real-life applications. The aim of Argamon and Koppel's (2013) study was to profile authors regarding sex, age, native language, and personality. First, the corpus included 19,320 blog authors, with an equal number of males and females, with an average of 7,250 words for each. The corpus was divided into three groups based on the reported ages of the authors. Second, the data was collected from authors from various countries, such as Russia, the Czech Republic, Bulgaria, France and Spain, who speak English as a second language. Finally, university students were instructed to write a short essay about 'stream of consciousness' in 25 minutes. At the end of the analysis, profiling languages achieved a 79.3 per cent success rate in using style and content features, while gender, age and personality achieved 76.1 per cent, 77.7 per cent and 63.1 per cent respectively. When the style features were excluded from the language class, 82.3 per cent accuracy was achieved, which is the highest of all the values.

Some researchers (e.g. Van Halteren, 2005) overcame this by collecting a large number of training texts for each author. However, Hirst and Feiguina (2007) used small texts with 1,000, 500 and 200 words but collected hundreds of texts for the research. Nevertheless, it was reported that the performance of the accuracy decreased significantly when the text size was reduced to fewer than 500 words. However, in recent years, authorship attribution studies focused on texts with fewer than 500 words.

In one of the first studies, Layton et al. (2010) analysed tweets with up to 140 characters and collected 14,000 Twitter users and their most recent 200 tweets. They applied a methodology called Source Code Author Profile, which is based on character n-grams, and included Twitter-specific features (e.g. @replies, hashtags and RTs). Their method performed well on Twitter messages and showed that 120 tweets per author were a threshold in short text authorship attribution studies.

Silva et al. (2011) collected four million messages in Portuguese from 200,000 users. From this corpus, they chose the most active 200 users with at least 2,000 tweets per author, excluding retweets. Various datasets were created depending on their size, ranging from 75, 250, and

1,250 to 2,000 tweets, to examine the text size effect in authorship attribution. For this study, along with the Twitter-specific features such as @replies and hashtags, some non-lexical features, for instance, emoticons, interjections and punctuation, were used. The results showed that emoticons were effective in attributing authorship in Portuguese tweets while punctuation marks showed average results. Furthermore, due to the effect of the combination of features, it was argued that 'all features carry relevant information, since using all groups of features simultaneously allows inferring more robust authorship classifiers than using any group of features individually' (Silva et al., 2011, pp.167-168).

Boutwell (2011) used character n-grams between 53 authors from Twitter. Different from the other studies, she combined the individual tweets into a single document. In a similar vein, Mikros and Perifanos (2013) aggregated several tweets into a single document in their study. This approach was criticised by Schwartz et al. (2013) and Rocha et al. (2016) as unrealistic and unfeasible, as it considered that the scenarios with a single tweet are in an actual investigation. The investigation may require a substantial amount of texts written by the same author. Schwartz et al. (2013) used a combination of character and word n-grams and achieved 49.5 accuracies when analysing 50 authors with 50 tweets each. This result increased to 69.7 per cent when the number of tweets was increased to 1,000 per author. However, when they increased the number of authors with 200 tweets each, the accuracy dropped dramatically to 30.3 per cent. Rocha et al. (2016) collected ten million tweets from 10,000 authors within six months in 2014 and achieved 70 per cent accuracy for 50 authors.

Mikros and Perifanos (2013) developed the first Modern Greek Twitter corpus, which included 12,973 tweets from ten popular users, and divided the data set into four different sizes, 100, 75, 50 and 25 words long, with the aim of exploring the effectiveness of n-gram features. In contrast to Layton et al. (2010), Mikros and Perifanos (2013) decided to exclude @replies, hashtags and RTs to develop a methodology based on linguistic features only. As mentioned earlier, they merged tweets into a single document, and they achieved the best success rate of 0.952 and 0.918 when using 100-word and 75-word texts.

Similarly, for weblogs, Koppel et al. (2011b) and Schwartz et al. (2013) ran an authorship analysis on micro texts. Since Twitter is a fruitful database and presents flexible patterns, it is becoming popular in authorship attribution studies. Schwartz et al. (2013) collected tweets for the analysis. It is worth noting that Twitter initially restricted tweets to 140 characters; however, this has been doubled over time (Twitter Blog, 2017). Therefore, the text size was limited to 140 characters when the data was collected for this study. Even though the character

size has doubled in Twitter, the texts are still short when compared with the literary texts. In their research, Schwartz et al. (2013) used a varying number of authors (between 50 and 1,000) selected from 50 to 1,000 tweets per author. Character n-grams and word n-grams which were between two and five words long were used. The results for 50 authors and different data set sizes showed that authors of short texts could be identified even with 50 tweets per author. A different dataset with 200 tweets per author demonstrated that authors could be detected when there are 1,000 candidate authors.

Bhargava et al. (2013) combined lexical, syntactic, Twitter-specific and metadata featured with the authors from ten to twenty and achieved up to 95 per cent correct classification. The results also showed that both removing lexical features and increasing the number of candidate authors lowered the accuracy. Due to the internet, text size has become a debatable factor in authorship studies. However, when the real-world cases are considered, the candidate author size is as important as the size.

2.4.2. Candidate Author Size

Recently, candidate author set size has started to attract the attention of authorship attribution studies in order not to overestimate the results based on scenarios when it is compared with the real-life cases. Juola (2008, p.320) recommended that 'the set of candidate authors must be chosen as carefully and rationally as possible, and the set of candidate writings must also be chosen with equal care and rationality'.

The effect of candidate author set size has been examined by many scholars. While computational studies (e.g. Koppel et al., 2011) focused on thousands of candidate authors, some scholars concentrated on a limited number of candidates. For instance, Grant (2007) conducted an analysis of three authors, Grieve (2007) 40 authors, and Rico-Sulayes (2011) ten authors. Regarding the limited candidate author size and text sizes in forensic linguistics studies, Olsson (2004) stated that:

> [M]any forensic inquiries present too little data to justify a full-scale statistical study. Often there are no more than two candidates, as few as three attested texts and perhaps no more than one or two questioned texts. This is why most methods described as authorship detection do not work in the forensic context. (Olsson, 2004, p.14)

Argamon et al. (2003) used 20 authors in a corpus of a newsgroup on various topics. Furthermore, Argamon et al. (2003b) found that if the number of authors is between two and 20, this causes a performance drop.

Moreover, Grieve (2007) reported that when the author set was increased from two to 40, the results showed a significant decrease. Luyckx (2008) tested the effect of many authors on feature selection on 20,000 words from 145 authors. The study demonstrated that when the number of authors was increased, the performance dropped dramatically. Specifically, when the number of authors was two, the accuracy rate was 96.90 per cent, while with five, ten and 20 authors, there was a gradual decrease of 88 per cent, 82 per cent and 76 per cent respectively. Furthermore, in less than half of the cases, a text from one of the 145 authors was assigned to the correct author correctly. Koppel et al. (2011, p. 85) collected 10,000 blogs from blogger.com. For each blog, 2,000 words of known texts were selected as snippets consisting of the last 500 words of the blog. Their method attributed a 500-word snippet to one of 1,000 authors with coverage of 42.2 per cent and accuracy of 93.2 per cent (Koppel et al., 2011, p.93).

Luyckx and Daelemans (2011) ran an analysis to test the effect of author set size and data size. For this purpose, they collected three datasets containing 268 texts in English and Dutch on various topics from 166 authors. To test the effect of author set size, the number of authors was gradually increased, which caused a significant decrease in performance. In total, 145 candidate authors achieved 11 per cent accuracy in authorship attribution, while five candidate authors achieved 70 per cent success. Although the candidate author size was small, limited texts are another limitation in authorship attribution studies. In the following section, these studies are presented.

2.4.3. Limited Texts per Author

The task of analysing authorship in forensic cases is generally not applicable to thousands of authors but rather a few authors, maybe two in some cases (e.g. Grant, 2013). Sometimes, the language data may not be enough to present meaningful results. Coulthard et al. (2017) stated:

> In an ideal world, there would be a substantial amount of data to work with, both disputed and known. However, the forensic world is rarely ideal, and the texts are often unhelpfully short; many of the texts which the forensic linguist is asked to examine are very short indeed – most suicide notes, ransom demands and threatening letters, for example, are well under 100 words long. (Coulthard et al., 2017, p.152)

Coulthard et al. (2017, p.205) added that 'one major difficulty for forensic linguists involved with authorship attribution is that typically the amount of questioned data is small and there are no useful population statistics' regarding the limited text numbers per author. In the Amanda Birks case discussed above, Grant (2013) had 204 short text messages belonging to the victim and 203 texts from the victim's husband. Considering the average size of each short text message, this does not provide a large database for authorship attribution.

Similarly, Kredens and Coulthard (2012) analysed one of the real-world cases depending on the short text messages in a small set of corpora used in this study. Moreover, in another example presented in Coulthard et al. (2017), the available disputed text was an email with only 140 words. However, there was a large number of known emails produced by two candidate authors. In this case, Coulthard used Grant's (2013) method and found 13 distinctive words and phrases from two authors at the end, 11 of which belonged to candidate author A and only two of which belonged to candidate author B.

In addition to the real cases, some other scholars tested the effect of the limited texts on authorship attribution applications. For instance, Hänlein (1999) collected between 13 and 17 texts from each author, which is a considerably small number for stylometric studies. Moreover, Stamatatos (2007) used four different corpora to test the extreme conditions. For instance, in one of the corpora, there were 50 texts per five authors and ten texts per 50 authors. In many cases, the results did not present a high level of accuracy and showed very low accuracy in all cases.

Overall, limited texts per author are related to the limited language data per author which is different from small text sizes. Various text sizes refer the number of words per text supposing that there are fifteen texts per author and each text has at least 400 words, however, limited texts per author indicates the insufficient number of texts per author. For instance, a suicidal may leave only a letter behind and it can be the only available document written by the person thus, it is called as limited text when it is compared with the candidate author's texts. A disputed author has three texts available in limited texts per author corpus. As it is mentioned above, one of the few studies focused on the limited number of texts (e.g. Stamatatos 2007) unlike an overestimated number of training texts in traditional stylometric studies. Along with text size, candidate author set size, limited texts per author, the cross-genre application is also important in authorship attribution studies.

2.5. Cross-genre Application in Authorship Attribution

The genre is one of the important factors in authorship problems since each text has a different composition style. For instance, online texts are considerably different from literary texts. Coulthard et al. (2017, p.152) stated that 'any known data used for comparison would ideally consist of large sets of texts of the same text-type as the questioned documents, written in the same register and at around the same time'. However, in real cases, it is not possible to control the effect of genre. For that reason, there have been some attempts to identify the cross-genre corpus effects. Stamatatos (2013, p.421) emphasised the effect of cross-genre application in authorship attribution and claimed that the problem of controlled genre and topic represents 'low-level representation' for the realistic studies since it is impossible to keep the conditions stable.

With the strong influence of the internet, new computer-mediated communication genres, including emails, online chats, blogs, short text messages and tweets, have emerged. The characteristic of these genres has attracted the interest of researchers for a long time (e.g. Crystal, 2001; Herring, 2001). Most studies focused on a single genre such as blogs, emails, and text messages.

Goldstein-Stewart et al. (2009) used text samples including essays, emails, blogs, and chat data. Audio samples which were collected from individual interviews and group discussions were also included in the study. For each genre, samples were collected for six different topics. The research showed a significant success rate for several cases, for instance, an accuracy rate of 71 per cent was achieved on samples from across genres. For person identification, 82 per cent success was achieved in five genres, and for topic identification, an average accuracy rate of 94 per cent was achieved.

A cross-genre study conducted by Kestemont et al. (2012) used the unmasking method in cross-genre authorship verification. The study focused on five representative contemporary English-language authors. The published texts were from two genres (literary prose and theatre play). First, a traditional intra-genre experiment was run on five authors and achieved 96.36 per cent overall accuracy. The cross-genre procedure was then carried out on the literary prose and theatre play genres. However, the result 'does not seem able to capture the overall difference between same-author and different author text pairs across two genres' (Kestemont et al., 2012, p.349).

Nini and Grant (2013) proposed that systemic functional linguistics framework (Halliday&Hasan, 1989) and multidimensional analysis (Biber, 1988) are possible methods to use in cross-genre authorship analysis even though their study is not based on cross-genre analysis. SFL context include three components as field (what is happening?), tenor (who is taking part?) and mode (what part language is playing?) and these can be used to identify linguistics differences across the text types such as threatening letters and business letters. Based on the results of their study, Nini and Grant (ibid) stated the possible outputs in cross-genre analysis as:

> …field affects just the content words of a text, tenor the number of declaratives/interrogatives/ imperatives and the expression of attitudes, mode affects cohesion and theme-rheme patterns. Therefore, if two genres for which field and tenor vary but for which mode is comparable are considered, then the measurement of the code of the author can be produced by only looking at those variables for which how much variation we expect for the genre is known, that is, cohesion and theme-rheme patterns. (p. 195).

Moreover, some studies which used SFL showed a correlation between contextual parameters and linguistic variables (ibid). Thus, it is possible to use such type of methodology in forensic cross-genre attribution problems.

Schwartz (2016) compared to blog posts and tweets by using word n-grams, character n-grams, and stop words. Nine bloggers with active Twitter accounts who had at least 200 tweets were selected for the study. Linear SVC performed better when using word unigrams and 1–9-character n-grams among three different algorithms in the cross-domain study.

Overdorf and Greenstadt (2016) compared blogs, Twitter feed and Reddit comments using the cross-domain approach in authorship attribution. They distinguished the cases by the amount of text in the unknown documents as 500, 2,000 and 4,500 words.

Stamatatos (2013) conducted various experiments to test realistic circumstances regarding candidate authors, topic, genres and distribution of texts over the candidates. Two main feature sets were used for the study. The first was a predefined list of words such as articles, prepositions and the most frequent features. The corpus was compiled from the texts published in The Guardian newspaper.

In the first test, the models based on character 3-grams were found to be more effective than models based on words in intra-topic attribution and achieved perfect classification accuracy. The second test, in the cross-topic scenario, the texts about politics, society, the world and the UK were examined, and it was found that character 3-grams were more effective than the

words, which is similar to the result in the previous test. Finally, newspaper articles were compared with another genre, book reviews, and again character n-grams performed better than word features. It is evident that character n-grams worked successfully for the corpus in Stamatatos's (2013) study in various conditions. However, he stated that in real-world applications, 'a one-model-fits-all approach' (Stamatatos, 2013, p.438) is not adequate in determining authorship of the texts under investigation.

Cross-genre authorship attribution is still an underexplored field in forensic linguistics although it is an anticipated situation where the disputed texts and known texts differ in the settings they have created. Moreover, it is a challenging task and standard methods may lead to misclassifications in cross-genre applications. When both disputed and undisputed texts are in the same genre in authorship attribution studies, less effort is needed during the analysis. However, in some cases, it is unrealistic to obtain the same amount of known and disputed texts from the same source. Furthermore, standard methods limit the practical usage of the methods when it is simply focused on a single genre. Some methods stated above (e.g. Schwartz, 2016) had promising results and provided multiple contributions in comparing genres including essays, blogs, tweets, Reddit comments to the field. However, there are still more tasks needed in order to discover different features and improve the ability of the features to attribute the authorship correctly in cross-genre scenarios.

Besides, a great majority of the studies presented above are done for English and continued to improve the success rate of the results over recent years, each improvement gaining an increase in the robustness of the methods. Along the same line, authorship attribution studies in other languages are increased in order to establish high accuracies. Next section furthers the contributions to the field by presenting Turkish authorship attribution studies.

2.6. Authorship Attribution Studies in Turkish

As mentioned in the introduction, there are very few authorship attribution studies in Turkish, and there is no study on forensic linguistics. One of the challenges in these studies is that all of them are focused on stylometric features such as word length, sentence length, etc. Such features may not be dependent on language and suitable to apply to other languages. However, the results are not reliable to present in the courtrooms due to the lack of linguistic explanations. Other common factors in the studies are data type and text size. To the best of the researcher's knowledge, to date, there has been no attempt to analyse small or online texts other than the

studies of Kucukyilmaz and Cambazoglu (2008) and Ekinci and Takci (2012) in emails in a Turkish language context.

In 2003, Amasyali and Diri (2003) attempted to classify texts in terms of the author's gender by using vocabulary richness, word type frequencies and distributions, word and sentence clause, paragraph length and distributions, syntactic analysis, co-occurrence and collocations, and content analysis. They collected data from one of the prominent Turkish newspapers on hurriyet.com.tr on various topics. The length of the texts was approximately 456 words. They achieved 84 per cent success by using those features in distinguishing the authors with the aid of a multilayer perceptron algorithm, which is a simple algorithm intended to perform binary classification (Amasyali and Diri, 2003).

Following this research, Can and Patton (2004) used the most frequent words in the two novels by two Turkish authors, namely Cetin Altan and Yasar Kemal. They used discriminant function analysis and achieved 92 per cent success in categorising the texts as new and old.

Amasyali and Diri (2006) used 630 single-authored documents from 18 authors to identify the author, the text's genre, and the gender of the author. They ran Naïve Bayes, Support Vector Machine, C 4.5. and Random Forest as computational identification methods. They achieved 83 per cent accuracy in attributing authorship, 93 per cent in the genre of the text, and 96 per cent in the gender of the author.

Turkoglu et al. (2007) used Amasyali and Diri's (2006) 630-document corpus to attribute authorship by using feature mining research. For the analysis, they divided the data sets into three sections and achieved success rates of 72.4 per cent, 80.0 per cent and 82.9 per cent respectively by using n-grams and various combinations of feature vectors. Following this, Tas and Gorur (2007) developed another classification method by using 35 style markers among 20 newspaper column writers. This method achieved an 80 per cent success rate between 20 authors.

Bozkurt et al. (2007) collected data from the Turkish newspaper *Milliyet* and ran an analysis depending on various features, such as a bag of words or frequency of function words, and achieved 95 per cent success by using a bag of words feature set.

Different from the previous studies based on newspaper and novel corpora, Kucukyilmaz and Cambazoglu (2008) employed data mining approaches on web chat messages to assign the gender of the author. Since the smileys are important features that are available in chat messages, they were considered as a feature along with the traditional stylistic features such as

character usage, message length, word length, punctuation usage, and so on. In total, 218,742 chat messages were collected from 1,616 unique users. This study found that in online environments, the use of slang words and misspellings are frequent in Turkish, which is in line with the other languages.

Takci and Ekinci (2012) used character n-grams as features, including punctuation marks, letters and some special marks. The dataset for the study comprised ten articles by authors from Sabah, a daily newspaper (one of the Turkish newspapers). The researchers concluded that 'character-based authorship is superior to other methods in respect to performance and effectiveness' (Takci and Ekinci, 2012, p.16). In the same year, Ekinci and Takci (2012) performed an analysis of emails which were obtained from five authors. In this study, 43 textual features were extracted, and data mining classification techniques were applied. On average, 83 per cent success was achieved on the available data set.

A recent study conducted by Bay and Celebi (2016) included 850 columns written by 17 columnists. They applied machine learning algorithms and achieved a 99.7 per cent success rate.

Finally, Agun et al. (2017) presented an authorship attribution method based on morpho-syntactic and POS tags in a data set which was constructed from Turkish newspaper articles written between 2015 and 2017. From the morphological point of view, Turkish is an agglutinative language which is rich in suffixes (this is discussed in the following section). The results indicated that neither part of speech tags nor morphosyntactic tags had the expected performance gain.

With the presentation of authorship attribution studies in Turkish, three important gaps are identified. First, the abovementioned studies were based on stylometric approaches and employed a list of various features ranging from word-lengths to character-grams without depending on linguistic explanations. One recent study considered the structure of the language in the target. Furthermore, the majority of the studies used newspaper columns as a data set rather than computer-mediated communication mediums, which is one of the most fruitful areas for authorship attribution studies. However, in conducting authorship analysis in real cases, the language of the texts may not include standard Turkish features like those in newspaper columns or literal texts. There is only one study that used data from web chat messages (Kucukyilmaz and Cambazoglu, 2008) and one that used data from emails (Ekinci and Takci, 2012).

A summary of the Turkish authorship attribution studies is outlined in Table 2-1. In the following section, the structure of the Turkish language will be explained.

Table 2-1: The summary of authorship studies in Turkish.

Study	Corpus	Method	Results
Diri and Amasyali (2003)	18 authors from www.hurriyet.com.tr Newspaper	22 Style Markers	84% Success
Can and Patton (2004)	Novels from two authors (Cetin Altan and Yasar Kemal)	Most frequent features, word lengths Discriminant Function Analysis	92% Success
Amasyali and Diri (2006)	630 singly authored documents from 18 authors from three Turkish newspapers	Naïve Bayes, Support Vector Machine	Author of text 83% Genre of a text 93% Gender of the author 96 %
Turkoglu et al. (2007)	35 texts per 18 different authors from Turkish newspaper www.hurriyet.com.tr www.vatanim.com.tr	Vocabulary richness, grammatical features, function words, n-grams Naïve Bayes, Support Vector Machine	The dataset I 72.4% Dataset II 80.0% Dataset III 82.9 %
Tas and Gorur (2007)	20 different authors 170000 words	35 style markers (e.g. number of punctuation marks, word-based features, pronoun, conjunctions) Naïve Bayes Multinomial	80% success
Bozkurt et.al (2007)	Turkish newspaper www.milliyet.com.tr 500 articles from 18 authors	Vocabulary diversity, bag of words, frequency of function words Principal Component analysis	95% success rate with bag of words feature set
Kucukyilmaz and Cambazoglu (2008)	20000 online chat messages from 100 authors	78 Turkish Function words	99.7 % success
Takci and Ekinci (2012)	10 articles from Turkish newspaper – SABAH	Character n-grams including punctuation and function words	53 % functional words 86 %-character n-grams
Ekinci and Takci (2012)	Electronic mails from 5 authors	43 features	83 % success

		Data mining techniques	
Bay and Celebi (2016)	Two different newspapers 850 columns from 17 columnists	Machine Learning Algorithms	99.7 % success
Agun et al. (2017)	Turkish Newspaper articles between two years	Morpho-syntactic tags	73.22 percent between 25 authors

2.7. Turkish Language

The Turkish language is the most widely spoken for many centuries across vast territories spread from the Balkans to China. It belongs to the Turkic family of languages. Turkish is the official language of the Republic of Turkey and Northern Cyprus, and thus it is predominantly spoken in those regions (Goksel and Kerslake, 2005). After the foundation of Modern Turkey, the point that 'Turkish is the official language of Turkey' was added to the 1923 Constitution. Ottoman scripts were replaced with the modern 29-letter Turkish alphabet by the Law on the Adoption and Implementation of the Turkish Alphabet (1928) which was one of the important parts of Mustafa Kemal Atatürk's Reforms. This alphabet is another version of the Roman alphabet and lacks the characters Q, W, X but includes Ç, Ş, Ü, Ö, İ, Ğ.

Theoretically, the Turkish spoken in Istanbul is accepted as standard Turkish which is spoken clearly, smoothly and slang is limited. Regardless of the region, the standard Turkish is taught in all regions including Northern Cyprus. The structure and the features of contemporary Turkish are explained in Table 2-2. (Goksel and Kerslake, 2005).

Table 2-2: The features and the structure of the contemporary Turkish.

Table 2-2 is created from 'Turkish, A Comprehensive Grammar' book written by Goksel and Kerslake (2005).

Orthography and Phonology	**Spelling is phonetic.**
	Optional circumflex accent (Optional circumflex accents can be used to distinguish the words which have the same spelling but different meaning.

	It has vowel harmony. (In Turkish vowels characterised by three features: front, high and rounded/unrounded. Two vowels cannot combine together in Turkish words)
Typology	Word order is SOV
Grammar	Lack of Grammatical Gender. For the third person, there is only one pronoun.
	All verbs are regular in all tenses.
	Postpositions (e.g. bu yana (since) Cumadan bu yana (Since Friday) Turkish postpositions usually correspond to prepositions in English
	No definite/ indefinite articles in nominative
	Gerunds for verbal constructions
	Impersonal passive a form that occurs in Turkish but not in English, namely the passive form of an intransitive verb (e.g. sevinilir 'one is pleased', 'people are pleased') Pluperfect is used more frequently than its English counterpart to show that one past event preceded another. (e.g. gelmistim means I was having come, 'sehre 10'da varmistik, burosuna saat 3'te gittik' 'we (had) arrived in the city at 10 and went to his office.
	Aorist refers a habitual aspect. (e.g. yapar-im 'I habitually do'
	It does not tend to form agentive passive sentences they not used as much as its English equivalent.
Verbs	There are no phrasal verbs, but there are some verbs that are very commonly used with nouns to make verbs such as -etmek, -olmak
Sentence Formation	Pronouns are omitted in a sentence since the person is implied in the adjectives or the verbs in sentences.
	The inverted sentence does not affect the meaning of the sentence.
Politeness Level	T-V distinction
	Apostrophe is used for separating a proper noun from its inflectional suffixes (e.g. Hulya'dan, from Hulya)
Word Formation	Processes of word formation create words that can be very long and sometimes it correspond the whole sentence in English. (e.g. Çekoslovakyalılaştıramadıklarımızdanmışsınız - You are apparently one of those we were not able to make a Czechoslovakian)
Negation	The negative marker –mA which precedes tense, mood and person affixes and follows reflexive, causative or passive affixes. (e.g. old-ma-di / It didn't happen)
Vocabulary	Reduplication is the repetition of a word of a word. It happens in three ways as follows: (i) emphatic reduplication: kipkirmizi 'stark red' (ii) m-

	reduplication: cirkin mirkin ‘ugly or anything like that’ (iii) doubling: yavas yavas ‘slowly’. M-doublets are the words which are followed by an echo of itself but with m replacing its initial consonant or preceding its initial vowel. The meaning of this form is ‘and so on and suchlike’ (e.g. dergi mergi okumuyor ‘s/he does not read journals or periodicals or magazines’). There is no such a usage in English.
	Vocabulary includes many load words mainly from Persian, Arabic and Greek.
Morphology	Agglutinative (Suffixes are attached to the end of the word whether it is verb or noun- even the simplest words such as in, at and on can add suffixes)
	No prefix (Apart from some Arabic loan words which have their prefixes)

Chapter 3: Literature Review: Internet Language and Computer-Mediated Communication

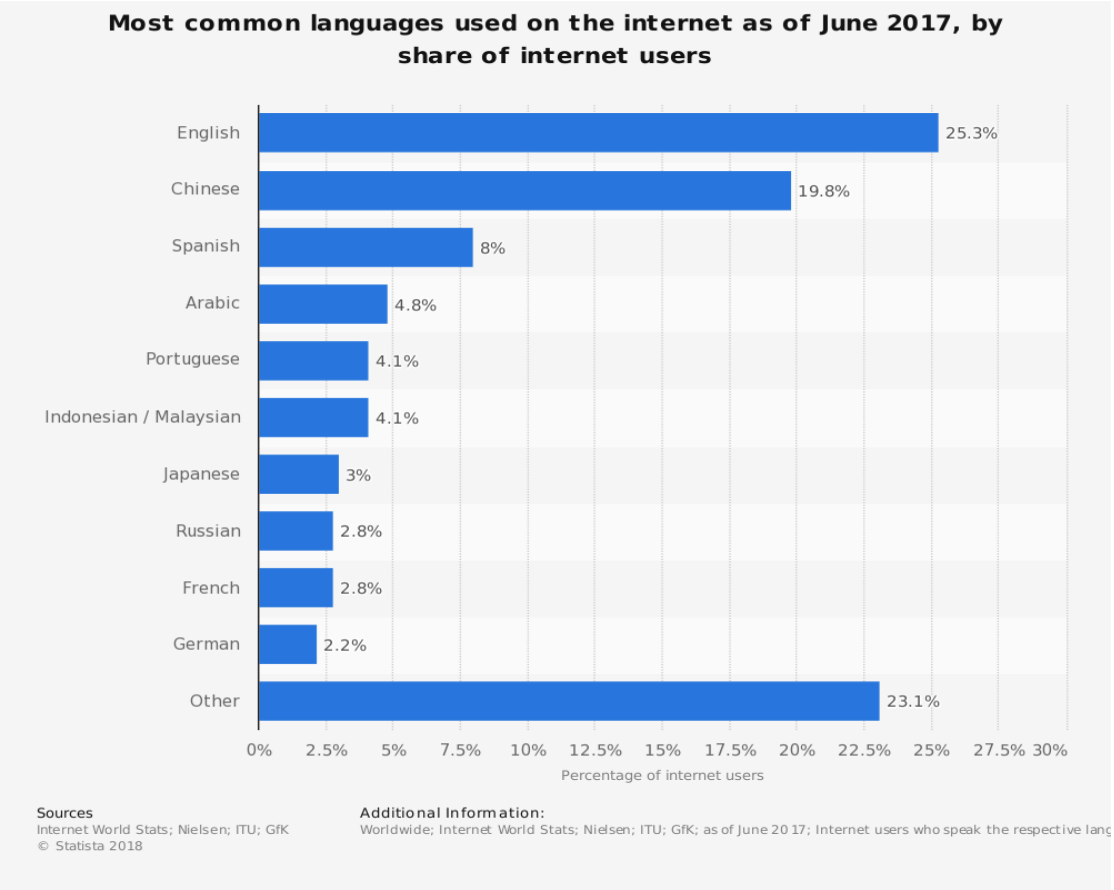
The Internet has been actuated by a movement that has developed new types of communication in virtual settings, such as e-mails, blogs, online chats, and many others. Communication via electronic devices on the Internet has affected the form of both languages and communication itself. In this era of information and communication technologies, computer-mediated communication (henceforth CMC) can be thought of as a fundamental aspect of the development of a new form for languages.

This chapter first defines Internet language (Section 3.1) and provides a subsection on Internet language in Turkish (3.1.1). The subsequent sections discuss computer-mediated communication and virtual communities (Section 3.2), Eksi Sozluk as a collaborative online encyclopaedia (Section 3.2.1), and content generated by Twitter users (Section 3.2.2).

3.1. Internet Language

With the invention of the Internet, new type of Internet-specific linguistic features has emerged. These features include new identity markers, such as chat room nicknames or web address extensions, creative keyboard usage, acronyms, and abbreviations. Despite the fact that the web is a multilingual corpus, a survey of the most common languages on the Internet as of June 2017 revealed that English is the most represented language online, used by 25.3% of Internet users worldwide due to its usage in globalised trade and relations (see Figure 3-1, 'Most Common Languages Used on the Internet'). Chinese is ranked after English with 19.8%, followed by Spanish with 8% and Arabic with 4.8 %; other languages, including Turkish, make up the remaining 42.1% . Based on the fact that English is the most predominant language on the Internet, it can be concluded that English has a consolidating, establishing power over Internet language.

Figure 3-1: Most common languages used on the Internet.



The dominance of English on the Internet is also due to technical reasons since the first protocols devised to carry data on the Internet were initially developed for the English alphabet (Crystal, 2003). Consequently, subsequent improvements (such as search engines) have been overwhelmingly dependent on English. Crystal (2007) pointed out that no matter what kind of environment the Internet creates, it has consequences for the type of language found there.

Furthermore, Internet language has common features with written languages, which can be classified as (i) grammatical features, which provide the user many possibilities in sentence structure; (ii) lexical features, which are the set of vocabulary; (iii) graphical features, which are the visual structure of a message (e.g. fonts, text size etc.); (iv) discourse features, which encompass the structural organisation of a text and are defined in terms of such factors as coherence, relevance, paragraph structure, and the logical progression of ideas (Crystal, 2004, pp.7-8).

The presence of the Internet added two more features connected to speech: namely, phonetic and phonological (ibid). Similarly, Internet language has the same features as written language or spoken language, which range from grammatical features to discourse features.

'Netspeak' is a term defined by Crystal (2001) as 'a type of language displaying features that are unique to the Internet and arising out of its character as a medium that is electronic, global and interactive' (Crystal, 2001, p.20). Netspeak is divided into subtypes that are related to the different types of communication, such as 'the language of e-mails' or the 'language of chat groups' (Crystal, 2001). The features of netspeak include word creations, abbreviations, distinctive orthographic graphology related to writing with different fonts, rebus-like abbreviations (e.g. *are > r, you > u, and > n*; p.164), emotional 'noises' (e.g. *hehehe*, *owowowowow*), rebus letter combinations (e.g. *NE1*, *2day*, *B4*, *C U l8r* ['later']; p. 229), capitalisation, and non-standard spelling and punctuation variations (Crystal, 2004). Similarly, Squires (2010) classified the characterisation of Internet language through abbreviations, non-standard typography (e.g. uncapitalised 'i', sentence-initials such as it. and do.), and respellings (p.467). Subsequently, Barton and Lee (2013, p.5) have also described the features of Internet language as including:

- Acronyms and initialisms;
- Word reductions;
- Letter/number homophones;
- Stylised spelling;
- Emoticons;
- Unconventional stylised punctuation.

Emoticons are the 'sequences of keyboard characters that represent facial expressions' (Herring, 2011, p.2). The term 'emoticon', a portmanteau of 'emotion' and 'icon', refers to visual signs such as a 'smiley' or 'angry' face that is used in textual CMC (Dresner and Herring, 2010). Dresner and Herring (2010, p.2) identify emoticons in three ways:

- As emotion indicators mapped directly onto facial expressions;
- As indicators of non-emotional meanings mapped conventionally onto facial expressions;
- As illocutionary force indicators that do not map conventionally onto a facial expression.

Usage of emoticons varies in different cultures. For instance, *kaomoji* signs, which are unique to Japanese culture (Katsuno and Yano 2007), are used in the same way. The emoticons are often seen in text messages (Crystal, 2008) and e-mails (Baron, 2003). However, these kinds of features are also in a state of constant flux. For instance, in the past decade, people tended to minimise the number of characters they used while writing short text messages for economic reasons; recently, however, many telecommunication providers have not charged for text messages, meaning that some abbreviations have been abandoned. Considering all of these forms and features, netspeak can be classified as a language with its own style and characteristics. Due to the constant changes in Internet languages, English-based netspeak improves and changes frequently. Although English is the focus of netspeak, owing to the varying characteristics of different languages, many languages have their own taxonomy for Internet language, including Turkish. Turkish Internet language is examined in the following section.

3.1.1. Internet Language in Turkish

The penetration of the Internet exists in all languages, including Turkish, which has been subjected to a very fast-changing process due to the spread of technology and communications devices. Nevertheless, because English has a considerable effect on the internet, it may unavoidably influence other languages. For instance, 'ç', 'ş', 'ğ' and 'ü' characters in Turkish cannot be used in virtual media unless the website allows this function. Cakir and Topcu (2006) note that new expressions, patterns, symbols and words are becoming more common in Turkish due to the Internet. Thus, the Turkish language on the Internet is becoming very different from standard Turkish, a case resembling what Crystal (2001) defines as 'electronic revolution'.

Furthermore, the pace of development of online social networks has caused a lot of foreign words to be used and adopted into the Turkish language. The fact that many people can communicate with each other, share ideas or chat simultaneously using e-mail, asynchronous and synchronous platforms has changed the language used.

In recent years, Erdogan and Yaman (2007) examined the linguistic features of online messages in Turkish from MSN messages. Subsequently, Temur and Vurus (2009) classified the uses of Turkish on the Internet and identified the features of Turkish Internet language (see Table 3-1).

Table 3-1: Turkish Internet language features.

Turkish	Example
Using English characters instead of Turkish ones: ş>s; ü>u; ç>c; ğ>g	Aslinda (indeed)
Using English characters to create Turkish sounds: ş>sh; ç>ch	Chanta / çanta (bag)
Using w instead of v or Q for K	Qimse – Kimse (nobody) / walla – vallahi (in truth)
Elision	bişi / bir şey (one thing)
Deletion of vowels	Tmm / tamam (okay)
Deletion of consonants	Diosam / diyorsam (if I say so)
Baby talk	Yusunden / yuzunden (because of)
Using English words instead of Turkish ones	Cok enteresan bir topic (It is an interesting topic.)
Writing the same characters repeatedly for emphasis	Hayiiiiiirrrrrrrrrrr (No)
Using capital letters and lower cases at the same time	Ardi arkaSI KESILMIYOR (It flows constantly)

It is possible to observe that some of these features are mainly derived from the English language rather than Turkish, since English dominates the internet. Furthermore, Akkoyunlu and Soylu (2011) conducted a study into incorrect language use on the Internet that focused on social communication networks. Results revealed the common non-standard usages of Turkish on the Internet, which are as follows (ibid, pp.446-449):

- Use of non-Turkish characters;
- Repetition of the same letters;
- Reduction of consonants as future and present continuous tense suffixes;
- Lack of punctuation marks in the sentence;
- Non-standard spelling of the interrogative particle (e.g. ‘sende mi?’);
- Non-standard spelling of -mi and -ki particles;
- Use of English words (e.g. ‘away’, ‘bye bye’, ‘thanks’);
- Use of abbreviations (e.g. ‘mrb’ instead of ‘merhaba’);
- The absence of space between words (e.g. ‘doumgununkutlu olsun’ (‘Happy birthday’))

Most of these findings are shared with the features of Turkish Internet language previously identified by Temur and Vurus (2009). Moreover, these features are mostly aligned with the Internet language classifications identified by Crystal (2004), which were originally made for English. However, the use of Turkish on the Internet needs more linguistic analysis.

As the Internet opened up a new type of communication that includes different linguistic features of different languages, the research field of linguistic studies was consequently broadened. Some scholars claim that internet language may affect phonetical, lexical, syntactical structure of Turkish in a negative way. For instance, Kara (2006) remarked that linguistic mistakes increased in written language due to the Internet. In the same vein, Aksut, Batur and Avsar (2006) found that grammatical rules were violated in the virtual world and written language is used improperly on the Internet (Yaman and Erdogan, 2007). However, studying on Turkish Internet language features would have a contributing role in terms of retrieving information about each author in authorship analysis studies since there are a number of possible designations for the language on the Internet. For instance, the specific use of punctuation as distinctive features or the use of reduction of consonants are employed in the current study in order to attribute the unknown author.

In the following section, the main aspects of the language used on technological platforms, along with the features of computer-mediated communication and virtual communities, are presented.

3.2. Computer-Mediated Communication and Virtual Communities

On the Internet, people interact and communicate within communities as they would in any other real-life discourse. This introduces a specific type of online human interaction, namely virtual communities. Different scholars from different fields have discussed the characteristics of virtual communities and their similarities and differences when compared to real communities. On the one hand, anthropologists have characterised communities based on having common interests with each other. On the other hand, sociologists have also proposed that communities do not belong to a place with a border; instead, they share the same habitat (Becker, 1984; Wellman and Gulia, 1999). Hillery (1955) put forward 94 definitions of ‘community’; moreover, a clear majority of researchers noted that ‘social interaction’ was an essential element of a community. The community is defined in the social sciences as a

‘multifaceted social relation that develops when people live in the same locality and interact, involuntarily with each other over time’ (Komito, 1998, p.97).

However, when it comes to virtual communities, Haylock and Muscarella (1999, p.73) define a virtual community as a ‘group of individuals who belong to a particular demographic, profession or share a particular personal interest’. Moreover, Wellman and Gulia (1999, p.18) state that ‘virtual communities provide possibilities for reversing the trend of less contact with community members because it is so easy to connect online with scores of people’. This describes the groups of people who communicate via the Internet (Rheingold, 1993; Wellman and Gulia, 1999; Blanchard, 2004). Baym (2007) carried out a study of the multinational online community of independent rock music fans from Sweden. She described the new form of online community as follows: ‘online communities may have more in common with the geographical place-based communities than previous online communities of interest’ (Baym, 2007, n.p.). Online communities are characterised by regular communication around a shared interest, the development of social roles and hierarchies, shared history, and an awareness of differences between other groups (Androutsopoulos, 2006). Oldenburg (1989) remarked that online communities lead to change the ways of communicating by providing them new place to meet people. Further to that Rheingold (1993) suggested that online communities are like social places which fill the need of social places in modern society. Thus, individuals want to construct an online representation of themselves like they do in traditional communities. Donath and boyd (2004) suggested that public displays of themselves serves as identity signals in which people define their identities through the information they provide in the online community. The notion of identity serves an important point to forensic linguistics research also. For instance, Grant and MacLeod (2016) run a research project into the online conversations of sex offenders and the children they abuse, and it was examined theories of idiolect and identity through analysis of the talk of perpetrators of online sexual abuse. In their research they mainly searched for the relationship between linguistic style and online identity performance and the most suitable linguistic analysis to describe an online linguistic persona.

In fact, every virtual community has a specific way of writing or speaking in communication which uniforms of genre types. Generally, the term ‘genre’ is used to refer to ‘a distinctive category of the discourse of any type, spoken or written, with or without literary aspirations’ (Swales, 1990, p.33). Swales (1990) claims that genre is usually named and known by the participants in the culture where it is found; furthermore, genres constitute the ‘production, reproduction, and modification of different types of organisational communication over time

and under different circumstances' (Yates and Orlikowski, 1992, p.301). In addition to the traditional written genres, online genres are structured depending on interaction between community members. Different scholars have investigated online genres and their characteristic types of communication. For example, the genre of digital documents was first introduced by Shepherd and Watters (1998). Later on, most studies in this field discussed the characteristics of web genres such as personal homepages, catalogues, e-mails and so on (Herring et al., 2004; Crowston et al., 2000).

The web is a vast and diverse community, as well as a new virtual environment where interaction happens between members. Because of the web's fluid, unstable and dynamic nature, it is hard to assign a single genre to web pages (Santini, 2007). Furthermore, genres are considered 'as organising structures [that] shape, but do not determine, how community members engage in everyday social interaction' (Yates and Orlikowski, 2002, p.15); this is valid for all genres, including online genres. The question of how community members engage in genres is related to computer-mediated communication.

CMC is produced when individuals interact with each other by transmitting messages via computers or smart gadgets and includes all modes of 'text-based human-human interaction mediated by networked computers or mobile telephony' (Herring, 2007, p.1).

There are many forms of CMC, including emails, shared network group folders, web pages, and discussion boards/forums. Communication of this kind is affected by some circumstantial features, which were classified by Thurlow, Lengel and Tomic (2004, p.32) as follows:

> • the type of channel (e.g. email, web page) and the modes of communication it enables (e g. text-based, graphics-based, audio-visual, or all three);
>
> • the type of participants (e.g. male or female, young or old) and the number of participants (e g. one-to-one, one-to-many, many-to-many);
>
> • the length (e.g. long-term or fleeting) and nature of participants' relationships (e.g. personal or professional);
>
> the topic (e.g. medical advice, romantic dating) and purpose of the exchange (e.g. scholarly, private or commercial);
>
> • whether the interaction is synchronous (i.e. in real time) or asynchronous (i.e. not in real time, with delayed interactions);
>
> • whether it is public or private (e.g. interpersonal, small group, or mass communication) and whether it is moderated (e.g. under someone's direct or indirect supervision) or not;
>
> • what participants' general attitude is towards communication on the Internet (e.g. enthusiastic or sceptical, half-hearted or committed) and how long they have been

> engaging in computer-mediated communication (e.g., are they newcomers or are they experienced?)

As outlined above, there are many features which give rise to different forms of CMC on the Internet. For instance, frequently updated web pages are defined as asynchronous, while chat and instant messaging are synchronous. In synchronous systems, the message exchange is faster than asynchronous online systems but is still slower than the spoken context in real life. Asynchronous online systems do not require that participants be logged on at the same time to send or receive messages; instead, they store the messages at the addressee's site until they can be read (Herring, 2007). The combination of written and spoken features, along with the differences between synchronous and asynchronous modes, constitute the three key issues in CMC, as stated by Herring (1996). Text-based CMC lacks prosodic and nonverbal cues, such as gestures and facial expressions, which contribute to the meaning. As a result, users tend to use recreational chat modes over both systems in order to convey their emotions.

According to Anderson and Kanuka (1997) and Hsiung (2000), asynchronous online forums are observable, easy to use, accessible and safe, unlike synchronous active chats.

Text-based asynchronous and synchronous types of CMC present new avenues for research in different fields including forensic linguistics. Herring (2001) proposed a specific methodology for computer-mediated discourse (CMD), defined as 'the communication produced when human beings interact with one another by transmitting messages via networked computers' (p.612).

The analysis of CMD, known as computer-mediated discourse analysis (CMDA), focuses on 'language and language use in computer networked environments, and by its use of methods of discourse analysis to address that focus' (ibid) from a linguistic perspective and includes 'any analysis of online behaviour that is grounded in empirical, textual observations' (Herring, 2004a, p.339).Herring (2004a) asserts that CMDA can be utilised as a suitable paradigm to facilitate the study of online interactive behaviour at both the macro and micro level of linguistic concepts. She explains the research on four levels, as noted in Table 3-2: (i) structure; (ii) meaning; (iii) interaction; (iv) social behaviour. Herring also provides a theoretical tool for analysis by asking the following questions:

> a) What are the discourse characteristics of a virtual community? b) What causes an online group to become a community? c) What causes a virtual community to die? d) How do virtual communities differ from face-to-face communities? e) What happens to face-to-face communities when they go online? f) In what ways do

communities constituted exclusively online differ from online communities that also meet face-to-face? (Herring, 2004, p.348).

Table 3-2: Four domains of language (Herring, 2004).

	Phenomena	Issues	Methods
Structure	typography, orthography, morphology, syntax, discourse schemata	genre characteristics, orality, efficiency, expressivity, complexity	Structural/Descriptive Linguistics, Text Analysis
Meaning	meaning of words, utterances (speech acts), macrosegments	what the speaker intends, what is accomplished through language	Semantics, Pragmatics
Interaction	turns, sequences, exchanges, threads	interactivity, timing, coherence, interaction as co-constructed, topic development	Conversation Analysis, Ethnomethodology
Social behavior	linguistic expressions of status, conflict, negotiation, face-management, play; discourse styles, etc.	social dynamics, power, influence, identity	Interactional Sociolinguistics, Critical Discourse Analysis

A CMDA approach is not required to answer all of these questions instead mostly related to the pragmatic, theoretical or methodological decisions made by the researcher. This framework 'applied methods adapted from language-focused disciplines such as linguistics, communication, and rhetoric to the analysis of computer-mediated communication' (Herring, 2001, p.2) and also it 'may be supplemented by surveys, interviews, ethnographic observation, or other methods (…) but what defines CMDA at its core is the analysis of logs of verbal interaction (characters, words, utterances, messages, exchanges, threads, archives etc.)' (Herring, 2004, p.2). This framework is grounded in linguistics and provides a methodological toolkit for online language use which demonstrates how CMC should be organised and analysed (Herring, 2007). Although, CMDA approach is focused on four domains of language as structure, meaning, interaction and social behaviour, it is possible to examine the one domain in a language research. For instance, Grant and MacLeod (2016) used social behaviour domain of analysis for CMDA on the online conversations of sex offenders and the children they abuse. The chat logs between the participants are explored by depending on the differing identities which are employed in an interaction. However, in one of the previous studies, Grant and MacLeod (2012) employed taxonomy of structural elements of tweets.

Structure domain of CMDA include ‘emoticons, abbreviation, lexical items (such as personal pronouns), word formations, syntactic patterns and quoting’ (Herring, 2004, p.360) and it is the primary concern in this study on authorship attribution. Computer-mediated language is less correct, complex and coherent than standard written language (Herring, 2004, p.5). The syntactical and semantic features of text allow the researcher to analyse authorship, conduct profiling and guess disputed sources in a forensic linguistics context. There are several advantages afforded by these phenomena if the researchers run online data analysis. As noted above, it is possible to identify some Internet-specific features (such as emoticons or certain abbreviations) that evolved after the invention of the Internet. Eksi Sozluk and Twitter are one of the best sources in CMC media to find Internet-specific features of the language used by Turkish speakers, which are presented in the following sections.

3.2.1. Eksi Sozluk – Online Collaborative Encyclopaedia

According to the Encyclopaedia Britannica, an encyclopaedia is a ‘reference work that contains information on all branches of knowledge or that treats a particular branch of knowledge in a comprehensive manner’ (Encyclopaedia Britannica, 2014). Olkiewicz (1988, p.8) further states that an encyclopaedia ‘emanates the spirit of its time, follows a particular system of values, as it mirrors the views and mentality of its creators, but also of the receivers to whom it is addressed’. There are some major guiding criteria that inform the classification of encyclopaedic works:

> **Scope**: *universal* – encompassing information from all areas of knowledge; *specialised* – concentrating on one discipline or a large concept;
>
> **Size**: large, multi-volume (containing extensive information); small, one volume (with a concise compendium of information);
>
> **Classification**: alphabetical;
>
> **Author**: *collective* – where different authors write the articles; *individual* – being the work of one author.

The collaborative approach to the production of encyclopaedias started in the early 18th century when encyclopaedias began to be written by groups of experts and professionals in different fields (Tereszkiewicz, 2013). In recent years, due to Internet-related developments, some traditional print encyclopaedias have been made available in a digital format. In addition, a new form of user-generated encyclopaedias has also emerged. User-generated collaborative

online encyclopaedias are repositories of encyclopaedic knowledge that are open to contributions from the public and differ from traditional print encyclopaedias (Emigh and Herring, 2005); this is evidence of a great difference regarding formality and contribution. For instance, traditional print encyclopaedias are expert-created, formal and have standardised language, while collaborative encyclopaedias allow everyday users to create a thread or contribute to the repository individually. Since the 'authorship of user-generated content differs from the traditional practices of authorship and covers the reader specialities' (Dogu et al., 2009, p.9), the collaborative structure of the texts gives rise to a new concept of authorship.

The author is defined as a 'person who writes books or the person who wrote a particular book' in the Oxford Advanced Learner's Dictionary (Hornby, 2005). Moreover, the COBUILD English Dictionary (Standop, 1988) states that 'the author of a piece of writing is the person who wrote it'. Goffman (1981, p.144) further explained the roles of the author (animator, author, and principal) in a spoken context; the animator is the sounding box from which utterances come, the author 'someone who has selected the sentiments that are being expressed and the words in which they are encoded', and the principal 'someone who is committed to what to words say'. Sometimes, these roles are played by the same people. In other circumstances—for instance, in media—the author, principal and animator might be different individuals; for example, the authors could be screenwriters, the principals could be producers, and the animator could be a singer or actor.

However, in a writing setting, Love (2002) asks what authorship means, describing authorship as a form of human work that validates individual agency and is 'always a textual performance' (Love, 2002). In addition, the functions of authorship need to be defined during the creation of the work rather than as a single, coherent activity (Love, 2002) and should be regarded as precursory, executive, declarative and revisionary.

Precursory Authorship is the function whereby 'a significant contribution from an earlier writer is incorporated into the new work' (Love, 2002, p.40). In other words, a previous text might be a source of inspiration or source for a later one. The second function is *Executive Authorship*; in Love's terminology, the executive author acts as the wordsmith, or the person who makes the actual text ready for publication (Love, 2002, p.43). The wordsmith could be a range of possible writers. The executive author may be the orderer, the deviser, the wordsmith and the reformulator, and the function may be performed solo or collaboratively (ibid).

The other function is *Declarative Authorship,* which places the declarative author's name to the text, after which the author will use their words and take the responsibility (Love, 2002, p.45). Press releases from companies or politicians may count as a good example of this function.

Revisionary Authorship is the function of the second writer, who rearranges or edits the text (whether individually or collaboratively) to make a final copy (Love, 2002, p.47). Functions of authorship can be applied to press releases, business letters, columns or collaborative Internet genres.

Wikipedia is one striking example to reflect on when considering the functions of authorship. It is open-source, the most widely-known wiki site, and a multi-language encyclopaedia designed 'to be read and edited by anyone'. Its *collaborative writing* and *peer collaboration* functions refer 'to projects where written works are created by many people together (collaboratively) rather than individually' (Wikipedia, 2018). On a wiki site, anyone who reads a page can also edit it with a click, and the history of the changes can be tracked by anyone who is interested in the edits. Authorship on Wikipedia is revisionary authorship according to Love (2002) since the author only edits and rearranges the text. Furthermore, Warschauer and Grimes (2007) discussed authorship on Web 2.0 platforms, especially on wikis, blogs and social networking websites, stating that the authors of wikis are 'not more than organisers of content' (ibid, p.9) and evaluating this as the fading away of prophecy authorship. However, for earlier versions of texts on Wikipedia, it is possible to exercise the precursory and executive authorship functions (rather than simple content organisation). For instance, one can write a text as an executive author, or add some sources into previously written texts as a precursory author, while another person may edit the texts as a revisionary author. That is to say, the functions of authorship may overlap in some contexts (Grant, 2008).

Wikipedia is the fifth most visited website in the world (Alexa, 2018). There are 301 Wikipedias in various languages, of which 291 are active, and 10 are not (Wikipedia, 2018). Currently, its English version has over 5,648,884 articles of any length; when these are combined with those in all other languages, the number exceeds more than 40 million articles (Wikicount, 2018).

Regardless of the accuracy of its content and reliability of its information, Wikipedia is highly useful for Internet users due to its ease of access, it's being free of charge, and the opportunity it provides to read content in a user's native language. Because of its live structure, Wikipedia has attracted the interest of many researchers in different fields, including authorship attribution

studies, due to its collaborative writing and editing functions (see, e.g. Paul et al., 2018). Emigh and Herring (2005) conducted a comparative study between Columbia Encyclopaedia (a traditional encyclopaedia) and two online encyclopaedias (Wikipedia and Everything2) using corpus linguistic methods and factor analysis of word count to examine features of formality and informality.

Moreover, Bell (2007) compared Wikipedia with the Online Encyclopaedia Britannica, looking at readability, syntax and use of factual measures. The minimal difference was found between Wikipedia and Encyclopaedia Britannica, despite the huge difference in the systems. Similarly, Elia (2009) compared Wikipedia and Britannica entries in terms of type/token ratio, word and sentence length and index of readability. This study showed that 'Wikipedia is not statistically distinguishable from Columbia Encyclopaedia in some features' (Elia, 2009, p.267).

Eksi Sozluk, which is the first Turkish online collaborative encyclopaedia similar to Wikipedia, is a collection of user-created text-based content posted on different threads and based on the concept of websites built upon user contribution. Structurally, Eksi Sozluk is among the group of new encyclopaedias (according to Tereszkiewicz's (2013) taxonomy) in terms of being a user-created portal; moreover, its content is not limited to catholic materials but includes different types of articles on culture, sports and social life. Since it is mainly used for entertainment purposes, it is not intended for training purposes or as an academic reference; besides, the authors do not claim expertise or to provide accurate information on anything. Consequently, people do not generally consider the source trustworthy. In addition, all kinds of forms of writing are allowed, and it is not compulsory to use a neutral point of view, unlike Wikipedia.

The origin of the site's name comes from Portishead's song 'Sour Times' and the heart of Eksi Sozluk is its entries, which are updated continuously by visiting authors. The entries are posted publicly available, and authors can read other entries or edit their contributions. It is important to note that each entry is written by a single author. Users need to wait for a certain period of time before they can become an author, after which the users become a part of the online community and can start adding entries.

Eksi Sozluk entries are posted as single units and do not necessarily refer to each other; thus, they can be referred to as *self-contained texts.* Moreover, the standard length of the texts is not controlled (unlike other types of online data such as Twitter, which only allows 280 characters, or Facebook, which allows 63206 characters). Entries might include quoted passages, real-life

experiences or descriptive information. It is also possible to rate the entries; there are two buttons for voting, and anyone who reads an entry can vote for it using the buttons to vote up or down regardless of their membership position.

Figure 3-2: Image of the emergent, user-created online encyclopaedia Eksi Sozluk.

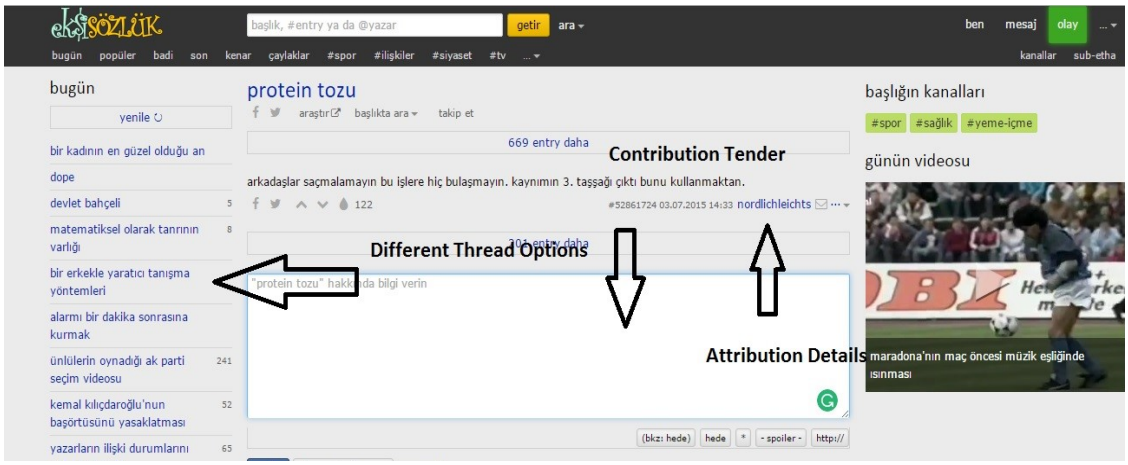


Figure 3-2 shows the typical entrance page of Eksi Sozluk. The featured titles and the first page have a consistent texture. In order to insert text, the contribution tender is used. Beside the contribution tender, metadata such as nickname, date, time etc. are displayed. The main word or phrase appears in the header of the page, along with the descriptions. Pages may number from 1-1000 depending on the popularity of the entry. For instance, the 'Recep Tayyip Erdogan' entry has 1862 pages of definitions, while the 'kedi' (cat) entry has 666 pages. The remainder of the page consists of a sidebar containing the other top entries of the day.

Moreover, as can be seen from the different thread options in Figure 3-2, the physical state of the website has a blog-like look with a chronological order. Thus, current threads are found on the first page, with other threads found on the following pages.

In 2016, more than 192 million people visited Eksi Sozluk (Eksi Sozluk, 2016). It is one of the 20 most visited websites in Turkey, and its global ranking is situated at 734 (Alexa, 2018). Although there is no clear evidence about the characteristics of Eksi Sozluk authors as regards their age, sex or socio-cultural status, some statistics websites (e.g. Alexa, 2018) provide an insight into the site's unique Internet visitors, along with their demographic features.

Figure 3-3: Percentage of visitors to Eksi Sozluk by country.

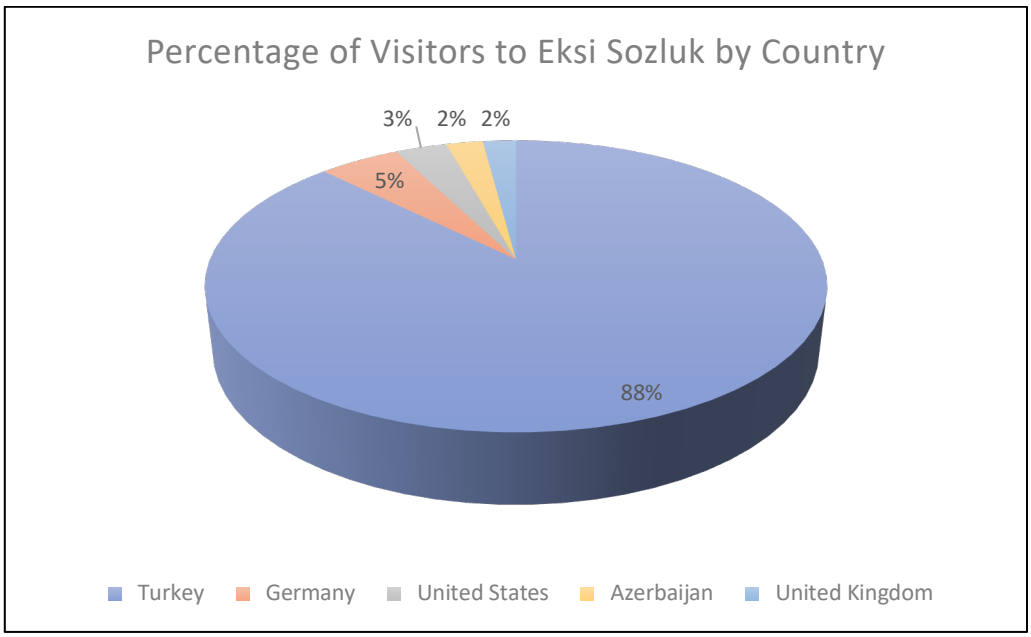
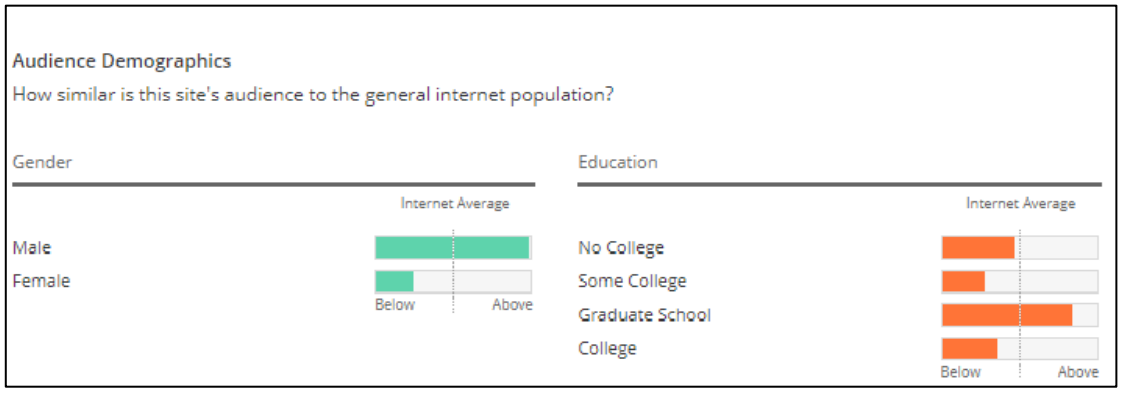


Figure 3-3 presents the distribution of the visitors' countries. The highest percentage of visitors are from Turkey, at 88%; the USA and Germany follow with 5%. Visitors from the remaining countries make up between 3% and 2%.

Figure 3-4: Audience demographics of Eksi Sozluk.



In Figure 3-4, the Eksi Sozluk audience population information from Alexa (2018) is compared with the general Internet population. According to the data, the male population is significantly higher, while females are under-represented. Moreover, the population who went to graduate school are highly represented in Eksi Sozluk.

It can also be seen that the online community of authors is heterogeneous in terms of their education level, as there are people who went to graduate school, only college or no college at all.

The diversity of the backgrounds of Eksi Sozluk users has an impact on the mixed presentation of the language and its various linguistic styles, which are balanced between spoken and written language. The language used on Eksi Sozluk features certain characteristics that are not restricted to conventional writing; this is because authors are not expected to follow conventional language rules, but rather to provide quality information. Thus, it is possible to find spoken language features; however, some entries use an informal style of writing that presents a flexible, rich style including nonstandard usages or some Internet-specific features. By contrast, Wikipedia promotes a formal, standardised language in order to establish credibility by supporting content with scientific references; this is not a necessity for Eksi Sozluk.

In addition, the texts are only under the control of those authors who are executive authors. Eksi Sozluk differs from Wikipedia in this way, since a standard Wikipedia article may be edited hundreds of times by different authors. In this sense, Wikipedia fulfils the 'collaborative' notion entirely; however, Eksi Sozluk is characterised by individual descriptions being written by many authors for any given entry. In other words, while Wikipedia involves large collaborative efforts to feature the best article, Eksi Sozluk features a large number of contributions on the same topic in an effort to identify the best source.

3.2.2. Twitter User-Generated Content

Twitter is a worldwide micro-blogging website that allows users to post messages of 280 characters or less in length (Twitter Blog, 2017). Microblogging is defined as 'a variant of blogging which allows users to quickly post short updates, providing an innovative communication method that can be seen as a hybrid of blogging, instant messaging, social networking and status notifications' (Ross et al., 2011); it is also perceived as an online genre that is not stable and may incorporate characteristics of other genres. As stated by Giltrow and Stein (2009, p.2) 'the CMC text… trespasses its borders, creating vectors on which genres can travel into one another's territories'.

Twitter features some unique communication mechanisms whereby users employ some specific provided orthographies to repost, highlight a topic and mention a user. Firstly, 'RT'

(commonly referred to as 'retweeting') is to spread the contents of a post from its original account to that of a user, while the dominant form is to use a 'hashtag', a new tagging format that associates a user-created tag with an event or a context via a prefix symbol (#) used to categorise topics (Chang, 2010). While retweeting 'allows members to relay or forward a tweet through their network' (Nagarajan et al. 2010, p.295), hashtags are form of conversational tagging which refers the topic of a tweet. Retweeting contributes 'to a conversational ecology in which conversations are composed of a public interplay of voices that give rise to an emotional sense of shared conversational context' (boyd, Golder and Lotan, 2010, p.1). Zappavigna (2012) defined the use of hashtags 'where the primary function appears to be affiliation via findability' (p.789). Moreover, such type of collaborative tagging provides the users to search for a specific subject in a tweet. However, same hashtags can appear in tweets unrelated topics for instance, when Black Mirror Bandersnatch was released many irrelevant tweets appeared simultaneously to raise a concern about current political crisis.

Finally, the @ symbol refers to 'mentioning' a tweet using another user's nickname. It is used to tweet a message to a specific user only. Figure 3-5 presents a tweet series that includes the hashtag and @mention orthographic examples. According to Honeycutt and Herring (2009) @ symbol helps to capture the attention of the addressee. Later on, Zappavigna (2012) remarked that mentions provide a way to present other voices into tweets. A set of example tweets are presented in Figure 3-5 which include hashtag and mention features.

Figure 3-5: A display of tweets.

As seen on Figure 3-5, each tweet has one main body containing the message itself and some complex features such as user name, geographical information and the device source. Here, it

is shown the most important features are presented the message, retweeting and mentioning functions. Some other useful metadata information is also available such as *retweet status* which shows the number of retweeting, *tweet created at* is the date when the user tweet the message and *the number of likes* shows the popularity of the message. However, these features are not related to the authorship attribution analysis in the current study.

Twitter users have two networks as following and followers however they can only receive updates from the persons they follow. The number of following and followers are not necessarily equal. One may follow more than a thousand accounts while having only a few followers. When the account is not protected an authorization is required from the account owner, and the account holder have right to make their account protected otherwise all tweets are to be publicly visible. Twitter messages are made not 'the transfer of information or status messages that are crucial factors, but rather, the opportunity to be part of someone else's process by reading, commenting, discussing or simply enhancing it' (Ebner et al, 2010 p.98). Such features make Twitter different than many other CMC mediums for that reason, it attracts many researcher's interest.

Several previous studies have analysed the ways in which people use Twitter (e.g. Honeycutt and Herring, 2009). Williams et al. (2013) classified Twitter-related academic papers and created an extensive literature review. In the end, 575 papers were identified that focused on Twitter research from different fields. Considering the popularity of Twitter, it is likely that more studies have been conducted over the past five years on top of these 575 papers. However, the majority of these articles focused on general information on Twitter as a social medium and the rest is focused on the use of Twitter in online higher education.

Recently, Hu et al. (2013) collected over 45 million tweets in Oct 2012 in order to gain a deeper understanding of the language of Twitter, presenting a set of linguistic features by which 'Twitter language' may be quantified:

- Twitter, in general, is surprisingly more conservative, formal and less conversational than SMS and online chat, although it shares similar brevity and interactivity. Its primary usage is to convey information (either for sharing news or broadcasting self-status).
- Twitter users appear to be developing linguistically unique styles when compared against other mediums—for example, both first-person and third-person pronouns are extensively used, whereas other mediums tend to stick to one type of pronoun.

- We find that Twitter exhibits usage of temporal references that are similar to those used in SMS and online chat.
- Twitter has less variations of affect when compared to email, blog, slate and news, and it tends more toward positive moods than other mediums (p. 245).

While the set of research using Twitter as data has been expanding rapidly, there is still a limited number of articles focused on using Twitter in forensic linguistics purposes. Moreover, when it is considered the differences and similarities between the other CMC mediums and Twitter it offers big opportunities to explore authorship attribution studies. Because of its massive popularity, the language on Twitter has a particular emphasis on the Internet language. That is to say, it covers a massive scale of information sharing and informal writing style. According to Mischaud (2007), Twitter users have 'appropriated this medium to reflect whatever use or style of communication they want' (p.38). Despite there is a character limit of the tweets, many account holders use a creative way of language. Namely, the texts contained in tweets is very different from standard languages which may contain misspellings, different punctuation, emoticons, hypertext links and more. Such an attitude is similar to the Eksi Sozluk even though there is no brevity force of the language in Eksi Sozluk. Crystal (2011) mentioned that non-standard spelling of shortenings is common in Twitter however, in Eksi Sozluk there is no need to shorten the words. The entries in Eksi Sozluk also reflect the author's style without any style limitations thus both contents contribute to establishing the identity of the authors and makes Twitter as an effective environment for linguistic research and also forensic linguistics. Structurally, Twitter have some common features with Eksi Sozluk: metadata (author's name which is a pseudonymous in Eksi Sozluk, time of posting and sharing options). Even though users can use only pseudonymous, most of the Twitter user share their real identity.

Overall, in this section the distinctive features of Twitter from other CMCs are outlined, focusing on the structural features and the language used in the platform. Since Twitter provide a great sample of language data due to its non-stop activity, it can be also a fruitful source for forensic linguistics.

Chapter 4: Methodology

The research methods employed in this study are presented in this chapter. The aim of the chapter is to generate a research framework for forensic authorship attribution. Section 4.1 gives the theoretical background to the study; Section 4.2 explains the data-collection criteria and the corpus design; Section 4.3 gives the characteristics of each corpus; Section 4.4 outlines the methodological approach applied; Section 4.5 gives the statistical design; and Section 4.6 raises a number of ethical considerations.

Online crime is a growing area, and many governments, including in the USA and the UK, are using forensic linguistics as a way of solving cases. Furthermore, awareness of forensic authorship attribution cases has risen since the explosion in social media as part of computer-mediated communication (CMC), as well as popular thriller books and television series. One of the popular law forum websites in Turkey, 'hukuki.net' (forensics.net), includes many questions about anonymous fraud or threat letters; hence, the need to consult a forensic linguist is increasing. However, few attempts to attribute authorship based on computational methods have been made in the Turkish language (see Section 2.8.). Those attempts raise the question of which features and approaches are better for authorship attribution studies in Turkish. Therefore, a suitable methodology that includes computational and forensic linguistic methods is needed for Turkish studies. As Grant (2008, p. 216) stated, 'there is no single question of authorship analysis', and for this reason 'there can be no single technique that should be universally adopted'. More research is needed in forensic linguistics in every language and for every genre.

On the other hand, in forensic linguistic studies, real-world data are generally inaccessible unless the person is an expert giving an opinion on a case. Even in real-world cases, 'finding comparable material from the suspected potential writers is not always a simple matter' (Johnson and Woolls, 2010, p. 114). For that reason, a tailor-made authorship attribution scenario is created for this study. The data set has the advantage of an informal style that is consistent with real-life data, which was explained in detail in Chapter 3. The theoretical background of the study is presented in the following section.

4.1. Theoretical background

This study aims to identify the author of anonymous text or texts where there is a limited number of possible closed set authors (Solan and Tiersma, 2005), depending on 'consistency and distinctiveness' principles (Grant, 2013). Grant (2013) developed a feature-based approach that depends on quantitative and statistical methods using lexical choices, punctuation, and shortenings in the short text messages and allied genres. Grant (2013, pp. 473–474) also proposed that all comparative authorship analysis methods are based on two assumptions: (i) 'there is a sufficient degree of consistency of style within relevant texts' and (ii) 'there may be a degree of distinctiveness between pairs of individuals or within smaller or larger groups'. Accordingly, this study, based on the theoretical assumptions, has two hypotheses:

- An author/authors' writing has a sufficient degree of consistency of style with the relevant texts.
- An author/authors' writing is sufficiently distinctive from other relevant authors.

Similar to Grant (2013), McMenamin (2002) has used the term *resemblance* with the same meaning as Grant's *distinctiveness*. Although, there are other terms that correspond with the same assumptions, *consistency* and *distinctiveness* are the terms used in this study. As stated by Grant (2013), it is possible to apply this approach to other genres on the Internet that have the same characteristics as short text messages. Eksi Sozluk is a different genre from the SMS and Twitter but it is an opportunity to check whether this approach 'may be generalized to other text types and other features' (Grant 2013, p. 472) According to Grant (2008), the question that should be asked in a comparative authorship analysis is as follows: *What is the relationship of a text to the comparison text?* This question is investigated in two phases with the aid of selected corpora based on consistency and distinctiveness principles. First, text vs. text comparison is applied in which each text is treated individually from the same author and unknown texts are attributed to one of a set of known authors. Second, in author vs. author comparison it is clustered the texts and it is measured author similarity by looking at how many authors have the closest distance between each other.

The above section has established the theoretical grounds for the research; the following section presents the corpus design and the controlled criteria for data collection.

4.2. Data collection and corpus design

Corpus linguistics is defined as 'a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study' (Hunston, 2002, pp. 1–2). As pointed out by Hunston (2002), the use of corpus linguistics and corpora revolutionised the study of language, and it has provided a more unbiased view of language than did instinct or anecdotes.

Corpus linguistics has several ranges of use, which are for language teaching, translation studies, general corpora to establish frequency and usage, sociolinguistics, discourse analysis studies, and forensic linguistics. In terms of forensic linguistics, 'corpus linguistics not only provides a new kind of data but also new ways of analysing the data on which forensic linguistics has traditionally concentrated' (Coulthard, 1994, p. 27). Coulthard (p. 40) also pointed out that in the situation of corpora for forensic linguistics, 'any improved methodology must depend to a large extent, on the setting up and analysis of corpora'. Additionally, Cotterill (2010, p. 578) pointed out, 'the issues of authorship and plagiarism are now growing matters within the field of forensic linguistics, for which corpora can prove a useful instrument of investigation'.

In general, 'any collection of more than one text can be called a corpus' (McEnery and Wilson, 1996, p. 21). In her study, Tognini-Bonelli (2001, p. 55) remarks that a corpus is 'a computerised collection of authentic texts, amenable to automatic or semi-automatic processing or analysis. The texts are selected according to explicit criteria to capture the regularities of a language, a language variety or a sub-language'. Some scholars have accepted corpus linguistics as a 'new physiological approach' (Leech, 1992, p.106), rather than 'a whole system of methods and principles of how to apply corpora in language studies' (McEnery et al., 2006).

Furthermore, *Corpus-based* and *corpus-driven* are terms introduced to corpus studies by Tognini-Bonelli (2001). Corpus-based study is used to 'refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before' (p. 65) whilst corpus-driven is spoken of when 'the commitment of the linguist is to the integrity of the data as a whole and descriptions aim to be comprehensive with respect to the corpus evidence' (p. 85).

According to Biber, Conrad and Reppen (1998, p. 4), the essential characteristics of corpus-based analysis are:

- it is empirical, analysing the actual patterns of use in natural texts;
- it utilises a large and principled collection of natural texts, known as a 'corpus,' as the basis for analysis;
- it makes extensive use of computers for analysis, using both automatic and interactive techniques; and
- it depends on both quantitative and qualitative analytical techniques.

Tognini-Bonelli (2010) and Gilquin and Gries (2009) also used corpus-driven to 'approach corpus data in an exploratory fashion, i.e. without rigorously formulated hypotheses', treating it 'as a methodological term, synonymous with 'bottom-up' (Gilquin and Gries (2009, p. 10).

In the bottom-up model, the researcher searches for the existence of patterns, units or some formulations related to the author's style, while a top-down model refers to the predefined stylistic items that would show the author's discriminative style (McMenamin, 2002, p. 54). With the top-down approach, the feature categories are predefined before being determined, whilst in the bottom-up perspective individual language and data-driven features can be mapped. One way to obtain reliable results when treating both data and theory is to combine bottom-up and top-down approaches. In other words, 'combining two approaches makes it possible to put data and theory on an equal footing' (Gilquin, 2010, p. 10).

Contrary to corpus-based and corpus-driven concepts, McEnery and Hardie (2012, pp. 148–149) stated that:

> While corpus-as-theory rejects any explanation of language patterns that does not derive from the analyst's interaction with the data, corpus-as-method considers corpora and corpus techniques to be sources of empirical data that may be deployed in support or refutation of any explanatory theory about language – even a theory devised in whole or in part without reference to corpus data.

After considering all these concepts and methods, there is an ambiguity in the concepts. Therefore, both corpus-as-theory and corpus-as-method analytical approaches are used, instead of using corpus-based and corpus-driven notions. As this study aims to test the corpus with the linguistic features, it can be said to take a corpus-based approach initially, or a corpus-as-method approach. In one sense, this study is corpus-based, which is why it is essential to state Biber et al.'s (1998, p. 4) approach to quantitative and qualitative techniques in corpus research as 'association patterns represent quantitative relations, measuring the extent to which features, and variants are associated with the contextual factor. However qualitative interpretation is also an essential step in any corpus-based analysis'.

Furthermore, designing a corpus for linguistic analysis accentuates some key points under four headings: sampling and representativeness; size; machine-readable form; and standard reference (McEnery and Wilson, 2001).

First, sampling and representativeness refer to 'the whole variety of a language rather than an individual text or author' and 'corpora are generally large, representative samples of a particular type of naturally occurring language', as in linguistic studies (Baker 2006, p. 2).

In general, the corpus represents the language variety as a standard reference. Careful design and creation of sampling criteria are essential for triangulating the study. However, this kind of corpora is related to general language use, while current corpora in this study are determined to represent a specific aspect of language that is developed for answering particular questions. Moreover, the size of the corpus depends on practical considerations and research focus (McEnery, Xiao and Tono, 2006). It is increasingly common that corpora should not be massive, as even small corpora can still provide useful and reliable results about the study. Kennedy (1998) argues that corpora of fewer than one million words can still offer useful insights to the discipline. This study draws on the impact of corpora with regard to its size and range. Although the corpus of the study is small when it is compared with stylometric studies, it is still significant enough to issue meaningful linguistic and statistical results in attributing authorship in Turkish. Also, in forensic linguistics, many real casework texts are concise, such as threat letters, suicide notes and Twitter posts, in an online setting. Finally, 'machine-readable form' (digital) refers to the availability of corpus online instead of paper print versions.

Beyond that, the Internet offers a huge corpus for language studies since 'it offers a home to all languages – as soon as their communities have a functioning computer technology' (Crystal 2006, p. 229). Furthermore, there are two main principles in building corpus from the Internet – first 'searching the whole web through a commercial engine and collecting the pages from the web (randomly or controlled)' and 'searching them locally' (Lüdeling et al., 2007, p. 8). The function of the web in this study is to create a corpus by searching locally from both websites. In building a corpus, Eksi Sozluk and Twitter provide an enormous database of many words and phrases that are authentic examples of the Turkish language. It is worth noting that, although it is a rich source in terms of language data, Twitter is not taken as the main corpus for this study; instead, it provides supplementary data for cross-genre comparison corpus. Despite Twitter has less impact than Eksi Sozluk in this study, it is an important language source since it provides an avenue to collect a large and diverse linguistic content. The

uninterrupted frequent interactions in Twitter leads to generate big amount of data in different topics.

As a matter of fact, using the Internet to create a corpus has many advantages for such kinds of studies. Fletcher (2012) specified the benefits of web corpus as being: (i) size, (ii) range, (iii) up-to-dateness, (iv) multimodality, and (v) availability.

In 2016, more than 192 million people (Alexa, 2018) visited Eksi Sozluk for both writing and reading purposes. The site is accessible 24 hours a day, seven days a week. This makes the website live all the time and the data are growing larger and larger day by day, which may eventually make the site unmanageable. Along with its availability and popularity, its existence is widely known. Nevertheless, its reputation does not depend only on its availability, range or up-to-dateness. The reason is the content of Eksi Sozluk, which generally targets a large segment of society. For instance, on 14 January 2014 the Radikal newspaper reported that 39 authors had been accused of writing harmful content and expressing 'humiliating religious values', and some other cases have been submitted to civil courts related to the content of Eksi Sozluk.

To the researcher's best knowledge, Eksi Sozluk does not share statistics related to daily entry numbers; however, as a member of Eksi Sozluk's virtual community it is possible to observe the non-stop writing activity. Despite the advantage of the web data in terms of size, it causes a reproducibility problem. Lüdeling et al. (2007) mentions reproducibility as an important aspect of using web corpora:

> … the web is constantly in flux, and so are the databases of all commercial search engines. Therefore, it is impossible to replicate an experiment in an exact way at a later time. Some pages will have been added, some updated, and some deleted since the original experiment. In addition, the indexing and search strategies of a commercial engine may be modified at any time without notice (p. 11).

However, since this study is focused on corpus-as-method rather than corpus-as-theory research, replicability of the analysis in terms of statistics and feature extraction has more importance than replicating the corpus.

There is one issue that does not correspond with the advantages of the web corpus. Multimodality refers to several modes of communication in terms of visual resources; however, Eksi Sozluk is a monotype website that does not allow visual resources to be included in the entry, apart from hyperlinks.

Although Twitter is different from Eksi Sozluk, the language data are the only core material for this study; thus, multimodality in Twitter does not cause a difference to the overall aims. In the following section, the criteria for creating a forensic linguistics scenario are presented in accordance with the research aims of this study.

4.2.1. Criteria for data collection

In the field of forensic linguistics, some corpora are available for forensic analysis itself, such as the case of Timothy Evans, described by Jan Svartvik, or the case of Derek Bentley described by Malcolm Coulthard (1994). However, as stated by Crystal (2011, p. 123), 'it is hard to obtain samples of authentic data to analyse to provide norms. This is a regular problem in forensic linguistics'. Similarly, although in a typical forensic scenario the analyst has access to a handful of texts of questioned authorship and several texts of one or more candidates (Hyland et al. 2012), it is a problem to find appropriate material and, in particular, appropriate comparative material (Johnson and Woolls, 2010). Considering the limitation of obtaining real-world data, an invented scenario that includes a synthetic corpus may be suitable for forensic linguistics research. Yannikos et al. (2013, p. 322) stated that:

> Unlike real-world corpora, synthetic corpora provide ground-truth data that is very important in digital forensics education and research ... The ability of the framework to generate synthetic corpora based on realistic scenarios can satisfy the need for test data in applications for which suitable real-world data corpora are not available.

In compiling a suitable corpus in terms of forensic linguistics, an authorship attribution scenario is generated from Eksi Sozluk and Twitter, which includes Internet language and language-specific features that are likely to occur in real-life cases. Scenario-based studies enable the linguistic features that are related to authorship attribution to be controlled and observed.

In this study, four specialised corpora were created after 2013 and 2014 data from Eksi Sozluk were used for analysis. The importance of the years 2013 and 2014 comes from the Gezi Park protests. During and after the Gezi Park protests in 2013 and 2014, social media usage had a breakthrough in Turkey. Protests started as a small social movement against the ruling party AKP (Adalet ve Kalkinma Partisi – Justice and Welfare Party). Although this movement began in Istanbul's Gezi Park, it expanded nationwide, largely due to social media. Eksi Sozluk was

one of the most effective forms of CMC media during the protests; for this reason, this study decided to collect texts that were produced during those years.

Grieve (2007) has pointed out that author-based corpora should contain 'questioned documents' and 'known documents' produced 'around the same point in time' (p. 255). In his corpus, he selected texts that were built over a five-year period. It is worth noting that, unlike Grieve (2007), questioned documents are called disputed texts in this study.

As Herring (2004) pointed out, in virtual community research, the richest possible context is required, and it should include periodic time-based sampling – for instance, several weeks at a time at regular intervals over a year. Therefore, for this research, texts produced in the same period can justifiably be considered to have similar composition dates. A purposive sampling method was taken for the collection of the data. Aarts and Meijs (1990, p. 22) proposed two possible methods for creating a corpus. The first one is 'to take whatever you can take' and the second one is to 'make a careful selection of texts'. As mentioned before, the computer-mediated media devices are always in motion; it is not possible to collect data randomly regardless of genre, format, size etc. Therefore, purposive sampling makes sense for this research rather than random sampling. For instance, many researchers in the field of CMC limit their sampling by a thread (e.g. Herring, Johnson and DiBenedetto, 1992), by a period, by demographics or by gender. However, none of those is a concern of this study, but size and written time are one of the central concepts of data collection along with some criteria depending on internal and external factors.

First, size is the most important issue to validate the authorship attribution methods in the Turkish language. For that reason, four different corpora created to discover the factors behind different sizes. Time is the second concern of the purposive sampling in this study, as it is mentioned above author-based corpora should consist of texts which were written around the same time. The purposive sampling approach allows sampling quickly a list of authors from both sources (Eksi Sozluk and Twitter) however, due to the nature of the study some authors left out of the sample depending on the questionnaire results which is explained in the following pages. Although the expectation in purposive sampling is to collect linguistically homogenous data, in this study size and time depending approach leads a more linguistically diverse set of authors.

The following criteria were used to collect the data, based on internal and external factors, which were controllable or not. External factors are mostly related to the characteristic of the website and the internal factors correspond to a set of criteria for author characteristics.

The internal factors concern the authors and include:

- social distinction – such as people are from different genders, from various age groups, from different education or ethnic backgrounds. In this corpus, the author's sex and gender are not known information since authors do not have a detailed personal profile page in Eksi Sozluk
- the frequency of posting: infrequent users (one or fewer posts in a week); daily users (one or two posts a day); frequent users (more than one or two posts a day)
- multiple authorship
- autocorrect on mobile devices
- bilingual influence – mainly Kurdish effect; the levels of familiarity with the language such as Turkish as native language vs Kurdish as a second language or Kurdish as a native language vs Turkish as a second language.

On the other hand, external factors correspond to: (i) genre, (ii) the website location and (iii) the format of the website. In this study, the effect of these factors is investigated through questionnaires. According to Brace (2008), there are two key tests for a questionnaire: reliability and validity. A questionnaire is reliable when there is a 'consistent distribution of responses from the same survey universe every time' (p. 276). Validity is related to whether or not it measures what we want it to measure. Brace also advises administering the questionnaire twice, at different times, in order to determine the reliability of the answers.

However, in this study questionnaire results are not the core of the study; instead, they are the first step in data collection. Moreover, the questions are generally based on mostly closed-end questions and the responses do not vary at different times. Testing a validity requires that 'we ask whether the questions posed adequately address the objectives of the study' (Brace, 2008, p. 277). Collecting data from online sources causes some inconveniences such as selection bias and a third-person effect, which may affect validity and reliability. When it comes to selection bias, all the authors who have the same prerequisite for the suitability of the research are equal to the researcher since the study does not require face-to-face interaction or gathering data via questionnaires. Thus, the questionnaire used in this study is both reliable and valid, according to Brace's (2008) definition, and the results of the questionnaire have established criteria that are in accord with the research aims.

At the first stage, the authors who had at least 1000 entries each were selected as potential authors for this study. Before the authors were chosen, a questionnaire was sent to them in order to check their appropriateness for the research.

For the protection of the authors, since many of them had authored politically sensitive content under pseudonyms or nicknames, the authors' names were replaced with numbers.

The semi-structured questionnaire sent to the respondents asked the following questions:

Part I

1) What is your native language?
    a. Turkish
    b. Fluent in Turkish/Bilingual (Native Language, e.g. Kurdish, Arabic, Persian, Bulgarian etc.)
    c. Other

   If your answer is *b,* please continue from the second question, if not please skip the following question and proceed from the third question.

2) Please choose a number from 1 to 5 in order to describe your proficiency in Turkish: 1 Not fluent; 5 Native-speaker level
3) How do you describe your authorship style in Eksi Sozluk?
    a. Text originator
    b. Multiple users for one account
    c. Known nickname but no one can access
4) Do you use the editing option often? If yes, what is your purpose in using it?
    a. Correcting sentence structure (e.g. grammar, vocabulary, punctuation)
    b. Adding extra information

Part II

5) What kind of device do you generally use while writing your entries?
    a. Desktop/Laptop
    b. Mobile Devices/Tablet
    c. Both
6) Do you use a spell-checker on your device?
    a. Yes
    b. No
7) What is your frequency of posting?

a. Infrequent user – one or fewer entries a day
b. Daily user – around one or two entries a day
c. Frequent user – more than one or two entries a day

This questionnaire was created in two parts using an online survey web tool (www.surveyplanet.com) with a link sent to 135 suitable authors who had at least 1000 entries. Some of them were not willing to share their entries with the researcher. Some of them agreed to participate but either did not answer the questionnaire questions or did not reply to the messages on time. It is worth to note that, the same questionnaire is not applied to the Twitter users since the users either linked their Twitter account with the Eksi Sozluk account or share their Twitter name under the 'Twitter accounts of the authors' title. In Eksi Sozluk, connecting both accounts are not a common attitude due to the urge of anonymity, the ones who share their Twitter address publicly is a sign of a frequent writer in both mediums. In such a case, it was not necessary to check the author's frequency on Twitter by applying the questionnaire to the participant.

In the end, 75 questionnaires were ready to check the corpus-compiling criteria for the current study. Since this study examined original Turkish texts, the first question was of vital importance. No one answered 'other', but 32 authors selected the 'Fluent in Turkish/Bilingual' option.

Among the other minorities in Turkey, Kurdish people have the highest population. The population of Turkey is approximately 79.50 million. According to a national poll named 'Who Are We?', Turkish is the mother tongue for 84.54% of residents. Kurdish is the second one at 11.97%; besides, 9.76% of Kurdish speakers use their language in their daily life, which means 2.21% of people speak a language other than Kurdish in their daily life (Konda, 2007). Kurdish is not accepted as an official language yet. Thus, Turkish citizens whose mother tongue is not Turkish are educated in the Turkish language when they enrol for school under the Ministry of National Education. Kurdish people have had to develop proficiency in Turkish in order to integrate into the Turkish-dominated system (Polat, 2007).

Roughly, it can be estimated that most Kurdish native speakers can speak advance level Turkish. However, school education may or may not affect the pupils, as it is expected that their proficiency in Turkish remains low. Even under these conditions, one can still get an account from Eksi Sozluk. For that reason, the second question is asked in order to grade their

level in Turkish from 1 (not fluent in Turkish) to 5 (native speaker level). Three out of 13 authors returned the answer 4, which means a high level of proficiency but not native speaker level. By taking this situation into consideration, these three authors were excluded from the author's list since the aim is to have Turkish speakers only for the study.

Later on, a question related to their authorship style is asked. To the researcher's best knowledge, based on experience on Eksi Sozluk as an author and a reader for more than 10 years, for some keeping their nickname a secret is an essential requirement for the people who are interested in Eksi Sozluk. There are many authors who hide their nicknames even from their partners or spouses. Since this situation cannot be generalised for all, the third question is asked to eliminate multiple users for the one account. This is one of the most important questions because at the core of the analyses it is hypothesised that one author is the only author and his/her style is consistent within the texts.

Twenty-six out of 132 authors returned with 'known nickname, but no access to it' and five authors chose multiple users for one account. Thus, five account holders were removed from the list.

In the last question of the first part, the reason for using the editing option is asked. Forty-seven authors chose the 'to correct sentence structures' option; 31 authors chose to add extra information as an afterthought to the entry.

The importance of non-edited texts was stated in a study of eight Dutch university students – some linguistic features carry important clues, provided the texts have not been subjected to editing (Baayen et al., 2002). For that reason, the authors who correct sentence structure were removed from the list to obtain not-planned and not-proofread texts in order to have similar conditions to real-life forensic cases. At the end of the first part, only 80 authors were left out of a total of 135 authors. The second part only concerns the 80 authors.

The second part of the questionnaire aims to understand the device type, spell-checker usage, and website-visiting habits. According to the results, the device type and the spell-checker usage are related to each other. For instance, 24 out of 80 authors chose 'mobile devices/tablet' and 24 out of 80 authors chose 'yes' to the spell-checker option. This situation can be explained by the fact that many smart devices, such as tablets and mobile phones, have smart keyboards to change the lexical spelling mistakes into standard usage based on the dictionary of the target language. Although the aim of this study is not focused on non-standard lexical preferences, these authors were excluded from the list. Eventually, 55 authors were left for the data

collection; however, when compiling the data, it was noticed that some authors preferred to write extremely short messages – for instance, two or three words or only *bknz* (see also) to another title.

Finally, 45 authors were selected in accordance with the research aims in terms of size. It is important to state that the questionnaires were not administered to the authors on Twitter. The suitable authors were then divided into three subsets based on their writing habits, which were *long or medium length or short texts*. 900 texts from 45 authors are split into four corpora with 225 texts each as follows; long texts are considered ≤550 ≥250, medium length ≤250 ≥150, and short texts are ≤150 ≥50 within the research context. The break is made at the sentence boundary; thus, it is not possible to collect the data with sharp, clear cuts. Actually, the average of the texts is respectively 347, 119 and 43 words, but, in order to keep it simple, corpus names are used to describe the data along with general size as Corpus 1 – Long Texts, Corpus 2 – Medium Length Texts and Corpus 3 – Short Texts. Corpus 4 is not centrally focused on data size; instead, it has a cross-genre application. For that reason, it is referred to as Corpus 4 – Cross-Genre Comparison. Moreover, there are three sub-samples of Corpus 1 in terms of candidate author, size, and the limited text size per author. More information regarding each corpus is provided in Section 4.3. and its sub-sections.

The reason for collecting data in three text sizes is that forensic texts tend to be between 400 and 700 words in length (Coulthard, 1994). Nevertheless 'no one really knows how small a sample one can reliably work with, at what size significant irregularities begin to emerge' (Coulthard, 1994, p. 13). Besides, 'the forensic world is rarely ideal, and the texts are often unhelpfully short' (Coulthard, Johnson and Wright, 2017, p. 152). Thus, there is a need of short texts for the analysis since the aim is to create a realistic authorship attribution scenario.

Also, real-world cases have different characteristics and flexibility in language, not like a simulated experiment. Nevertheless, when linguistic diversity is considered, Eksi Sozluk presents a highly realistic background. For that reason, it enables a fundamental requirement for authorship attribution corpus.

In a forensic scenario, the extracted texts are assumed as evidence for the case. Known and unknown texts are compared between the possible authors. Accordingly, in each test there is a set of unknown texts by an unknown author, which is needed to compare to the rest of the authors. To minimise linguistic changes over the years, the texts are collected from a particular time range, as mentioned above. By analysing the performance of various authorship

attribution scenarios on the multiple corpora, the study aimed to determine the number of authors, the role of features, the length of the texts, the number of available texts and the cross-genre data effects of the results. The results should be an important contribution to the potential prospective authorship attribution studies in Turkish within the context of forensic linguistics.

Finally, collecting pages from the web has also two sub-sections in terms of construction of the corpus. First, 'one can construct a corpus automatically by downloading pages from the web' and second, 'one can collect a corpus by manual or semi-automatic selection of pages downloaded from the web, according to precisely specified design criteria' (Lüdeling et al., 2007, p. 8). For the current data, the automatic function was not possible to run according to the specified criteria.

In this study, data were collected manually from Eksi Sozluk and semi-automatically from Twitter. While it is possible to construct a corpus from the websites automatically, Eksi Sozluk is not suitable for such functionality since there are several texts from several authors, regardless of the specific criteria such as writing date, editing history, text size.

For instance, in each title there are many author entries. If one wants to download one author's entry, one is likely to download the rest. Since such a method needs a cleaning stage, which is time-consuming, it was decided to collect the Eksi Sozluk corpus manually. At the earlier stages, the entries were copied and pasted into a Microsoft Word file and then inserted into NVivo 11 as a source. However, the manual download is a limitation of any study regarding size. In the future, such downloading should ideally be done automatically.

On the other hand, Twitter allows downloading whole data belonging to a particular author. It is accessible via applications and one does not need to register to the website as a user. NVivo 11's NCapture function was used to get the texts for cross-genre comparison corpus. NCapture is a free web browser extension that enables material to be gathered from the web to import into NVivo (NVivo, 2018). This approach was repeated for each author in the Twitter corpus. The collected texts from both sources were adjusted to the text format, and meta-data was deleted to avoid confusion in feature coding. In order to achieve this, Sinclair's (1991, p. 21) clean text policy was followed for 'the safest policy is to keep the text as it is, unprocessed and clean of any other codes'. The Internet data may include irrelevant elements such as dates or location. Accordingly, author names, dates, social media sharing buttons, popularity buttons, and hashtags were not necessary to add because they were not the production of the authors. Furthermore, there are three forms of tweets as; standard messages that shared to the

community, directed tweets that focused on a particular Twitter user i.e. '@tom I don't agree with you.' In such a case, only the recipient's username is removed, and the rest is used in the study. As it is mentioned above, retweets are originally posted by another user and any user can repost the text. These messages show RT on top of the tweets since the account holder is not the originator of the tweet these ones are excluded from the study.

In the following section, the corpora are presented, including the characteristics of each corpus and its sub-samples.

4.3. Corpora

For the main four corpora, the length of the data was controlled in order to see how much text size difference would affect the accuracy. It is possible to find many texts at different text lengths from various topics such as the topic of distribution in Eksi Sozluk. There are 27 different topics along with the uncategorised ones. The topic organisation is made by experienced authors in Eksi Sozluk; however, there is no guarantee that the entry is exclusively assigned to a particular topic. For instance, an author may write an entry about politics and insert some memories related to his/her childhood. Since the entry is written under the particular topic that is categorised as politics by admin, it will be considered as politics. It is worth stating that there are many hidden factors in terms of topic classification in the current data set, more than is considered. Furthermore, this topic diversity is an excellent practice when the authorship attribution methods are applied in real-life scenarios.

In a similar manner, it is possible to collect many texts in various topics in Twitter. As it is mentioned above, mentions and hashtags serve as a topic organiser within Twitter and a tweet can be about anything.

Although a diversity of topics is an advantage of Eksi Sozluk and Twitter, it is not a particular interest of the current study. Nine hundred texts were collected in total from the 45 authors who met the collection criteria, based on their responses. Two hundred and twenty-five texts each with various sizes were compiled for the corpora – Corpus 1, Corpus 2 and Corpus 3 – from the 45 authors. Finally, for Corpus 4, 7 authors with 15 texts each from Corpus 3 and 8 authors with 15 texts from Twitter were collected. The total size of corpora is presented in Table 4-1. It should be noted that the cross-topic issue is autogenously welded in this study and in Table 4-1 it is presented to demonstrate the unbalanced distribution of texts.

Table 4-3: Classification of the corpora

	Corpus 1 – Long Texts	Corpus 2 – Medium Length Texts	Corpus 3 – Short Texts	Corpus 4 – Cross-Genre
Number of known texts	210	210	210	210
Number of disputed texts	15	15	15	15
Total number of texts	225	225	225	225
Total number of authors	15 authors including the disputed author (14 distinct authors in total)	15 authors including the disputed author (14 distinct authors in total)	15 authors including the disputed author (14 distinct authors in total)	7 authors from Twitter and 8 authors from Eksi Sozluk (14 distinct authors in total, same author has texts in both sources.)
Text Size	≤550 ≥250	≤220 150≥	≤100 50≥	≤50 5≥
Total word counts	78291	26908	9802	7625

As seen in the table, there are 210 known texts while there are only 15 unknown texts in each corpus. According to the scenario, there is only one questioned author, and 15+15 texts belong to that author, whilst 13 authors have just 15 texts per author in the remaining parts. In total there are 14 distinct authors in each corpus. It is worth to emphasise that; the disputed author is also represented within known texts. There is one certain author who corresponds the disputed author however, disputed author and the potential candidate author have separate author IDs. In each corpora the corresponding and the disputed author are selected randomly in order to feature the difference between corpora and emphasise that all disputed authors are different from each other. Keeping the same sequence may lead a misunderstanding between the corpora.

Finally, this study needed multiple training texts samples per author in different lengths in order to develop an accurate model for Turkish since some authors tend to produce shorter texts and the others longer texts while less but longer texts for that reason segmenting the corpora increased the chance of various authors and writing types. Thus, various corpora offer the possibility to test methods in different cases. In order to investigate the effect of the authors and text size in an attribution problem, four corpora provided containing 15 and 30 authors

subsets. Different data sets are helping the diversity for each scenario different corpora are used and randomly selected the texts based on the date they were written in each dataset. As it is mentioned above, topic diversity is an advantage in both sources (Eksi Sozluk and Twitter) thus, a diverse set of topics and sizes give a boost in performance and it is better than using the same data for different tests. The specialised corpora depending on various parameters are presented in the following sections.

4.3.1. Corpus 1 – Long Size Texts

Corpus 1 is balanced in terms of the number of texts per author, and the average size of texts is 347 words per text. The texts in this corpus were written by 14 different authors on 28 topics; the text size range is between ≤550 ≥250 and the top three most written topics per author are presented in Table 4-2.

Table 4-4: Text sizes and topics per author in Corpus1.

Authors	*Text Sizes in Total*	*Top Three Most Written Topics & Entry Numbers Per Topic*
*Author 1**	5555 words	Relationships 163 Questionnaires 131 Sports 108
Author 2	5265 words	Sports 574 Questionnaires 459 Relationships 252
Author 3	4910 words	Politics 594 Technology 352 News 345
Author 4	3825 words	Politics 752 News 490 History 340
Author 5	5378 words	Sports 1313 Politics 769 Music 537
Author 6	5527 words	Politics 413 Questionnaires 242

		Relationships 208
Author 7	5168 words	Television 731
		Sports 279
		Cinema 184
Author 8	5953 words	History 377
		Politics 356
		Relationships 225
Author 9	5940 words	Relationships 328
		Questionnaires 293
		Eksi Sozluk 109
Author 10	4496 words	Relationships 554
		Music 368
		Literature 316
Author 11	5257 words	Politics 4965
		Questionnaires 2362
		News 2140
Author 12	5517 words	Music 277
		Questionnaires 276
		News 179
Author 13	4939 words	Politics 2817
		History 1254
		News 972
Author 14	5842 words	Relationships 165
		Questionnaires 162
		Music 95
*Author 15**	4719 words	Relationships 163
		Questionnaires 131
		Sports 108
Total:	78291 words	

The disputed author (Author 1) has 163 texts on the *relationships* topic, 131 texts on the *questionnaires* (equals with opinions) topic, and 108 texts on the *sports* topic, whilst Author 4 has 257 texts on *politics*, 490 texts on *news* and 340 texts on the *history* titles. It is worth noting that the *questionnaires* topic might include various sub-topics from politics to sports. In one sense, it needs a deep decomposition when the topic is a controlled issue for the research.

As can be seen in Table 4-2, the smallest sample size in terms of number of words belongs to Author 4 with 3825 words in total for 15 texts. Author 9 has the largest number of words, a total of 5940. According to this scenario, Author 1 and Author 15 are identical authors in Corpus1 – Long Texts. An asterisk symbol used to refer the corresponding authors in Table 4-2. Each author has 15 texts however, disputed author (Author 1) and corresponding author (Author 15) have 30 texts in total. These authors are presented with an asterisk next to the author number. The sub-samples derived from Corpus 1, for that reason disputed and corresponding authors are the same in the next two sections.

4.3.1.1 The number of candidate authors

This test obtained data from Corpus 1 and Corpus 2. Although the default candidate author number is 15 in this research, the number is doubled only for this sample and the first and second 15 authors set are sourced from Corpus 1 and 2. In order to test the various conditions, this sample is gathered to estimate the impact of the candidate author sizes in this study. Since it consists of the same data as Corpus 1 and Corpus 2, it is not entitled to have a different corpus name. Instead, it is referred to as '30 Candidate Authors' in the discussion section. There are some factors to be attentive to in combining both corpora. First, there are two disputed authors in the corpora. For that reason, one set of texts that belong to the disputed author in Corpus 2 is excluded and one new author is inserted instead. Second, Corpus 2 is selected to insert into Corpus 1 in order not to have a corpus with a world of difference in terms of text size. Corpus 1 and Corpus 2 have the closest text sizes within the whole corpora. In this sample, there are 30 authors with 15 texts each and one author is the disputed author (Author 1). This sample is created with the aim of finding the best parameters for the study.

4.3.1.2. The number of texts per author

Similarly, this sample is also derived from Corpus 1 in order to test the role of the available texts per author. In general, the standard number of texts is 15 for this study. However, it is necessary to test the limits regarding available texts per author. Thus, the number of texts is gradually decreased from ten to five texts per author. This is not called a separate corpus, rather a sample to test the various conditions. Moreover, these samples are entitled 'Five Texts per

Author' and 'Ten Texts per Author' in Chapter 6. In the same way as Corpus 1, the disputed author is Author 1 in both samples.

4.3.2. Corpus 2 – Medium Length Texts

Corpus 2 is balanced in terms of text numbers – 15 texts per author and text size with an average of 119 words per text. Table 4-3 provides the total number of texts per author for 15 texts and the most popular topics for the author. As seen in Table 4-3, Author 12 (disputed author) mostly wrote on politics, sports, and news. The top three most written topics varied among the remaining authors. In the corpus, Author 13 is the corresponding author of Author 12. An asterisk symbol inserted to refer the same authors in Table 4-3.

Table 4-5: Text sizes and topics per author in Corpus2

Authors	*Text Sizes in Total*	*Top Three Most Written Topics & Entry Numbers Per Topic*
Author 1	1772 words	Politics 435 Sports 391 News 346
Author 2	1702 words	Questionnaires 492 Relationships 335 Cars 244
Author 3	1774 words	Relationships 1025 Questionnaires 908 Music 576
Author 4	1781 words	Questionnaires 1418 Literature 981 Eksi Sozluk 920
Author 5	1819 words	Relationships 1979 Questionnaires 1868 Health 497
Author 6	1721 words	Sports 783 Science 251 Television 171
Author 7	1858 words	History 1410 Literature 1201 Politics 1085
Author 8	1852 words	Sports 361 Politics 247 Technology 232
Author 9	1911 words	Questionnaires 189 Television 168 Relationships 164
Author 10	1840 words	Relationships 488 Music 330 Questionnaires 326
Author 11	1911 words	Politics 297 Eksi Sozluk 257

		News 111
*Author 12**	1639 words	Music 242
		Television 170
		Relationships 130
*Author 13**	1702 words	Music 242
		Television 170
		Relationships 130
Author 14	1807 words	Eksi Sozluk 693
		Questionnaires 651
		Relationships 561
Author 15	1819 words	Questionnaires 471
		Relationships 273
		Television 146
Total:	26908 words	

4.3.3. Corpus 3 – Short Size Texts

Corpus 3 is balanced in terms of text numbers – 15 texts per author and with an average of 43 words per text. Table 4-4 provides the total number of texts per author for 15 texts and the most popular topics for the author. Table 4-4, Author 6 (disputed author) mostly wrote on literature, relationships, and questionnaires. The top three most written topics varied among the remaining authors. In the corpus, Author 8 is the corresponding author of Author 6. In Table 4-4 those authors (A8 and A6) presented with an asterisk symbol aside.

Table 4-6:Text sizes and topics per author in Corpus3.

Authors	*Text Sizes in Total*	*Top Three Most Written Topics & Entry Numbers Per Topic*
Author 1	664 words	Relationships 238 News 200 Politics 188
Author 2	601 words	Technology 263 Music 230 Questionnaires 138
Author 3	619 words	Politics 748 News 690 Economics 505
Author 4	608 words	Politics 817 Sports 387 News 386
Author 5	651 words	Literature 395 Relationships 322 Questionnaires 293
*Author 6**	703 words	Literature 395 Relationships 322 Questionnaires 293
Author 7	681 words	Sports 962 Politics 831 Questionnaires 738
*Author 8**	662 words	Music 1178 Questionnaires 719 Politics 655
Author 9	641 words	Technology 998 Music 350 Television 323
Author 10	682 words	Questionnaires 724 Relationships 512 Eksi Sozluk 183
Author 11	667 words	Questionnaires 427 Relationships 316

		Economics 235
Author 12	591 words	Questionnaires 439
		Relationships 424
		Television 391
Author 13	666 words	Questionnaires 593
		Relationships 479
		Eksi Sozluk 288
Author 14	670 words	News 668
		Sports 622
		Politics 608
Author 15	696 words	Television 217
		Politics 209
		Questionnaires 207
Total:	9802 words	

4.3.4. Corpus 4 – Cross-Genre Comparison

In a real-world authorship attribution task, it is not always possible to find the same text types. The given text may be written in a specific genre and the other texts from another genre. For instance, in the case of suicide letter, it is highly possible that this is the only suicidal letter written by the author. In this case, when there is no comparison material, the material may be from other platforms. As Chaski (2007) stated, text type or register is one of the challenges in authorship attribution since it is rare to find the same as the register of the questioned document in real cases. For instance, in a forensic examination of suspicious tweets, other data sets may include emails, blog posts or online comments. In such cases, the data under investigation may not present same prototypical features as the comparison data set. Thus, an authorship attribution method for cross-genre comparison is a necessity for Turkish. There are only some studies focused on the effects of cross-genre data on authorship attribution and none in Turkish. There is, therefore, a clear need to analyse texts from different genres.

The cross-genre comparison corpus was compiled by searching Twitter for the authors who linked their Eksi Sozluk account with their Twitter account. The disputed and comparison corpora do not share all properties even though both of them are CMC mediums. Seven Twitter accounts with at least 1000 tweets were selected. The datasets from these users were captured with the NVivo 11 NCapture function. The drawback of this function is downloading whole

data from the first tweet to the latest one, regardless of their writing dates. For that reason, only the tweets which were written around the same time with the entries compared from Eksi Sozluk. Furthermore, Twitter allows the user to post messages to a maximum of 280 characters in length (Twitter Blog, 2017). Although producing different sizes of texts with 280 characters depends on the author's performance, Corpus 3 – Short Size Texts are the most similar data with an average of 43 words per text. Some tweets have only retweets or hyperlinks, which have little importance for authorship attribution cases. The tweets that are structured as retweets and hyperlinks, besides the ones with limited word count (for instance, five or fewer words), are excluded from the data sampling. Retweets do not include any information about the user's writing style. After collecting 105 tweets from 7 different Twitter accounts and 120 texts from 8 Eksi Sozluk authors from Corpus 3, depending on the criteria, the texts are inserted into NVivo 11 as an internal source. Tweets and Eksi Sozluk entries are grouped by author username and changed into A1, A2 and so on, as is described for Eksi Sozluk. Similar to Eksi Sozluk data, the topic is neither a controlled nor sampling issue for Twitter. In this corpus, disputed author (Author 8) is from Eksi Sozluk. However, the corresponding author, Author 9, is from Twitter. In Table 4-5, text sizes are provided in Corpus 4. It is important to note that Author 8 (Eksi Sozluk), Author 9 (Twitter) are the same authors but the texts are collected from different sources.

Table 4-7: Text sizes per author in Corpus4.

Authors	*Text Sizes in Total*
Author 1	664 words
Author 2	601 words
Author 3	619 words
Author 4	608 words
Author 5	651 words
Author 6	703 words
Author 7	681 words

*Author 8**	662 words
*Author 9**	363 words
Author 10	489 words
Author 11	339 words
Author 12	267 words
Author 13	313 words
Author 14	391 words
Author 15	274 words
Total:	7625 words

Authors from 1 to 8 are selected from Eksi Sozluk, which has 5189 words in total from 8 authors. The average of the texts is 43 words per text. On the other hand, texts from authors 9 to 15 are collected from Twitter, which consists of 2436 words in total. The average of the texts is 23 words per text. When the data are combined, all in all, the average of the texts is 33 words in this corpus.

In a real case work, the situation can be difficult to obtain same amount of data, the disputed author has more texts in one genre and less in the other one. Since, this is a possible scenario in attributing authorship, the size of the texts is not equally distributed between Eksi Sozluk and Twitter.

4.4. Methodological Approach

As stated in previous chapters, there is no methodology for authorship attribution in Turkish texts apart from the methods applied in stylometric studies. However, there are many attempts in forensic linguistics to combine the approaches, including descriptive and statistical methods.

Initially, Grant (2010; 2013) offered a quantitative approach, which is combined with statistical testing in the research design. Following that, MacLeod and Grant (2012), Nini and Grant (2013), Wright (2014; 2017), Johnson and Wright (2014), Larner (2014), and finally Nini (2018) used such an approach in authorship attribution studies. Following Grant (2010), MacLeod and Grant (2012) performed a qualitative analysis with a WordSmith (2012) tool to

identify the occurrences of some features. First, the lexicons were identified from micro-messages, and then the Delta distance metric test was used for statistical testing. This method demonstrated positive results in attributing authorship for short messages. Nini and Grant (2013) used systemic functional linguistic as a framework in order to identify the author's stylistic choices. The variables were coded as conjunctions, modality, mood, nominalisation, theme and transitivity. They used ANOVA testing as statistical testing to compare the frequencies of variables between authors.

Their method was successful enough to reduce potential biases in stylistic approaches and provide a scientific method. Wright (2014; 2017) applied the method to identify the idiolectal effects of word combinations, namely word n-grams in a relevant population data. Similar to Grant (2013), Wright (2014; 2017) measured with the Jaccard co-efficient statistic test, and the results were promising when attributing e-mails. Similarly, Johnson and Wright (2014) focused on word n-grams for the analysis and measured the similarity between any two tests using the Jaccard coefficient test. Different from the previous studies, Larner (2014) collected formulaic sequence tokens as distinctive style markers and used the Jaccard coefficient test; however, the results presented low success. Finally, Nini (2018) used word 2-grams in analysing *Jack the Ripper* letters and applied the Jaccard coefficient test as statistical testing and found an editorial link between the authors.

In light of these studies, based on descriptive and statistical methods, the method applied here is similar to Grant (2010; 2013) and MacLeod and Grant (2012) in feature selection, since it mostly depends on data-driven features along with the characterises of short text messages and Internet features. There are no predefined style markers, such as formulaic sequences (Larner, 2014), grammatical theories (Nini and Grant, 2013) or word sequences (Wright, 2014). Regarding statistical testing, most of the studies presented here applied the Jaccard coefficient test in order to measure the distance between the texts. The common ground of these studies is they have used relatively short texts such as tweets or short text messages when compared with academic assignments (e.g. Nini and Grant, 2013). Due to the Jaccard coefficient test being based on the presence and absence notion, it works fine with short texts.

The Jaccard coefficient measures the binary similarity and dissimilarity between text pairs, which is a classification system in problems of pattern analysis (Choi et al., 2011). It is applied in many fields from biology to authorship identification (e.g. Grant 2010; 2013). Since it doesn't depend on frequencies, such kind of statistical testing is suitable for short texts like the

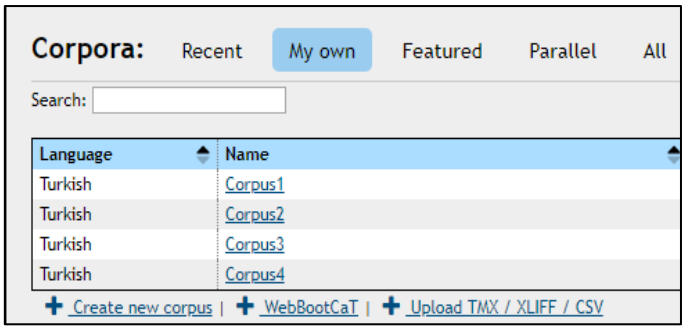
data in this study. Furthermore, Grant (2013, p. 417) proposed a methodological protocol for classification problems in authorship attribution.

According to the protocol, the first step is to identify the linguistic features in the texts, which is the first and the most critical stage of the analysis. First of all, Herring's computer-mediated discourse analysis (1996) framework, which was presented previously, is adapted for this study. This framework provides a description of the interactive practices.

The structural domain includes 'emoticons, abbreviations, lexical items (such as personal pronouns), word formations, syntactic patterns and quoting' (Herring, 2004, p.18), which can be detected by the Sketch Engine corpus tool and reading the texts from the beginning to the end. Sketch Engine (2018) is a corpus tool to explore how language works and to identify instantly what is typical in language and what is rare, unusual or emerging usage, which is also designed for text analysis or text-mining applications. Due to its availability online and ease to use, Sketch Engine is used for corpus analysis in this study.

The first stage of the analysis after compiling the corpora is started from the Sketch Engine (2018) after it has been converted into the required file format. When the corpus has been uploaded, it shows the basic statistical information about the corpus – for instance, the number of words, sentences, and number of tokens. It allows uploading several corpora into an online database and identifying wordlists based on frequency in the data, keyword lists. Comparing with a larger corpus also, it is possible to observe the grammatical relations with collocations.

Figure 4-6: A display of corpora in Sketch Engine.



Wordlists are described as 'one essential starting point for a systemic textual analysis' (Stubbs, 2005, p. 11) and provide all the words in the corpus based on their frequency, starting with the most frequent words, which are usually grammatical words (Baker, 2006). Furthermore,

wordlists can generate the word grams list, which is a sequence of words (bigram = 2 structures, trigram = 3 structures … n-gram = n structures) using the data (Sketch Engine, 2018).

Table 4.6 presents a typical wordlist, which is sorted according to frequency.

Table 4-8: An example of the first 20 words on the wordlist.

WORDLIST

word (175 items)

	Word	Frequency		Word	Frequency		Word	Frequency
1	bi	121	18	yok	28	35	her	15
2	bir	75	19	böyle	28	36	sen	15
3	da	60	20	ama	28	37	zaman	15
4	bu	57	21	en	28	38	kötü	14
5	o	57	22	ki	27	39	biri	14
6	de	56	23	şey	26	40	insan	14
7	ben	50	24	adam	26	41	değil	14
8	çok	46	25	bile	24	42	olarak	14
9	lan	42	26	ya	24	43	onu	13
10	ne	42	27	iyi	21	44	beni	12
11	gibi	38	28	diye	19	45	güzel	12
12	var	35	29	mi	19	46	bana	12
13	kadar	32	30	koyım	18	47	bunu	12
14	için	31	31	amına	18	48	olm	12
15	ve	30	32	falan	18	49	hiç	12
16	sonra	30	33	öyle	17	50	tek	12
17	daha	29	34	zaten	17			

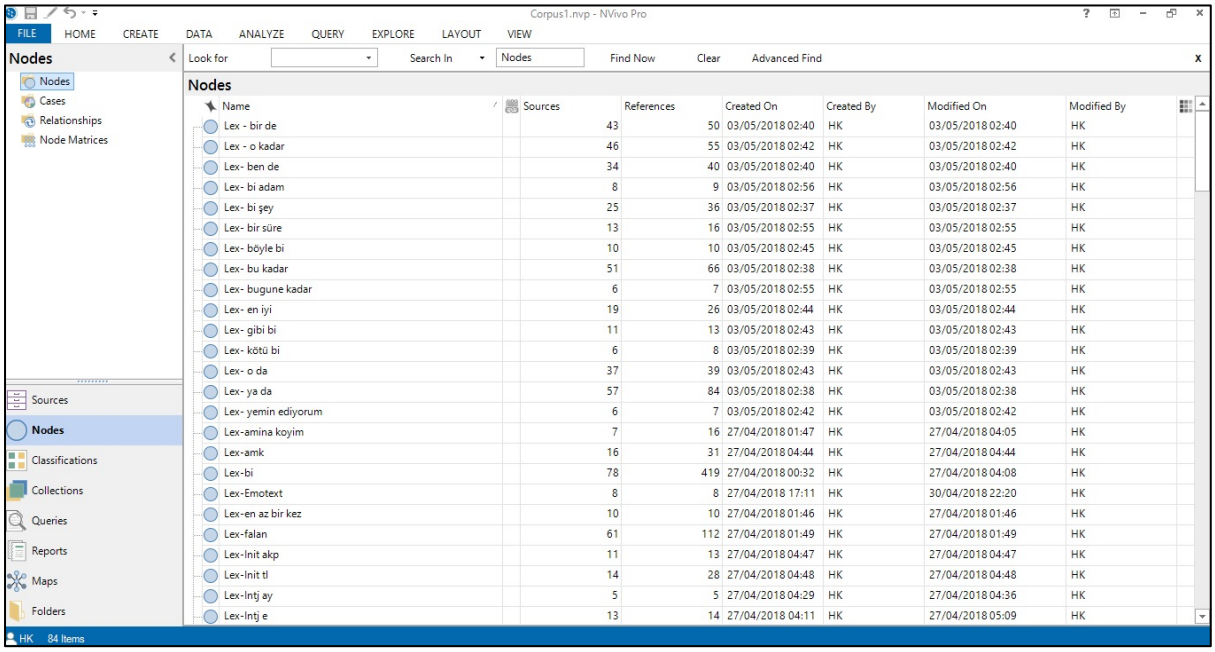
Although wordlists provide an insight into the data, the list may be considerably longer in long texts. In order to make it manageable, it is necessary to decide which features carry more information about the data and rules to apply. Stubbs (2005) used the top 50 frequent words that occurred 20 times in his study, which was focused on a stylistic analysis of a literary book. The first words with high frequency provide information about the data. As seen, a frequency threshold is set in order to exclude the non-informative features in this study (Stubbs, 2005). In a similar way, Grant (2013) approached the features with the minimum 10 frequencies on the wordlist. On the other hand, sometimes a feature is found to be highly frequent but has content-related features that may cause non-replicable subjective results in forensic linguistics – for instance, the proper noun *Turkiye* (Turkey) may frequently occur in the corpus and be positioned on the top frequency list, but the text may be a political essay written about Turkey. The frequency list is mostly overwhelmed by content words and function words. Content-related features may be useful in predicting the author, but also it decreases the applicability of the model for the following features and makes the results *one-time only,* rather than small-scale online text in Turkish. For that reason, it is necessary to exclude proper nouns from the

lists, such as city and country names, since these are mostly topic-dependent. However, this study is benefited from word n-grams as features even though these are related to the content. That is to say, word n-grams are taken as mathematical formula rather than its content related meaning.

Furthermore, concordance lines can be generated for every word in the corpus listed in the wordlist. This both provides information for the semantic meaning of a word (Sinclair, 2004, p. 19) and reveals the grammatical patterning of the word. In this study, it is used to reveal the grammatical patterning and is allowed to extract 'patterns of language use' (Baker, 2006, p. 77) in selecting the linguistic features.

Beyond wordlists, McMenamin (2002) stated the methods for studying variation in two models as bottom-up and top-down in identifying the linguistic features in text, as mentioned above. On the one hand, the bottom-up model 'searches for recurrent patterns, distributions, and forms of organisation in writing' with the purpose of finding evidence in the target text group (McMenamin, 2002, p. 74). On the other hand, a top-down approach requires research for 'predetermined taxonomy of stylistic items which would allow for the discrimination of writers' in a specific speech community (McMenamin, 2002). Based on these two approaches, a top-down approach was applied first to find linguistic features from previous studies. Then, to find data-driven features, a bottom-up approach was used for coding the features. Wordlists were useful in identifying the data-driven features. Moreover, the linguistic features divided according to their functions as lexical, syntactic and structural on NVivo 11, which is a qualitative data analysis tool, in order to organise, classify and find the insights in the data (Wikipedia, 2018). A display of NVivo 11 is presented in Figure 4.2.

Figure 4-7: A display coded nodes in NVivo 11.



Name	Sources	References	Created On	Created By	Modified On	Modified By
Lex - bir de	43	50	03/05/2018 02:40	HK	03/05/2018 02:40	HK
Lex - o kadar	46	55	03/05/2018 02:42	HK	03/05/2018 02:42	HK
Lex- ben de	34	40	03/05/2018 02:40	HK	03/05/2018 02:40	HK
Lex- bi adam	8	9	03/05/2018 02:56	HK	03/05/2018 02:56	HK
Lex- bi şey	25	36	03/05/2018 02:37	HK	03/05/2018 02:37	HK
Lex- bir süre	13	16	03/05/2018 02:55	HK	03/05/2018 02:55	HK
Lex- böyle bi	10	10	03/05/2018 02:45	HK	03/05/2018 02:45	HK
Lex- bu kadar	51	66	03/05/2018 02:38	HK	03/05/2018 02:38	HK
Lex- bugune kadar	6	7	03/05/2018 02:55	HK	03/05/2018 02:55	HK
Lex- en iyi	19	26	03/05/2018 02:44	HK	03/05/2018 02:44	HK
Lex- gibi bi	11	13	03/05/2018 02:43	HK	03/05/2018 02:43	HK
Lex- kötü bi	6	8	03/05/2018 02:39	HK	03/05/2018 02:39	HK
Lex- o da	37	39	03/05/2018 02:43	HK	03/05/2018 02:43	HK
Lex- ya da	57	84	03/05/2018 02:38	HK	03/05/2018 02:38	HK
Lex- yemin ediyorum	6	7	03/05/2018 02:42	HK	03/05/2018 02:42	HK
Lex-amina koyim	7	16	27/04/2018 01:47	HK	27/04/2018 04:05	HK
Lex-amk	16	31	27/04/2018 04:44	HK	27/04/2018 04:44	HK
Lex-bi	78	419	27/04/2018 00:32	HK	27/04/2018 04:08	HK
Lex-Emotext	8	8	27/04/2018 17:11	HK	30/04/2018 22:20	HK
Lex-en az bir kez	10	10	27/04/2018 01:46	HK	27/04/2018 01:46	HK
Lex-falan	61	112	27/04/2018 01:49	HK	27/04/2018 01:49	HK
Lex-Init akp	11	13	27/04/2018 04:47	HK	27/04/2018 04:47	HK
Lex-Init tl	14	28	27/04/2018 04:48	HK	27/04/2018 04:48	HK
Lex-Intj ay	5	5	27/04/2018 04:29	HK	27/04/2018 04:36	HK
Lex-Intj e	13	14	27/04/2018 04:11	HK	27/04/2018 05:09	HK

Although the feature selection method here depends on a wordlist that is created automatically, a second coder was asked at this stage to contribute to the reliability of the analysis. A description of linguistic features used in this study is presented in Chapter 5.

In the second stage of the protocol, Grant (2013) proposed describing the features of the known texts first. However, this was not followed in this study. Instead, all texts were treated as equal without prioritising any of them. Wordlists were created for all texts, and the important features were coded without considering whether they were known or unknown texts, in order to decrease researcher bias.

The following step is related to the examination of the query texts for the identified features. With this purpose, a statistical analysis is run by comparing all texts with each other. In this step, each text is considered as a single unit.

Finally, a protocol was proposed to draw conclusions based on the consistency and distinctiveness of the disputed texts with the known texts. These conclusions are presented in Chapter 6 – Discussion and Analysis.

4.5. Statistical Design

In forensic linguistics, linguistic evidence is significant because it may contribute to the conviction of culprits or the release of innocents – this demonstrates the significance of reliability. Thus, in recent years there have been some attempts to overcome the bias and develop a method that depends on qualitative and statistical approaches.

The advantage of using both methods is stated by Creswell and Plano Clark (2007, p. 5) as follows: 'the use of quantitative and qualitative approaches in combination provides a better understanding of research problems than either approach alone.' The importance of the quantitation method regarding accuracy is highlighted well in Schmied's (1993) 'Qualitative and Quantitative Research Approaches to English Relative Constructions' study along these lines:

> Statistical tools may be employed to highlight and underpin impressions gained by a sensitive analyst. Probably, however, the qualitative analysis will never be sufficiently systematic and the quantitative one never delicate enough to replace the other completely. This one more illustrates my point that both approaches have to be combined in a detailed corpus linguistics analyses.

As Denzin (1989 p.307) corroborated 'by combining multiple observers, theories, methods and data sources can hope to overcome the intrinsic bias that comes from single methods, a single observer, and single theory studies.'

Considering the advantages of both approaches, this study used both quantitative and qualitative methods aligned with Grant (2013) and MacLeod and Grant (2012) in order to improve the reliability of the results.

For the statistical approach employed here first, the matrix query results are exported into an XLS file that makes the data ready for statistical analysis from NVivo11 and the results were converted into the binary system. Matrix query enables to see coding intersections between two lists of items, and it can be used to ask a wide range of questions about patterns in the coded data and gain access to the content which shows those patterns (Nvivo11, 2018). The next step was to use statistical analysis to run the capability of discriminating between several authors.

In this study, the Jaccard distance test was employed, which is complementary to the Jaccard coefficient. It shows the dissimilarity with the rest of the authors and similarity depends on the shared feature numbers and takes into consideration whether a particular feature is found or not in both the unknown texts and known texts rather than quantifying the instances. The

Jaccard distance test was applied to the chosen numerical data set of codes to classify the author texts. The use of the Jaccard measure is common among various disciplines from ecology to linguistics and it has been introduced into forensic authorship analysis as a way of measuring the similarity or distance between questioned and known documents based on the different linguistic features (e.g. Grant, 2010, 2013; Wright, 2014; Larner, 2014; Juola, 2013). The quantitative data were analysed using R-Studio, which is 'a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics' (R Studio, 2011). R-Studio can be used by linguists in order to analyse the relationship between linguistic variables and visualise meaningful plots for the results. Grant (2004; 2007; 2013) used SPSS in his authorship attribution studies. For example, in Corpus 1 there are 15 authors and 15 samples from each author, which gives a total of 225 comparison texts using the single occurrence of linguistic features. Once the data have been prepared and converted into a coding matrix file, each test using R-Studio does not take long to analyse.

In the Jaccard distance test, the absence of a feature does not increase or decrease the results regarding distance value. In the distance measure, when a value is close to zero this indicates those texts are different from the rest and similar to each other. In other words, in the Jaccard distance test, Text A and Text B have the following interpretations:

- The distance is small if Text A/Author A and Text B/Author B are similar.
- The distance is larger if they are not similar.
- The distance is 0 if they are the same.
- The distance is 1 if the texts are completely different.

As it is mentioned above, Jaccard distance test is for comparing two binary value sets. The formula used to calculate Jaccard distance is;

$$d_{ij} = \frac{q + r}{p + q + r}$$

Where

p = the number of variables that positive for both objects (shared items)

q = the number of variables that positive for the i th objects and negative for the j th object (items unique to sample)

r = the number of variables that negative for the i th objects and positive for the j th object (items unique to comparison sample)

s = the number of variables that negative for both objects

For instance, when a value is 1 it indicates the presence and 0 is absence. Suppose that there are two different author sets i.e. A1 = {1, 1, 1, 1} and A2 = {0, 1, 0, 0} for four features. Since there are four features, it is accepted that these features have four dimensions. Jaccard distance test takes the number of unique variables for the A1 and unique items for the A2. Later divide this number to the total number of positive variables for both features, q and r values. At the end Jaccard distance is calculated as 0.75 between two authors in the example (Kardi, 2015) which means they share only one feature within four features. If there are fewer features, it is simple to find similarities between. For instance, if Author1 used only five features overall from the feature set, there would be a match up with the other author who used two/three of these features. Since the comparison made between limited data, the statistical formula would assign a match although it is three matches out of five.

As seen in Table 4-8, the diagonal values are zero since the same author is being compared with itself. That is to say, there is no distance between same author. Decimal numbers between zero and one represent the distance value.

Table 4-9: An example of the Jaccard distance measure.

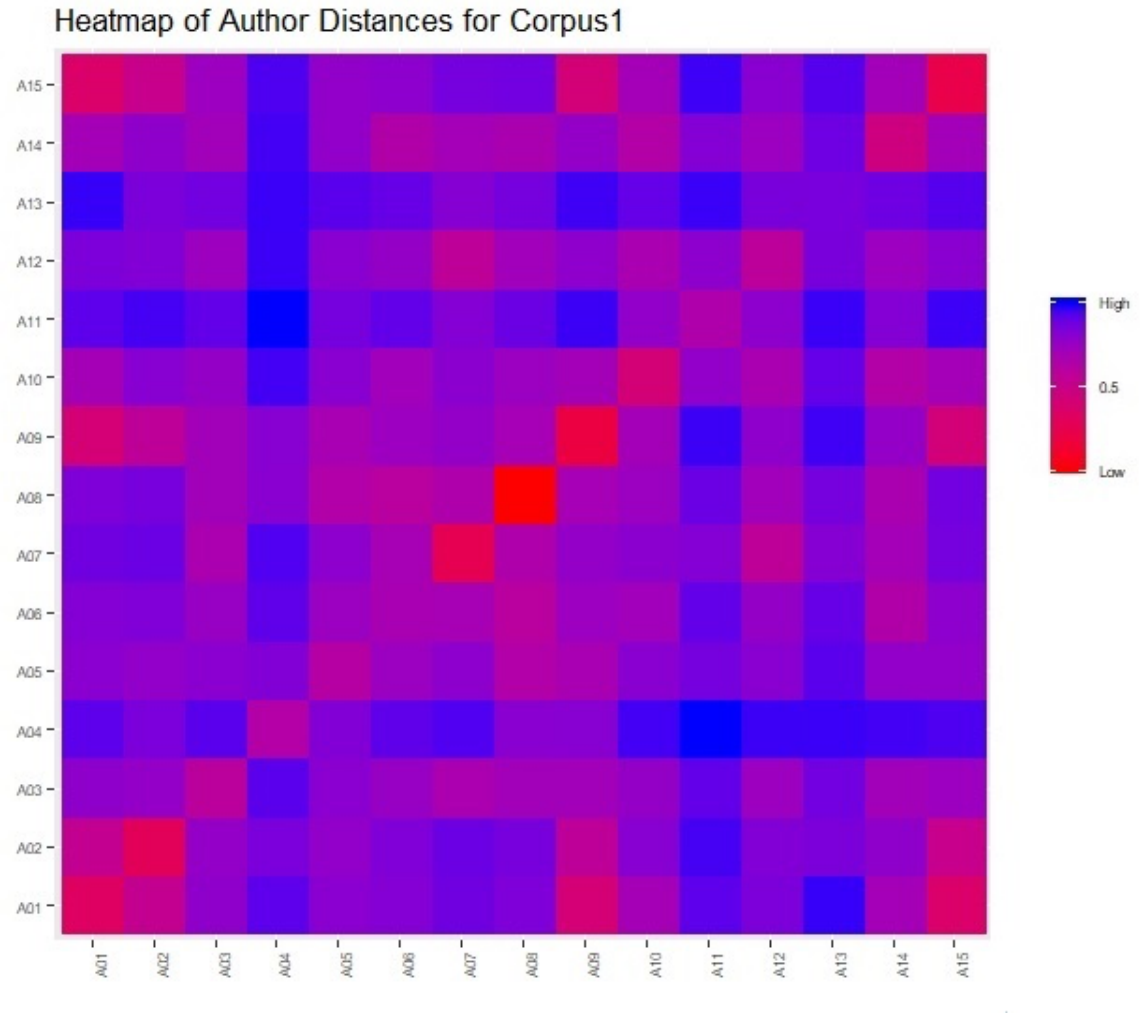
	A1T1	A1T2	A1T3	A1T4	A1T5	A1T6	A1T7	A1T8	A1T9	A1T10
A1T1	0	0,882353	0,75	0,714286	0,411765	0,578947	0,571429	0,708333	0,764706	0,705882
A1T2	0,882353	0	0,909091	0,666667	0,857143	0,769231	0,882353	0,823529	0,571429	0,428571
A1T3	0,75	0,909091	0	0,666667	0,666667	0,769231	0,75	0,758621	0,818182	0,772727
A1T4	0,714286	0,666667	0,666667	0	0,588235	0,666667	0,826087	0,6	0,615385	0,416667
A1T5	0,411765	0,857143	0,666667	0,588235	0	0,588235	0,578947	0,782609	0,714286	0,642857
A1T6	0,578947	0,769231	0,769231	0,666667	0,588235	0	0,578947	0,727273	0,714286	0,538462
A1T7	0,571429	0,882353	0,75	0,826087	0,578947	0,578947	0	0,807692	0,894737	0,777778
A1T8	0,708333	0,823529	0,758621	0,6	0,782609	0,727273	0,807692	0	0,777778	0,722222
A1T9	0,764706	0,571429	0,818182	0,615385	0,714286	0,714286	0,894737	0,777778	0	0,375
A1T10	0,705882	0,428571	0,772727	0,416667	0,642857	0,538462	0,777778	0,722222	0,375	0

After the analysis is done, the most straightforward interpretation of the results is available by producing heat maps, which show the way of assigning the correct author. The analysis in Chapter 6 aims to verify the applicability of Jaccard distance between texts considering the linguistic features. Heat maps are described in detail in the following section.

4.5.1. Data visualisation – Heat maps

Heat maps are for ‘graphical representations of the data’, and the matrix-like values present in colour (Silhavy et al., 2016, p. 218). Heat maps display the results based on the gradient of the colours and can be useful for those who do not have a statistical background. From this point, illustrating the results in an easy way for understanding is necessary in forensic linguistics. The purpose of this study is to create an easy and understandable procedure and obvious results for the court in Turkish authorship attribution studies. One of the main advantages of the heat maps is in showing big numerical data. For instance, it is time-saving, as it provides a high level of visibility and ‘speeds up the analysis process’ (Silhavy et al., 2016, p. 219). Figure 4-3 shows an example of a heat map.

Figure 4-8: A display of a HeatMap



This heap map describes the distance between authors. Figure 4-2 is the output of the numeric results of the analysis. All rows and columns can be seen at once by colours on the heat map. This leads to a better understanding of the data in contrast to more complicated plots.

Such types of plots may help in writing reports for the court as an expert. The heat map is a useful way of arranging the table of numbers by using colours. Colour clustering shows the relationship of the values in those tiles. With the aid of a colour gradient, it is effortless to identify highest and lowest values in each column, which in this study represents similar and dissimilar texts. Heat maps are generated by the ggplot2 package from R-Studio, which allows

the user to visualise the data. This package provides various visualisation options, but for this study, only heat maps are produced.

For this study, all distance heat maps were represented by red and blue colours. Similar texts appear close to red rather than blue within the gradient colour scheme. Red tint represents the minimum distance in the entire map while the blue colour is close to more significant distances. Furthermore, heat map reading should start from the density of the colours depending on the legend which shows the high/low value with colours. Along each axis, all the texts in the study are arranged in groups by the author. The diagonal line represents where the author/author and text/text would have been compared against itself. At each side of the diagonal line, it can be seen that there are more red patches showing which texts are similar to one another using the Jaccard measure. Where an author has a consistently identifiable style, this can be seen as a rough red square against the background red and blue colour of the rest of the comparisons. In short, blue represents greater distance while red indicates shorter distance.

4.6. Ethical considerations

This section explains the ethical considerations in this study. Virtual communities provide a way for a group of peers or people who have an interest in the same topic to interact with each other. They include discussion boards, online chat rooms, mailing lists or collaborative online encyclopaedias, as in the current study. In recent years, a boom in Internet-based community research has led to increasing interest in online research ethics. Online virtual spaces present many new opportunities for researchers but also present further ethical considerations. In other words, ethical issues in online research have become sensitive. This section discusses the ethical considerations for this research.

All the data from the Internet used in this study have been taken from a public online collaborative encyclopaedia and Twitter. Some online forum pages or chat rooms require registration to access them, but the data for this study is an open-access website where only registered users are allowed to post content, although anyone may create an account with no verification. In order to be registered as a user, there is a waiting period for newcomers, which might be months or even years.

Madge (2007) asserts that for private and semi-private virtual sources, such as email or closed chat rooms, informed consent should be considered essential, but in open access forums, informed consent may not always be required.

When considering the online data, Sixsmith and Murray (2001, p. 14) asked an important question that needs to be thought about: *To whom do the posts belong?* Do they belong to the poster (author), electronic group (community), or to any observer (including researchers)? This might be an issue for the online communities that have sensitive matters such as health problems, gender discussions etc., because in this type of data, authors write about anything, mostly on politics or prominent political figures. Although it is accepted that people are free to share their ideas, sometimes their contributions can cause unpleasantness. Such cases rarely occur, but it is not impossible. For that reason, an acknowledgement e-mail is sent to protect them from any sort of harm.

Nevertheless, the current data set does not touch these kinds of personal issues, and participators also confirm the terms of service before joining the website. Furthermore, in order to ensure the anonymity of users, each author would take a number – for instance, A1 for Author 1, A2 for Author 2 and so on.

According to Creswell (2003), a qualitative study reports detailed views of participants in a study in its natural setting. In online settings, participants usually take part in social interactions by using nicknames and therefore their identity is unknown. Hence, a problem of natural setting occurs, as the guidelines for online research, given by Kinkus (2002), indicate that online researchers must verify the identities of online users. In order to overcome this issue, the participants in this study were asked to report in the questionnaire who they were and what nicknames they used on blog forums. Prior to conducting the study, they were given copies of a plain language statement that contained information about the aim of the present study so that they were aware of what was required of them if they chose to participate. However, participants may act in a different way if they know they are being observed. In the current study, for example, they could edit their entries or change their style of writing on purpose, which could affect the results. To overcome this disadvantage, there is a way not to let people know that they or their output are being observed. But this creates another ethical problem. In order to avoid this potential problem, authors who had more than 1000 entries were primarily selected.

In this way, even though they were aware they were under observation, they could not know which entries were chosen for the study.

As previously mentioned, with online data collection various computer-mediated communication systems are easy to access and observe. That is why an asynchronous online

setting has been selected to provide a comfortable way to collect data, principally because of its feasibility. In addition to this, participants in this environment do not assume that their communication is private, and they are aware that the website has publicly accessible archives. Also, by accepting the membership agreement, they admit giving their copyright to the website owner. The website owner or anyone who reads the posts can share them on social networking websites such as Facebook, Twitter or blogs; indeed, each post has a share button below it.

Another essential point mentioned in the corpus design variable section is that, although this study is not an authorship profiling study, it is important to know some components of composing texts by potential authors, such as using predictive writing on a mobile device. Eksi Sozluk has a mobile application and it is also possible to contribute via mobile devices such as tablets or mobile phones. In order to obviate these issues, an online questionnaire was administered to potential participants while two months of data were collected during November 2014. The questions were structured in the form of closed-ended questions that consisted of general information about the selected participants. The questions asked participants to fill in the boxes according to their general inquiries and authoring experiences. The same questions were asked of all participants. This was the only interaction with the participants.

Many online users prefer to adopt a pseudonym in virtual environments in order not to face any problems, and this gives a high level of anonymity to users.

In view of avoiding risks, and even though participants used pseudonyms, any personally identifiable information (PII) that did not affect the nature of the research was altered. Such material included names, dates, and places and all were replaced with apocrypha ones. No matter what a person is discussing, including intimate topics such as sexual experience/choice/fantasy or psychological issues, the publication of this research would not result in any shame, social status loss or harassment because of the anonymous identification number used for each contributor. Thus, it will be impossible to link results to individuals. In any case, it is not common for participants to share their personal characteristics, distinguishing features and biometric information or unique personal numbers.

Overall, this study was approved by the School of Languages and Social Sciences Ethics Committee at Aston University in May 2015 with the guidance of my supervisor.

Chapter 5: Feature Selection

This chapter focuses on feature selection and the coding approach for authorship attribution in Turkish. In the following sections, the coding approach (Section 5.1); classifications of the features according to lexical, syntactic, and structural divisions (Section 5.2); and the reliability and inter-coder reliability test results (Section 5.3) are presented. Section 5.4 then provides a summary of the results and the findings.

In authorship attribution, finding the most suitable features is as important as the statistical method applied to the research. There is still a debate over whether the valid markers are character length; sentence length; word length; n-grams; function words; punctuation marks; lexical, syntactic, and structural features; or a combination of them all. Moreover, there is no agreement about which marker set has more discriminative ability as this varies across different text types. For instance, as discussed above, Chaski (2001; 2005) used frequency of punctuation marks as discriminators in the texts based on their roles in the sentence as sentential, clausal, phrasal, appositive, or word internal. Her results were criticised, however, as the data set was too small to verify these as reliable markers (Grant and Baker, 2001). Mosteller and Wallace (1964) used function words as discriminators when analysing Federalist Papers because of their unconscious production, the high number of counts for the statistical analysis, and the content-independent nature.

Moreover, McMenamin (2002) claimed that style markers, including text format, numbers and symbols, abbreviation, punctuation, capitalisation, spelling, word formation, syntax, discourse, errors, and highly occurring words are distinctive features when attributing authorship. Grant and Baker (2001, p.77), however, said that 'it is not enough to simply show that a particular marker works or does not operate in a particular case'. In other words, it should not be concluded that selected markers would or would not work in all casework. Despite Coulthard (1994, p.40) mentioning that 'the future for forensic linguists must lie in the creation of a better standardised and more widely used methodology', it is still questionable to use a standardised method that includes the same features for every genre in the same way.

After all, looking for general agreement on feature usage is unlikely for all languages and different sets of features may work in various genres and languages.

Also, it is too early to come to conclusions regarding Turkish authorship attribution features as the number of studies is scarce in stylometric studies and there are none in the field of forensic linguistics, as mentioned above. Therefore, this study will fill a gap in the literature by describing the feature sets for such a study. The coding approach that is applied during the coding process is described in the following section.

5.1. The Coding Approach

The coding approach is divided into several steps that begin with deciding upon the relevant language features. To get the real data, structure tags such as the date, share or complaint buttons, and nicknames are removed as these kinds of additional structural details can affect the results of the study. In forensic authorship cases, language features should discriminate the authorship of a text from a limited number of authors or determine whether a message can be attributed to a particular author. For this reason, it is essential to set up reliable features for the analysis.

Due to the emergence of computer-mediated communication (CMC), there is a need to focus on internet-based languages as these have different norm structures and can be characterised as informal and expressive (Crystal, 2001). The impact of the internet has blurred the distinction between spoken and written language and internet-based language 'includes speech and writing, regional and class dialects, occupational genres (such as legal and scientific language), creative linguistic expression (as in literature), and a wide range of other styles of expression' (Crystal, 2004, p.6).

It is possible to distinguish several core properties between spoken language and different written internet genres, such as email, online chat messages, and SMS. Crystal (2011, p.21) also stated that 'internet language is identical in neither speech nor writing but selectively and adaptively displays properties of both'. Moreover, Crystal (2008, p.35-62) identified six distinctive characteristics of text messages: logograms, pictograms, initialism, omitted letters, nonstandard spellings, and shortenings/abbreviations. Abbreviations can also be defined as 'the shortening of the written form of a word or words without concomitant shortening of pronunciation' and categorised into six parts: initialism, acronyms, clipping, blends, awkward cases, and facetious forms (Crystal, 2003, p.120).

Although the nature of the data here is different from text messages, CMC does include e-mails, chats, virtual communities, instant messaging, short text messages, web pages, etc.

(Edens and Heinman, 2011, p.89), and these mediums may share features between them (Herring, 2009). Accordingly, Macleod and Grant (2013, p.218) proposed 'a multi-level sophisticated feature categorisation system' for Twitter messages that was sourced from previous studies (Smith et al., 2009) that were applied to text messages.

In this study, Internet language features (Crystal, 2001; 2008), previously tested features on short messages and Twitter messages (MacLeod and Grant, 2012; Grant, 2013), along with the data driven features are found potentially useful in analysing authorship. The existing framework has to be modified according to Turkish language features as it was initially created for English texts. Furthermore, data-driven features are selected as they are sourced from various individual choices and the web corpus has 'new word usages, and collocations the Web wins out against other corpora because of its sheer size and because it is always updated' (Leech, 2006, p.13), which may be a potentially discriminative feature for an authorship attribution study. Word n-grams are considered as one of the data-driven features. Besides, some features are found with bottom-up analysis while others top-down.

Before starting the coding process, it is important to decide how to code the features according to the nature of the data set. For instance, it is difficult to find many features and to produce reliable frequencies from short texts. In this study, the maximum and the minimum text sizes are between 519 and 35 words long, respectively. Therefore, questioning the presence or absence of a feature is more reliable than frequency-based analysis. For some features such as function words, however, their frequency is still high within the texts, e.g. the total frequency of the function word *bir* is 1853 in Corpus1.

With this presence and absence approach, even small texts include enough features to represent the texts. Grant (2010; 2013) proposed a formula for the presence or absence of features in short messages using the Jaccard coefficient test as this test does not use frequency when calculating dissimilarities. The zero occurrences show the absence of the feature and the non-zero occurrences represent its existence.

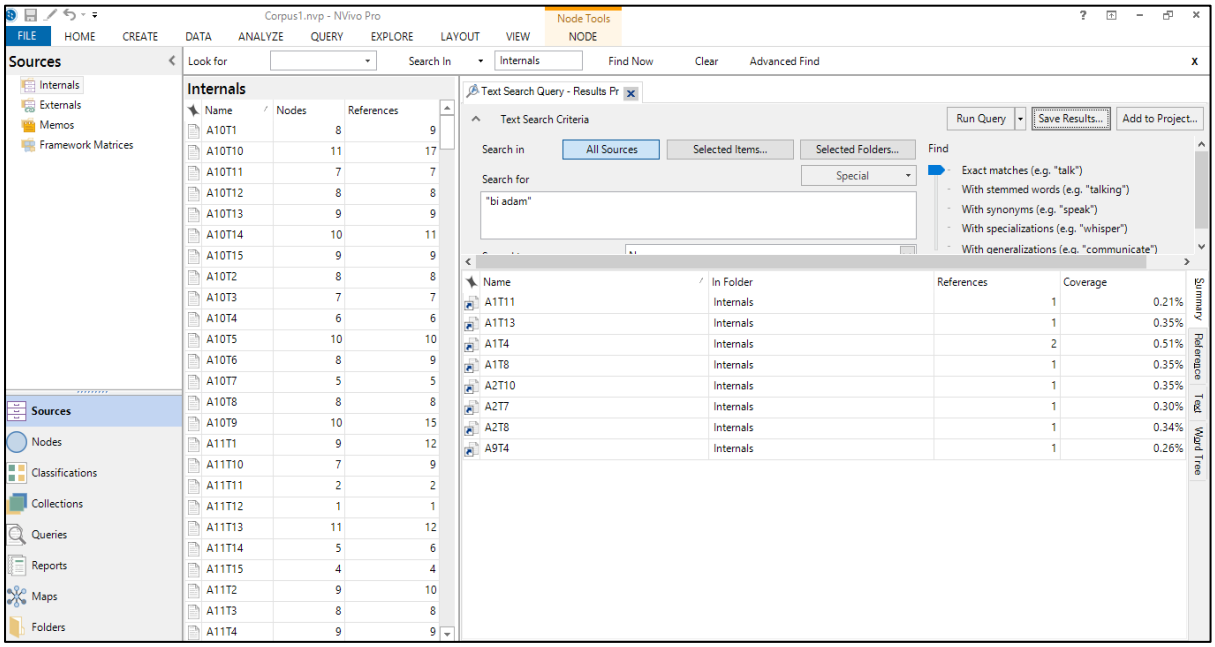
The Jaccard coefficient test does not produce meaningful results if one of the features occurs fewer than two times, however. For that reason, the frequency threshold is set to a minimum of two when selecting features. Any feature occurring fewer than two times is excluded as it would not be productive in attributing authorship. Because according to the Jaccard algorithm, there is no possibility to calculate the distance of a feature between the texts when it is occurred only once. For instance, an author used *yapiyo* – R Clipping feature one time only, and it was

not occurred again neither in his/her texts nor in the rest of the texts which belongs to the other authors. Detailed information is given about the Jaccard coefficient test in Section 4.4.

The maximum frequency threshold is set to 45, which is 20% of the texts. For instance, when a linguistic feature occurs 45 times in the corpus, it may be possible to find the feature more than one author's texts because each author has maximum fifteen texts. Each feature occurred more than 45 times show that that feature can be appear in other two authors also when it is considered the 20 percent ratio. Twenty percent of the total number of texts (225) is 45. It is expected that different features have different performance values. For instance, when a feature is used often, e.g. 120 times, which is 53% of the 225 texts, this means more than half of the authors have this feature in their texts and such a feature is not feasible in authorship attribution. It is worth noting that thresholds are set up manually rather than automatically based on the characteristics of the data and the limitations of the statistical method. Regardless of the frequency, there is no elimination system applied to exclude the less common features due to the small size of the texts. Considering the initial information regarding the rules involved in the coding process, top-down and bottom-up approaches will be used to find the linguistic features.

First, the top-down coding process is used for pre-defined style markers, such as the multi-level sophisticated features (Grant, 2013; Grant and MacLeod, 2012) and SMS and internet language features (Crystal, 2001; 2008). Although MacLeod and Grant's (2013) feature set for short text messages is used as one of the starting points for the current study, not all of them are applicable due to the genre and language origin restrictions. To select the data-driven features, the Sketch Engine corpus analysis tool will be used to extract the word lists and obtain an insight into the linguistic patterns and features automatically rather than intuition-based feature selection. After that selected features are found with using text query section within texts and automatically coded with the AutoCode function in Nvivo11. The text query function at Nvivo11 is presented at Figure 5-1.

Figure 5-1: A Display of Nvivo11 Text Query Function



As it is seen from the Figure 5-1, first the feature is inserted to the search section, then it provides the texts including that feature and finally it is possible to save the results as nodes.

The word lists provide the most frequent items occurring twice in the texts in order to obtain the important features of the study. The importance of automatic approaches in forensic linguistic analysis is mentioned by Larner (2014, p.8): 'automating the approach guards against human error, and this technique is also feasible for use with larger sets of data' as it speeds up the analysis no matter the size. The bottom-up coding process is used for the features that are neither pre-defined nor extracted from the word lists. For these kinds of features, an inter-coder reliability test is applied to reduce subjectivity (see Section 5.3).

After coding four datasets, numerous common features were found, therefore, a common linguistic features reference list was created.

Although it is hard to develop a taxonomy for each kind of language data, a reference list can be generated that shows the common features in Turkish online messages after the coding stage. Rudman (1998 p.360) stated that 'one of the most important facts to keep in mind is that each authorship study is different' and each studying require a unique expertise and experimental

design based on each author, each genre, each language and each time period. It is worth to note that, the reference list idea is not contrary to Rudman's approach.

A reference list may reduce the time spent coding by linguists analysing the texts. It should be noted that the proposed reference list developed from the current data set does not look for a generally accepted attitude and it may vary in different texts. There is no such thing as the best and most highly discriminative feature list as the reference list is mostly created to provide a starting point for future studies. Besides, this study does not represent a huge range of language communities. Due to the dynamic nature of language, it is not possible to fit a language study into a specific framework, therefore, the reference list should be open to improvements and amendments.

Finally, each feature was categorised as lexical, syntactic, or structural. This makes it possible to gauge the performance of each feature set in the analysis section. All texts were coded at the same time to reduce inconsistency between the codes for all corpora. All Nvivo projects for this study can be found in Appendices B on a CD.

5.2. Feature Classification

Feature classification is the most important step for authorship studies. In this section, the question: 'What stylistic features can discriminate between these texts by different authors?' (Grant and Baker, 2001, p.68) is asked to establish grounds for forensic linguistic studies in Turkish. As mentioned above, the features are classified into three groups (i.e. lexical, structural, and syntactic) to try and determine the discriminative capability of each feature in authorship analysis studies on Turkish texts. In contrast to previous authorship studies in Turkish, the feature sets are not fully automated and do not depend on computerised methods. Instead, this study aims to explain the features linguistically. The features found during the coding process are presented in the following sections.

5.2.1. Lexical Features

Lexical features are linked to words, word strings, vocabulary richness, n-grams, and word/sentence length. When only word/sentence length and vocabulary richness are considered, it is possible to do a language-independent analysis, however, applying the same method to different genres may be ineffective in small groups of texts. Besides, vocabulary

richness has been found to be an ineffective feature as 'the vocabulary used in short documents is usually limited and relatively unstable' (Zheng et al., 2006, p.382).

Grant (2007, p.6) also commented that, 'any classification of the query text…on the basis of the sentence and word length would, therefore, be based on a fallacy and so be incorrect'. Moreover, these complexity measures are inadequate when used on their own (Grieve, 2007).

Apart from complexity measures based on word length and sentence length, many studies have shown high accuracy with character n-grams, such as capitalisation or punctuation (Stamatatos, 2009). Word n-grams have been shown to be successful by many researchers in forensic linguistics (e.g. Wright, 2014; Nini, 2018). In contrast to previous studies based only on word n-grams, this research includes other types of features, such as lexical, structural, and syntactic features and word n-grams as text sizes are gradually decreasing and capturing the useful features for attributing authorship is important in short texts. It is worth noting that Wright (2014) used word n-grams for emails only when the text size was considerably larger than the current datasets.

As an initial step, word grams between two and six strings were chosen from the whole corpus, however, the findings returned with 600 results for Corpus1. Taking all word n-grams as features is therefore unmanageable and does not correlate with the research aim of identifying the role of feature types in authorship attribution in Turkish. It was decided to search for word n-grams that occurred between 10 and 45 times within the texts and include them in the lexical features list. Ten was used as the minimum frequency threshold by Grant (2013) in the analysis of text messages accordingly in this study, when the data is filtered with at least ten frequencies at the end 54 word-grams left. Surprisingly, some word n-grams are only used by a single author, however this is not the disputed author, e.g. *en az bir kez* (see Figure 5-1) is used by the same author ten times, therefore, this feature is excluded from the list as it does not have discriminative power between authors. This phrase appears ten times in two or three texts of Author 8 from Corpus1 and it does not occur anywhere else in the corpus. A8 is the only author in Eksi Sozluk corpus who used *en az bir kez* feature and this feature is unique to A8 which is not shared with anyone else. However, when it comes to the disputed author such type of features are not excluded from the list in order to find the distinctive features.

Figure 5-2 Word n-grams of between two and six words.

Word list

Corpus: Corpus1

<< First | < Previous Page 2

word (n-grams)	frequency
dün akşam	11
böyle bi	11
bu yüzden	11
bir insan	11
bir anda	11
benim için	11
az bir	11
andan beri	11
ama çok	11
yıl boyunca	10
ve daha	10
vay efendim	10
var ya	10
tek tek	10
olsa gerek	10
o an	10
için çok	10
en güzel	10
en fazla	10
en az bir kez	10
dediğim gibi	10
daha önce	10
bir kez	10
bir dönemde	10
bi insan	10
az bir kez	10
aynı zamanda	10

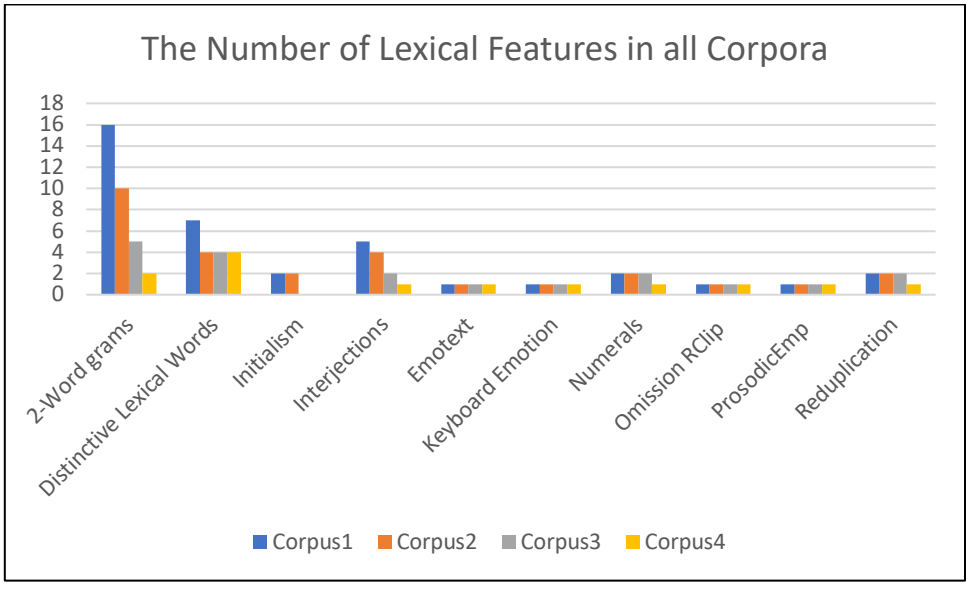
After discarding the word n-grams that were used by the same authors, sixteen two-word grams were left for the feature set. It is worth noting that the selection criteria for word n-grams is strictly limited because although a large number of word n-grams means the correct attribution of authorship is more likely (e.g. Wright, 2014), the aim of the study is to establish the efficiency of the feature sets in authorship attribution studies in Turkish. Ten two-grams, eight two-grams, and 2 two-grams were added to the feature list from Corpus2, Corpus3, and Corpus4, respectively. It is possible to observe the decrease in features depending on the text size.

R clipping, prosodic emphasisers, keyboard emoticons, and numeral features, (MacLeod and Grant 2012; Grant's 2013; Crystal 2004) were found with a top-down approach.

Among these features keyboard emoticons refer the characters which are 'typed in sequence on a single line and placed after the final punctuation mark of a sentence' in order to convey the feeling. (Crystal 2004 p. 36). The most fruitful data selection was based on data-driven features that were selected from the word lists, however.

To find discriminative lexical features, a top-down approach was applied depending on the analysis of the word lists. Although most of the lexical features were found from the word lists, some of the features in the word lists were excluded. For instance, idiosyncratic spelling was not considered to be a discriminative feature as Juola (2006, p.120) stated that 'people are also not consistent in their language and may (mis)spell words differently at different times' and 'one may be lucky enough to find a queried text with a large number of idiosyncratic spellings or grammatical constructions, but it makes little sense to test for a specific misspelling in every text' (Grant, 2004, p.48). Therefore, determining more sophisticated and reliable features is necessary for authorship attribution studies.

Figure 5-3 The number of lexical features used in the study.



As it is seen on Figure 5-3. at the end, 10 sub-categories were created under the name of lexical features as; word n-grams, distinctive lexical words, initialism, interjections, emotexts, keyboard emotion, numeral, omission of R sound, prosodic emphasiser and reduplication. Each category is presented with their general names for instance, even though there two different types of initialism and 16 different word n-grams, it is not presented individually in here.

However, in the analysis section, each one is represented as one feature. From the word lists, reduplications specific to the Turkish language were found, as well as interjections, which were used as discriminative features by Silva et al. (2011).

Following that, 28 features from Corpus2, 19 features from Corpus3, and 13 features from Corpus4 were elicited from the texts including ten, eight and two 2-word grams respectively. As the features were most common between corpora, a general reference list is presented with the aim of explaining the distinctive features used in the current study.

Despite the long lists of features in stylometric studies (e.g. Abbasi and Chen, 2008), some studies use short lists (e.g. Grant, 2013). Accordingly, the current research aimed to identify the most discriminative core features rather than thousands of features. The general definitions of the lexical features for all datasets are provided in Table 5-1.

Table 5-1: The reference list of lexical features.

Linguistic Features	Explanation	Example
Lexical Category\R clipping	Dropping the final 'r' of words	*Yapıyo, ediyo, gidiyo* (He goes)
Lexical Category\Abbreviation	The shortened form of a word or phrase.	*v.s. (ve saire)* (etc.) *v.b. (ve benzeri)*
Lexical Category\Initialism	The shortening of the written form.	Adalet ve Kalkinma Partisi *(AKP)* Welfare and Justice Party (WJP)
Lexical Category\Emotext	The written version of a laugh.	*Haha, hehe, aha ha ha, saasadsdsfdfdfd*
Lexical Category\Prosodic Emphasisers	Conveying specific pronunciation through spelling.	*Çooook* soğuk (Veeeeeery cold).
Lexical Category\Reduplication\Doubling in Lexical Format	This type of reduplication means the doubling of an entire word.	*Kıpır kıpır* durmadı. (He was like a jack-in-the-box.)
Lexical Category\Interjections	The features that are used to express a strong sense of feeling.	*Yahu! Ulan!* (Hey! Buddy!)
Lexical Category\Reduplication\M-Reduplication	M-reduplication also involves full reduplication, but the first consonant in the reduplicant is replaced with m-.	(*Sari* is the yellow colour in Turkish, *mari* has no particular meaning and is only used to improve the emphasis.) *Sari mari* giyme! (Don't wear yellow clothes.)
Lexical Category\Keyboard Emotions	Keyboard emotions refer to the emoticons conveyed with punctuation marks. It is stated above as Eksi Sozluk does not allow the use of emoticons so keyboard emotions are the only way to convey feelings.	*: (:)*
Lexical Category\Numbers	There are two types of number used in the features. First, mathematical objects to count and measure, and second, words used to imply the numbers.	*1, bir, 2, iki* (one, two)
Lexical Category\Word Grams Between Two and Six	Word strings between two and six words.	*Hayatta herkesin* (Everyone in their life) 2-word grams *Hayatta herkesin en az* (Everyone in their life at least) 3-word grams *Hayatta herkesin en az bir kez* (Everyone in their life at least once) 5-word grams

As it is mentioned above, regarding feature selection there is not a certain one type approach in this study. Both top-down and bottom-up approaches were used to select features for the task. First, top-down features were formed from the previous studies (e.g. MacLeod and Grant, 2012, Wright 2014) which have tested their efficiency in an authorship attribution task. Despite several studies used n-grams previously, word n- grams were included to the list to identify the distinctive word sequences. Word n-grams from 1 to 5 words were found with top-down approach. Furthermore, R clipping, prosodic emphasizers, keyboard emoticons, numerals, and emotexts features were used in previous studies and this study was further identified those features from the texts with a top-down approach. Even though abbreviation and initialism were on the same list, due to different stylistic usages these two features were found with a bottom-up feature selection. Finally, reduplications and interjections have highly discriminative power and it is not possible to guess the words and to do a top-down search in advance therefore, those features were found by bottom-up feature selection.

5.2.2. Syntactic Features

Syntactical construction choices can reveal clues about authorship (e.g. Stamatatos et al., 2000; 2001) and are linked with the punctuation and part-of-speech items. Many stylometric studies have shown that syntactic features are reliable in authorship attribution problems since they do not depend on conscious choices (e.g. punctuation, Baayen et al., 2002; Chaski, 2001).

Chaski (2005) classified the main punctuation types as simple punctuation marks and syntactically classified punctuation. Simple punctuation marks refers to the frequency of punctuation marks, while syntactically classified punctuation is counted by its syntactic meanings in the sentence. Punctuation is also used as a discriminative linguistic feature by McMenamin (2002) in forensic stylistics. In some studies, punctuation features taken as a unique category (e.g. McMenamin 2002) while some of them considered punctuation as an element of the syntactic category (e.g. Chaski 2001).

Aslan (2007) carried on a content analysis on 231 online forum postings between Turkish Language and Literature teachers and concluded that the non-standard usage of punctuation marks is done most frequently by Turkish users on the Internet. Thus, in the present study, punctuation marks are taken as single units (McMenamin, 2002) and standard and non-standard usages are extracted from the texts.

The bottom-up reading method was applied for both standard and non-standard forms of semicolons, colons, exclamation marks, question marks, apostrophes, ellipses, quotation marks, and full stops. Due to the high standard usage of the full stop, it is not considered as a standard feature. Non-standard usage demonstrates mistakes and errors, for example, Muysken (1998, p.57) stated that, 'Mistakes and errors might be creative and random behaviours that have the unintended effect of profoundly altering the relationship with a system'.

Unlike these features, spelling mistakes are excluded from the feature list as they are unreliable indicators of authorship (Foster, 2001) and it is difficult to replicate a study based on mistakes.

Non-standard versions of punctuation are different to non-standard spelling variations, however. To avoid any problems with coding the punctuation marks, manual coding was done for this feature. Mixed typographic punctuation, multi-typographic punctuation, exclamation marks, and question marks (MacLeod and Grant, 2012; Grant, 2013) were also found using a bottom-up approach. The multiple uses of punctuation in CMC has been defined as an 'emergent practice and convention of digital writing' (Danet, 2001, p.17).

Similar to the lexical features, the frequency threshold of this study is set at two, which is the required minimum value for the Jaccard test. Furthermore, part-of-speech (POS) tagging, which marks a word according to its category, provides a good source for capturing authors' styles, even in short texts. It is 'a system that cannot distinguish between contraction apostrophes and closing single quotes or that can only tag with 95% accuracy will conflate entirely different syntactic constructs' (Juola, 2008, p.265). Automated POS tagging applications are still relatively new, however, and struggle with Turkish due to its agglutinating structure. This property of the language leads to different tags being assigned, from nouns to adjectives and adverbs to propositions. Also, many words in Turkish can have an infinite number of suffixes, e.g. *'muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmişsinizcesine'* (as though you are from those whom we may not be able to make into a maker of unsuccessful ones easily) is derived from the noun 'muvaffakiyet' (success) and has 17 suffixes attached to the word itself.

Online texts are also more challenging for POS tagging than conventional texts due to the former's non-standard structure. POS tagging is used for tweets (Gimpel et al., 2011, p.46) and 'the system also struggles with the miscellaneous category, which covers many rare tokens, including obscure symbols and artefacts of tokenisation errors'. To improve the previous results, Owoputi et al. (2013) created a new Twitter POS tagger that includes large-scale

distributional features, however, this study increased the accuracy three per cent more than the previous (Gimpel et al., 2011) research.

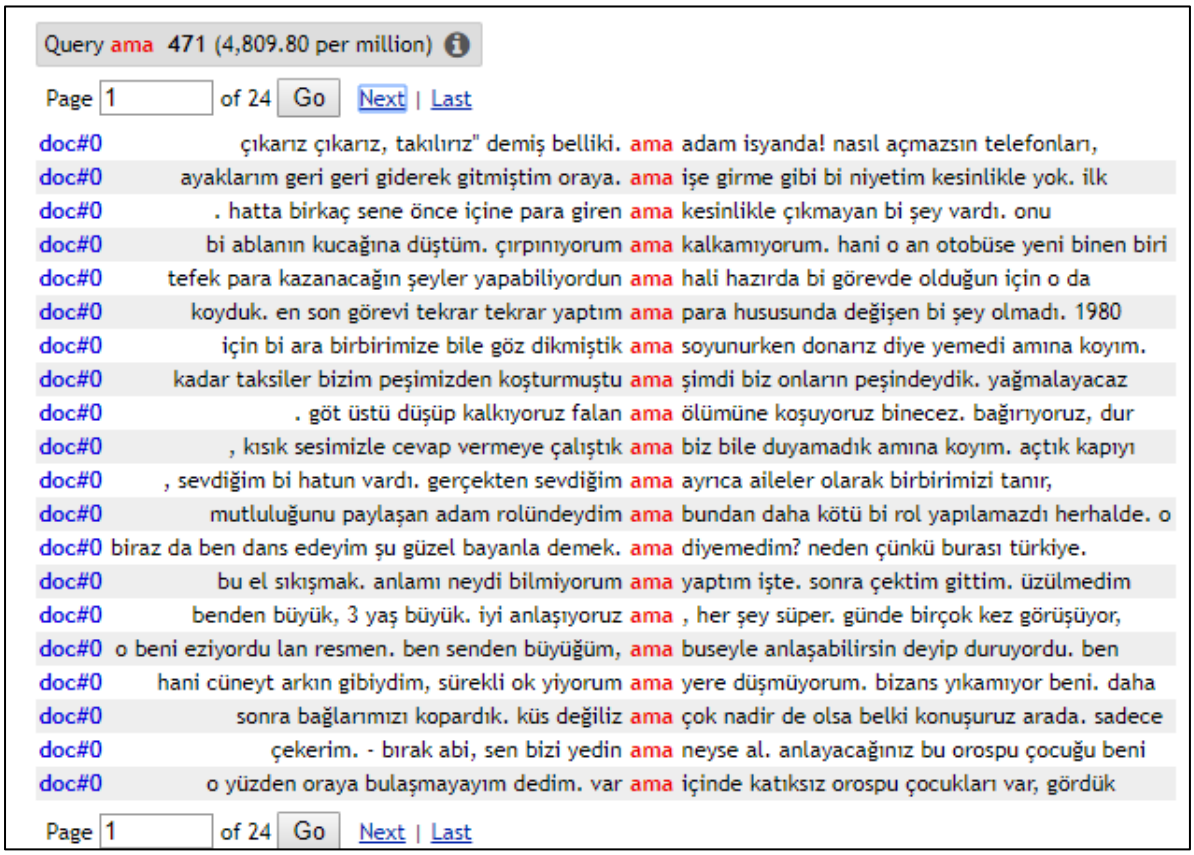
Therefore, rather than using POS tags, some syntactic features are derived from the word lists. According to the wordlists, most of the conjunctions are considerably frequent within the whole list. Conjunctions are linguistic elements that connect two variables in the text (Halliday and Matthiessen, 2004) and they are 'an indication of how the author organises concept and relates them to each other' (Argamon and Koppel, 2013, p.303).

There are different types of conjunctions according to their role within the textual context. When conjunctions have rhetorically connected the sentences, this is the function of internal conjunction, while for external conjunction, there is no conjunction for relating two sentences (Halliday and Hasan, 1976).

The presence or absence of explicit conjunctions is one of the principal variables in English discourse, both between registers and between texts in the same register (Halliday and Matthiessen, 2004). Halliday and Matthiessen (2004) described the top three types of conjunction relationships in the system as (i) elaboration, e.g. 'for example', 'for instance', and 'to illustrate'; (ii) extending, e.g. 'and', 'moreover', and 'furthermore'; and (iii) enhancing, e.g. 'hence' and 'consequently'.

Similar to Halliday and Matthiessen (2004), Argamon and Koppel (2013, p.302) categorised conjunctions as 'elaboration, extension', and 'enhancement'. For this kind of categorisation, however, it is necessary to use the semantic meaning of the conjunctions. Instead of, syntactically classified conjunctions, the positions of the conjunctions are focussed on within the sentence. The Sketch Engine concordance function was used to find the position of the conjunction without considering the literal meanings. In Figure 5-2, the concordance of the conjunction 'ama' (but) is presented and for the study, it is coded according to its position in the sentence, e.g. StrConj (starting the sentence conjunction), ConConj (connecting the words/sentences in the middle), and EndConj (ending the sentence Conjunction).

Figure 5-4 Concordance of ama.



Query ama 471 (4,809.80 per million)

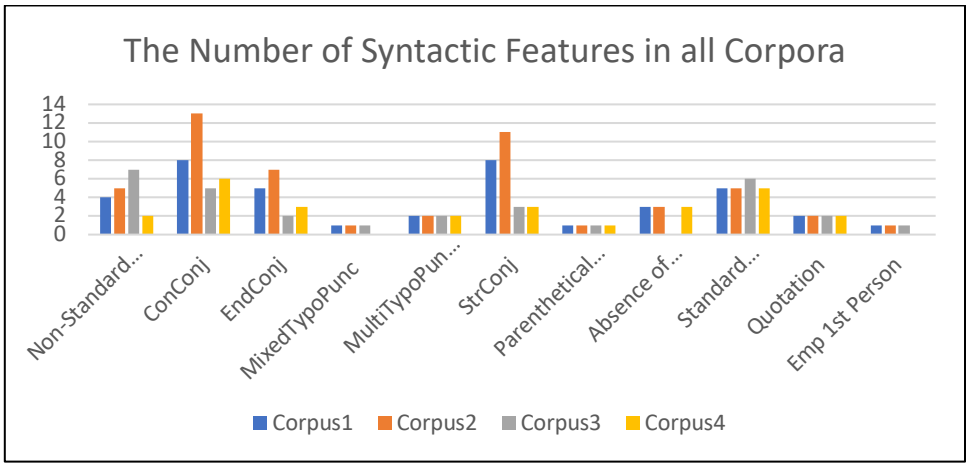
Page 1 of 24 Go Next | Last

doc#0	çıkarız çıkarız, takılırız" demiş belliki.	ama	adam isyanda! nasıl açmazsın telefonları,
doc#0	ayaklarım geri geri giderek gitmiştim oraya.	ama	işe girme gibi bi niyetim kesinlikle yok. ilk
doc#0	. hatta birkaç sene önce içine para giren	ama	kesinlikle çıkmayan bi şey vardı. onu
doc#0	bi ablanın kucağına düştüm. çırpınıyorum	ama	kalkamıyorum. hani o an otobüse yeni binen biri
doc#0	tefek para kazanacağın şeyler yapabiliyordun	ama	hali hazırda bi görevde olduğun için o da
doc#0	koyduk. en son görevi tekrar tekrar yaptım	ama	para hususunda değişen bi şey olmadı. 1980
doc#0	için bi ara birbirimize bile göz dikmiştik	ama	soyunurken donarız diye yemedi amına koyım.
doc#0	kadar taksiler bizim peşimizden koşturmuştu	ama	şimdi biz onların peşindeydik. yağmalayacaz
doc#0	. göt üstü düşüp kalkıyoruz falan	ama	ölümüne koşuyoruz binecez. bağırıyoruz, dur
doc#0	, kısık sesimizle cevap vermeye çalıştık	ama	biz bile duyamadık amına koyım. açtık kapıyı
doc#0	, sevdiğim bi hatun vardı. gerçekten sevdiğim	ama	ayrıca aileler olarak birbirimizi tanır,
doc#0	mutluluğunu paylaşan adam rolündeydim	ama	bundan daha kötü bi rol yapılamazdı herhalde. o
doc#0	biraz da ben dans edeyim şu güzel bayanla demek.	ama	diyemedim? neden çünkü burası türkiye.
doc#0	bu el sıkışmak. anlamı neydi bilmiyorum	ama	yaptım işte. sonra çektim gittim. üzülmedim
doc#0	benden büyük, 3 yaş büyük. iyi anlaşıyoruz	ama	, her şey süper. günde birçok kez görüşüyor,
doc#0	o beni eziyordu lan resmen. ben senden büyüğüm,	ama	buseyle anlaşabilirsin deyip duruyordu. ben
doc#0	hani cüneyt arkın gibiydim, sürekli ok yiyorum	ama	yere düşmüyorum. bizans yıkamıyor beni. daha
doc#0	sonra bağlarımızı kopardık. küs değiliz	ama	çok nadir de olsa belki konuşuruz arada. sadece
doc#0	çekerim. - bırak abi, sen bizi yedin	ama	neyse al. anlayacağınız bu orospu çocuğu beni
doc#0	o yüzden oraya bulaşmayayım dedim. var	ama	içinde katıksız orospu çocukları var, gördük

Page 1 of 24 Go Next | Last

After coding the defined features with top-down approach, a bottom-up reading was performed to identify the features with the high discriminatory power which was reported in the previous studies and also language dependent. Thus, features such as dropping *ben* (I), the first-person pronoun, parenthetical clauses, and quotations were located and it was observed that such features were unique to some authors and not a concern for others. Ustunova (2002) discussed dropping *ben* (I), the first-person pronoun, within the sentence and it is stated that there is a tendency to drop the pronoun in Turkish as the verb already includes a suffix that presents the agent. For instance, there is no semantic difference between *Ben yaptim* (I did) and *yaptim* (I did) in Turkish. The final *m* sound indicates the first-person pronoun, nevertheless, some authors tend to emphasise the verb with the first-person pronoun while others do not.

Figure 5-5 The number of syntactic features used in the study.



As it is seen on Figure 5-5- after adding these features, a total of 40 features from Corpus1, 51 features from Corpus2, 30 features from Corpus3, and 27 syntactical features from Corpus4 were elicited from the texts. As the features were most common between corpora, a general reference list is presented with the aim of explaining the distinctive features used in the current study. It is worth noting that general feature names are used in the reference lists rather than stating each feature. For instance, *ama StrConj* is not presented, but starting conjunctions are mentioned.

This is because this type of clarification is useless for future studies as the conjunction type may vary in other texts, but the function of it within the sentence does not. The general definitions of the syntactic features for all datasets are provided in Table 5-2.

Table 5-2: The reference list of syntactic features.

Linguistic Features	**Explanation**	**Examples**
Syntactic Category\Syntactic Category\Parenthetical Clause	A parenthetical clause includes a clause that is inserted into the sentence and that interrupts another phrase or clause.	Iki kedi *(beyaz, siyah)* vardi. (There were two cats (white, black).
Syntactic Category\Punctuation\Absence of Punctuation - Apostrophe	Omission of the apostrophe.	*Huseyine* gel demistim, gelmedi. (I told Huseyin to come, but he didn't.)

Syntactic Category\Punctuation\Absence of Punctuation - Full stop	Omission of a full stop.	Huseyin'e gel demistim *gelmedi (*I told Huseyin to come, but he didn't)
Syntactic Category\Punctuation\Absence of Space after Punctuation	The absence of a space after punctuation.	Huseyin'e gel dedim *gelmedi.Biz* de ayrildik. (I told Huseyin to come, but he didn't, so we broke up)
Syntactic Category\Punctuation\Exclamation Mark in Parenthesis	In Turkish, an exclamation mark in parentheses may imply a hint.	Cok basarililar *(!)* ya! (There are so successful (!))
Syntactic Category\Punctuation\Mixed Typo. Exclamation	Use of mixed characters to convey an exclamation.	Yaa*!!?!!!?*
Punctuation\Multiple Question Mark	Use of multiple question marks.	Ne*??????*
Punctuation\Multiple Typographic Exclamation	Use of multiple exclamation marks	Hadi canim*!!!!!* (No way!!!!!)
Syntactic Category\Punctuation\Non-Standard Punctuation - Double full stop	Use of a double full stop instead of a full stop at the end of the sentence.	Geliyorum.. (I am coming..)
Syntactic Category\Punctuation\Standard Punctuation - Semicolon	It indicates a pause, typically between two main clauses.	*Gelirken;* ekmek, peynir al. (On your way back, buy bread and cheese)
Syntactic Category\Punctuation\Standard Punctuation - Slash	An oblique slanting line.	*Evet/Hayir* (Yes/No)
Syntactic Category\Punctuation\Standard Punctuation - Ellipsis	It indicates where words have been omitted from quoted text, or (informally) to represent a pause, hesitation, or trailing off in thought or speech (Farlex Grammar Book, 2016).	Bekliyorum... (I am waiting…)
Syntactic Category\Conjunctions\Starting Conjunctions	It compares phrases and clauses.	*Ama* ben oyle demedim. (But, I didn't say so.)
Syntactic Category\Conjunctions\Connecting Conjunctions	It joins words, phrases, and clauses together that are usually grammatically equal.	Onu al *ama* bunu alma. (Take this but not the other one.)
Syntactic Category\Conjunctions\Ending Conjunctions	They are used to join equal sentence elements together.	Gel dedim *ama.* (I said, 'come' but.)
Syntactic Category\Quotes	Quotes are used to indicate a quotation. As mentioned before,	*'Asla'* dedi. (He said 'never'.)

	quotes are not considered during the coding as they are not the production of the authors. Using quotes often may lead to intra-author distinctiveness both in the current corpus and in Turkish text, however. Quotes are divided into two sections: word quotes that quote only one word, and sentence quotes that are fully quoted.	*'Ben artik yuruemem'* diye agladi. (He cried that I would never walk again.)
Syntactic Category\Emphasis 1st Pronoun	Using the first-person pronoun in a sentence.	*Ben* geldim. (I arrived.)

Some of the studies mentioned above have tested the performance of punctuation features and have obtained good results in authorship attribution tasks (e.g McMenamin 2002). Moreover, MacLeod and Grant (2012) performed mixed typo exclamations, multiple typographic exclamation and multiple question mark features in their study and achieved high level of accuracy. As it was explained before, Chaski (2002) used the syntactic classifications of punctuation marks (e.g. End of Clause Punctuation, End of Phrase Punctuation) however, the results were found problematic by some researchers. Further that, Argamon and Koppel (2013) successfully employed the conjunctions in their study however, they categorised the conjunctions according to their semantic meanings. Therefore, it is decided to acknowledge both methods and used the most suitable features based on the research aims. That is to say, semantic meanings of the conjunctions are disregarded, and these features are classified according to their syntactic positions i.e. *StrConj, ConConj, EndConj*. as Chaski (2002) classified punctuation in her study. Finally, Emphasis 1st Pronoun feature is considered as a potentially useful feature depending on the nature of the language.

5.2.3. Structural Features

Structural features are related to the genre of a text and this can vary from text to text, e.g. sentence and paragraph organisation. For instance, signatures, farewells, or greetings in emails or text frames that address another subject may be regarded as structural features. A subcategory of structural features used for the study is technical structure features, which includes font, hyperlinks, and embedded image characteristics (Abbasi and Chen, 2005). According to Abbasi and Chen (2005), inserting images, icons, or hyperlinks into text may

reveal the author's technical ability. At the end of their study, it was concluded that structural features are important in identifying messages in web forum messages.

Furthermore, Zheng et al. (2006, p.380) stated that 'people have different habits when organizing an article' and these habits may be strong evidence of authorship. These habits also have 'less content information but more flexible structures or richer stylistic information' (Zheng et al., 2006, p.380) in online texts. Structural features may be especially important in online texts like e-mails, which include farewells and signatures (e.g. De Vel et al., 2001). Moreover, some web pages 'give authors control over text formatting and layout, including specific aspects such as font choice font sizing, placement of figures and clip art, or the use of colour' (Juola, 2008, p.266). Eksi Sozluk entries do not have certain structures, e.g. greetings, however, and it is not possible to arrange the text in terms of font type, size, capitalisation, etc. Although Twitter has specific structural features, such as @reply, retweets, and hashtags, none of them were selected for Corpus4 - Cross-Genre Analysis as these features do not have equivalents in Eksi Sozluk.

Paragraph organisation is one of the limited structural features allowed by the website, but the largest dataset in this study has a maximum of 505 words, which is very short when divided into several paragraphs. For that reason, paragraph length is not a feature for this study however there are some Eksi Sozluk specific features such as *bkz* (see also), *swh* (sarcastic laugh) which are used as structural features. Structural features, including *bknz, spoiler, hyperlinks,* and *swh* were coded via the Auto Code function in Nvivo11. To capture forms of itemisation other than bullet points, however, a bottom-up approach was used in the feature selection.

Table 5-3: The number of structural features used in the study.

The Number of Structural Features for all Corpora				
	Corpus1	**Corpus2**	**Corpus3**	**Corpus4**
bkz	1	1	1	1
spoiler	1	1	1	1
Hyperlink	1	1	1	1
Itemize	1	1	1	1
swh	1	1	N/A	N/A

As seen on Figure 5-6, only five structural features were found in the datasets, obtaining meaningful results when the distance measures are calculated could be problematic. Therefore, structural features are included only because they improve the performance in general.

Although their contribution has a less discriminative effect than lexical and syntactic features, they can still have a significant cumulative impact.

Finally, the reference list of structural features is not going to provide information for the future studies unless the data set is collected from Eksi Sozluk. A helpful structural reference list is only possible to outperform different features in a wide margin from e-mails to online blogs. The general definitions of the structural features for all datasets are provided in Table 5-3.

Table 5-4: The reference list of structural features.

Linguistic Features	Explanation	Examples
Structural Category\Itemise	Listing individual items within a text. Some authors tend to give definitions as a sequenced list. This feature occurs on Twitter occasionally while the users talk about serial stories. For that reason, it can be counted as one of the features in the reference list.	Bullet pointing e.g. Ogretmenlerin Ozellikeri (The Characterisations of Teachers) • *Az calisip, cok tatil yapmak* (They work less and have long holidays)
Structural Category\Hyper Links	Embedded hyperlinks in the text.	In Eksi Sozluk, there is a smart tool that enables external web links to be added into a text without writing the whole address. The Sketch Engine corpus tool still recognised the hyperlinks inserted in the texts, however.
Structural Category\Bknz Redirecting	Referring to another title or information.	*Bkz.eksi elmalar* (See also: Sour Apples – which is a popular Turkish movie name)
Structural Category\Spoiler	Apart from the meaning of the cinema industry, it refers to giving previous information in anything. The standard usage of 'spoiler' is presented as in the example.	*Spoiler* =========
Structural Category\Swh	Swh refers to a kind of sarcastic laughter in Eksi Sozluk. Generally, authors write their entries and refer (*) to it as if it is a separate title. This type of laughter does not exist in any CMC medium, however, so it is selected as a structural feature.	*swh*

5.3. Reliability

This study measures the validity of the proposed linguistic features by measuring the level of agreement between four different coders in the field of linguistics. Due to the subjectivity risk in selecting and coding features, the inter-coder reliability test is essential for improving the

reliability of the coding scheme. Finding features based on previous feature lists and applying an inter-coder reliability procedure may reduce the bias in coding. Moreover, the codes in three different datasets share similar linguistic structures, which indicates the consistency in the coding. In real-life cases, inter-coder reliability may be a problem due to the issue of confidentiality, however, for this study, it was necessary to set up an external control mechanism to decrease the subjectivity level.

5.3.1. Inter-coder Reliability Test

According to Krippendorf (2004a, p.215), 'agreement is what we measure; reliability is what we wish to infer from it'. In this section, the reliability of the feature selection through the coding of the texts was tested using an inter-coder reliability test to enhance the quality of this study and reduce the subjectivity of the researcher. The inter-coder reliability test aims to produce the same selection as the other coders as coders' judgements may vary among individuals. It refers to 'the extent to which the different judges tend to assign the same rating to each object' (Tinsley and Weiss, 2000, p.98).

First, a codebook is prepared that includes the feature definitions and examples to strengthen the notion. According to the codebook, the second coders read the instructions and code the features by entering their answers into an online survey web tool. The aim is to demonstrate the reliability of the coding by assigning the same value between coders for the given texts. Inter-coder reliability survey questions asked in Turkish by using the same question format as follows: 'Please find the related linguistic feature or features from the text below.' As a sample sentence would have more than one linguistic feature, each question had five options and it was possible to give multiple answers.

Second, selecting enough material for the test and third applying best reliability test which is suitable for the number of the coder. Setting up a reliable and valid method is critical because in this study, it is hypothesised that results may be presented at the courtroom. Furthermore, having an agreement regarding the linguistic feature coding would also help establish a useful categorisation systematic for future studies.

It is worth noting that, since there are some features which coded with bottom-down approach it is necessary to find out the reliability of those features. Furthermore, an automatic feature selection approach is not applied to the study thus, it is acknowledged that they may be a

potential problem in coding. However, a second look from another researcher is used to minimise the researcher's bias in feature selection in this study.

For this study, three coders were recruited who had similar educational backgrounds. Ten percent of the data (22 out of 225 random texts) was selected as samples to be given to the second coders. The coding process started after the coders had read the codebook instructions, including explanations of the features. The second coders then went to the online survey web page.

Table 5-5: Description of the second coders.

Coders	Educational Background	Native Language
Coder 1 – Researcher	PhD Researcher in Forensic Linguistics	Turkish
Coder 2	PhD Researcher in Language Documentation	Turkish
Coder 3	PhD Researcher in Applied Linguistics	Turkish
Coder 4	PhD Researcher in English Drama and Literature	Turkish

Although the coders had similar educational backgrounds, a codebook was prepared to avoid any misconceptions (Eisenmann et al., 2013). In the codebook, coders could find a list of the features, their definitions, and an example to improve their comprehension. The three coders and the researcher then coded 22 randomly selected texts from the Corpus1 dataset in terms of syntactic, structural, and lexical features. Once the coding was complete, the results were converted into .csv files, with 1 indicating agreement between answers and 0 indicating disagreement.

We can say that the lexical features are reliable for the current study, but who ensures the reliability of lexical features in the first place. To avoid the risk of subjectivity, an inter-coder reliability test was applied to three coders. Recal3 (Reliability Calculator for three or more coders) is 'an online utility that computes intercoder/interrater reliability of coefficients for nominal data coded by three or more coders' (Dfreelon.org, 2013).

5.3.2. Intercoder Reliability Test Results

To establish the coding reliability, pairwise percent agreement was calculated for the study.

In pairwise percent agreement, 'coefficients of .90 or greater are nearly always acceptable, .80 or greater is acceptable in most situations, and .70 may be appropriate in some exploratory studies for some indices' (Neuendorf, 2002, p.145).

Lombard et al. (2002) suggested that agreement should be at least 90 percent due to the weaknesses of the analysis. The average pairwise percent agreement was 83.33%, which is acceptable for Neuendorf, but not for Lombard et al. (2002). There are criticisms of average pairwise percent agreement as a reliability measure, however, as it calculates the overall agreement while ignoring some disagreements. Every value is important in inter-coder tests.

Table 5-6: Average pairwise percent agreement results.

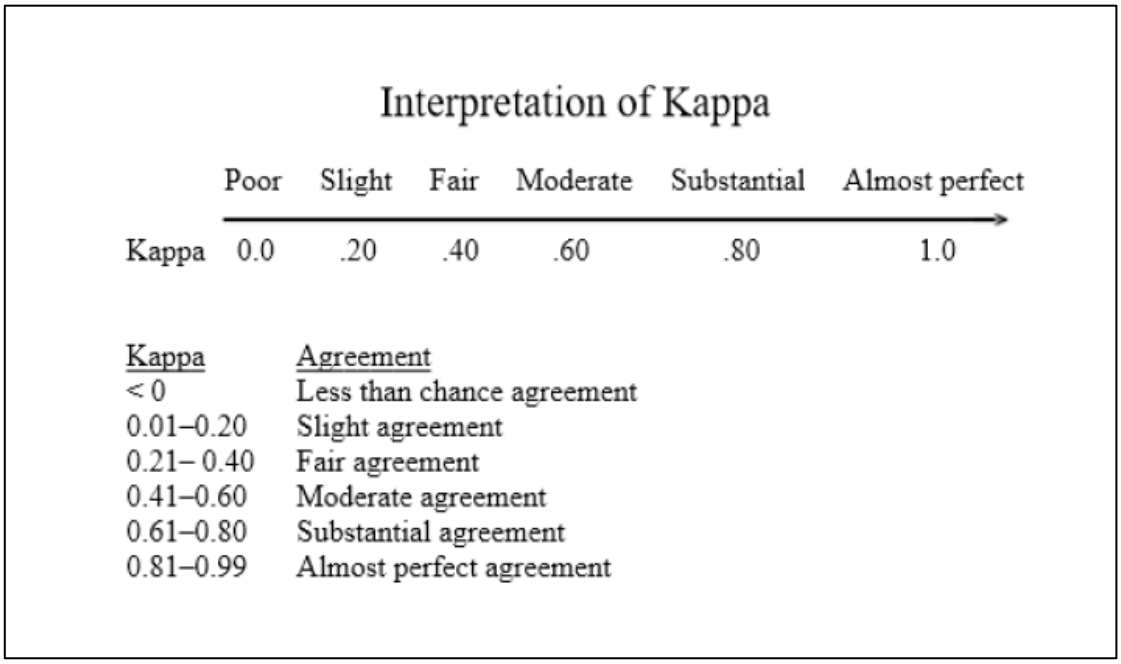
Average Pairwise Percent Agreement						
Average pairwise percent agr.	Pairwise pct. agr. cols 1 & 4	Pairwise pct. agr. cols 1 & 3	Pairwise pct. agr. cols 1 & 2	Pairwise pct. agr. cols 2 & 4	Pairwise pct. agr. cols 2 & 3	Pairwise pct. agr. cols 3 & 4
83.333%	86.364%	95.455%	81.818%	77.273%	77.273%	81.818%

Table 5-7: Fleiss' Kappa results.

Fleiss' Kappa		
Fleiss' Kappa	Observed Agreement	Expected Agreement
-0.008	0.833	0.835

Second test is Fleiss' Kappa which is eligible for more than two coders. The equation for Fleiss Kappa is given in Table 5-6. According to the interpretation of Kappa (see Figure 5-7) there is almost perfect agreement between coders.

Figure 5-6: The interpretation of Kappa



5.4. Summary of Feature Selection

In traditional forensic investigations, matching shoe marks, blood types, or DNA are the primary forms of evidence.

In forensic linguistics, however, there is no definitive comparison method for future cases. As this study claims to be the first study related to forensic linguistics in the target language, it should be valid and reliable in terms of reproducibility. Therefore, the most important features for attributing authorship, including the data-driven features specific to the language and genre, have been investigated.

To decrease subjectivity, a second coder reliability test was applied. In real-life cases, however, a second coder may be a problem due to the issue of confidentiality. At this juncture, applying a second coder reliability test to develop a reliable initial reference list for synthetic forensic linguistic cases may guide real-life cases. Moreover, it is evident from the results presented above that the second coders have sufficient agreement to ensure the reliability of the feature selection. The performance of the selected feature sets is explained regarding different text lengths in the following chapter.

Chapter 6: Analysis and Discussion

The first approach of the analysis compares three different combinations of the features presented in the previous chapter, to evaluate their performance for attributing authorship. Section 6.1 presents the evaluation procedure before interpreting the results. The role of lexical features (Section 6.2.1) is examined in the first study. Then syntactic (Section 6.2.2) and structural features (Section 6.2.3) are tested. Section 6.3. presents an overall conclusion of the section.

The second approach addresses text size (Section 6.4), including Corpus 1- Long Text (Section 6.4.1.1.), Corpus 2 – Medium Size Texts (Section 6.4.1.2.) and Corpus 3 – Small Size Texts (Section 6.4.1.3.) sections. This approach also investigates the functions of the number of candidate author set size (Section 6.4.2) and limited texts available per author (Section 6.4.3) in authorship attribution in Turkish. Finally, the results of the cross-genre analysis (Section 6.6) are presented. In the end, a general conclusion is drawn upon the overall results (Section 6.7.)

In the previous chapter, potential linguistic features are presented according to their function in a sentence. Three main analytical approaches and their corresponding findings were outlined in this chapter. The overall purpose of this study is to develop a methodology which combines stylistic and statistical approaches for the Turkish language. Since stylometric approaches are considered to have a lack of linguistic theory behind them (Grant, 2008), they are unreliable in this aspect. This study aims to apply statistical methods and interpret the results linguistically. With this method, we can eliminate intuition and subjectivity and establish a meaningful interpretation in terms of linguistics from the analysis. First, the linguistic material was acquired from a reliable source parallel with the research aims and their extraction from the website and coding process was semi-automated. Then, the feature selection approach was useful in avoiding subjectivity with using the predefined feature lists and Turkish Internet language features which were found with a bottom-up approach.

Moreover, the second coders proposed their opinions in feature coding which were found substantial agreement between the coders. Finally, the analytic method with the statistical presentation of the results reduced the potential of subjectivity and intuition. This study aims to identify a linguistically defensible, forensically feasible method that is sensitive to real-world case limitations and has a statistically testable approach, by following the combined method

proposed by Grant (2013). A real-world forensic linguistic task will be simulated, involving a set of known texts and a disputed author. The texts include different numbers of words and are short when compared to traditional authorship attribution data. First, each text is treated individually and then cumulatively per author. All available texts from one author are combined and presented a combined representation of that author's style from these concentrated texts. Thus, concentrated texts may produce reliable results when it is compared with individual short texts. In treating every text individually, the similarity between other available text samples rather than the disputed author in the corpus is ignored. Furthermore, the approaches are intended to test the various conditions, as actual cases may represent similar characteristics regarding the text size, the number of candidate author set size or an available number of texts per author or cross-genre applications.

According to the authorship attribution scenario presented here, there are fifteen texts from each known author and the same number of texts from one unknown author in all studies. In total there are 14 distinct authors. The unknown author is represented between known authors; thus thirty texts in total belong to the disputed author. In each corpus, disputed authors are different due to the reason of text size differences. Each author has different writing habits some write longer texts while others prefer shorter ones instead of dividing the texts into the two parts, the texts intrinsically produced according to the need of the study chosen. For that reason, 4 different disputed authors depend on their writing habits in each corpus. Moreover, the order of the corresponding and disputed author is determined differently in order not to cause confusion between corpora and randomly selected. For instance, in Corpus1 the disputed author is Author1 while in Corpus2 it is Author12, it is followed a mixed sequence in choosing the disputed author.

In determining the actual author, Grant (2007, p.21) proposed three categories as 'correct classification, no classification and misclassification' in defining the results of log-likelihood. It is possible to use such a classification when there are results based on percentages. Similarly, most of the stylometric studies interpret the results as being at a *high level of accuracy* or a *low level of accuracy.* In some stylistic studies, the linguistic expert uses a scale which shows a level of exclusion, identification or inconclusiveness in presenting the results to the court. Coulthard et al. (2017) interpreted results on an 11-level scale: from five, meaning 'satisfied'; to minus five, meaning 'not satisfied'.

Similarly, McMenamin (2002) used a nine-point scale to represent the resemblance between questioned and known authors. Chaski generalised the possible conclusions in different forensic science fields:

> '(i) the known and questioned items belong to the same person (identification/inclusion); (ii) the known and questioned items belong to different persons (elimination/exclusion); (iii) the analyst is unable to state a conclusion about authorship based on the current set of known items (inconclusive evidence/no conclusion)' (2001, p.5).

Furthermore, Love (2002) categorised the possible outputs as follows: (1) an assured attribution – when there is no doubt; (2) a confident attribution – supported with evidence of a range of good positive results with one anomalous negative; (3) a tentative attribution – the result of a process of arbitration between one body of evidence in favour and another against; (4) a plausible speculation – a piece of argument which falls a good way short of establishing an attribution but might prove valuable if further data were to emerge (p. 216). Anything below this level is not worth further investigation (Love 2002 p.216).

The current study aims to test whether the known and questioned items belong to the same person, to give probability judgements in authorship attribution. Grant (2013) presented the results based on notions of *consistency* and *distinctiveness*, which means an author has a sufficient degree of consistency of style and the writing is distinctive from the rest of the authors and stated that absolute consistency is not necessary in such cases. The notions of consistency and distinctiveness across the features are considered as an indicator of authorship. That is to say, any feature which occurs a single time in a single text is not considered as an indicator; however, when it repeatedly occurs throughout the texts is a remarkable phenomenon. Sketch Engine software was enabled to create wordlists to find consistent linguistic features, and Jaccard distance test is one of the most suitable tests in order to analyse the consistency and distinctiveness throughout the texts. Furthermore, in order to detect consistency and distinctiveness between authors two-phase analysis was required in this study as; text vs text and author vs author. Thus, the results of this study will be presented in a similar manner.

6.1. Evaluation Procedure

To conduct a fair comparison between texts, it is necessary to state the basic statistical evaluation procedures applied in the study.

In the Jaccard Distance test, the distance between vectors depends on the vector space, which is 0 – identical and 1 – different. In this case, any value which is close to 0 between others may be a starting point for further progress. Since the maximum distance value should provide an upper bound for texts (1), and the minimum distances should provide the lower bound (0) for the texts. The probability of becoming the disputed author is considered that if a value has the closest distance to 1, it means a misattribution while the values lower than this has accurately attributed potential. Those close values to 1 are unlikely to be the candidate author, while those far from the 1 are likely to be the candidate author.

Furthermore, heatmaps suggested a tendency for similar patterns, in which smaller distances between texts in the distance matrix. This highlights the fact that closer distance is a good substitute for reliable distance values in attributing authorship.

Moreover, in some studies, a threshold is established in the Jaccard coefficient test. For instance, Niwattanakul et al. (2013) applied the Jaccard Coefficient in a study that performed well in measuring the similarity of words when comparing each letter of the word. The acceptance criteria for similarity was settled at 0.75 and reported that the similarity is computed effectively in this study. Moreover, Wright (2014) discussed the development of a threshold for Jaccard scores which can identify a level of similarities and for the rest it could be concluded as unreliable and continued '[t]his would provide a more nuanced set of results as opposed to binary correct/incorrect attribution and would reduce the number of misattributions' (p.188).

However, in this study, the smallest mean Jaccard scores for text vs text comparison and raw Jaccard distance values for author vs author comparison are used to estimate the disputed author correctly. In text vs text comparison, the Jaccard mean distance values were calculated for all texts from all authors. The smallest mean distance values between disputed author and one of the possible authors are transferred into another file for further investigation. When the disputed author has the smallest score with one of the other authors and this author is the corresponding author who established before then this attribution test had been correct classification; otherwise, it is called misclassification. In some cases, there are some mean Jaccard values which are small among the texts. In such a case, when an author hit the lowest mean at least 5 times is considered as an evidence for attribution.

In author vs author comparison, the smallest raw Jaccard distances between disputed author and the other authors are hypothesised to be the author of the unknown texts. The more dissimilar texts have a higher distance value compared to one another.

Furthermore, shared features between disputed and the potential authors are presented in two ways. First, depending on its size the frequency threshold is applied in each corpus, for instance in Corpus1-Long Texts the features which are used at least ten times between texts and in Corpus2- Medium Size and Corpus3- Short Size texts at least five shared features are presented, while in Corpus4- Cross-Genre all available common features are shown due to restrictions of its size and the frequency. The same procedure is applied for the sub-training texts such as Ten Texts per author, Five Texts per author similar to Corpus4-Cross-Genre. However, for the 30 authors test the frequency threshold for shared features are limited to ten again. This procedure is applied due to the likelihood of presence any feature is correlated with text length. Second, when two authors do not use a feature at the same time, it is considered as a shared feature; thus shared zero values are also mentioned in the tables. It is worth to state that feature lists are differently designed for each corpus, any feature number does not correspond the same feature across corpora.

The following procedure is related to the visualising of the data. The general information is provided in the previous chapters (Section 4.5.1). It should be noted that heat maps use the same categorical information in the Jaccard distance matrix with a colour for each pair of texts. In Jaccard distance matrix, any value less than 1 means those two authors are getting similar to each other, and the boxes are represented with red colour, the other way around the dissimilar authors are tinted with the blue colour. The <1 distance values are featured on a visualisation method which leads to figure out general tendencies in a real-world forensic case.

In the following studies, it is assumed that the disputed texts are written by one of the fourteen authors. Distance matrix results excel outputs are partially added in the written part of the book due to the size restrictions. However, full distance matrix results can be found in Appendix D: Jaccard Outputs

6.2. The Role of Feature Types in Attributing Authorship

In this section, the importance of feature types in authorship studies when analysing forensic text was investigated. Three groups of the feature set were implemented to determine the most efficient features in analysing authorship in Turkish text. In the following sections, the results of the three feature groups are presented; they illustrate discriminative potential and increase the probability of assigning actual authorship. Sketch Engine was used to obtain all possible linguistic features. Firstly, a top-down approach was applied to determine the pre-defined features and secondly; a bottom-up approach was taken to code the data-driven features. At the

end of this process, a total of 38 lexical features were derived for Corpus1, 28 from Corpus2, 19 from Corpus3 and 13 from Corpus4, respectively. Once the features were selected and coded within the text, the next step was to examine the effectiveness of the features in three settings. However, the role of the feature types test was only applied to Corpus1, in response to the related research questions. The feature selection and analysing methods were presented in Chapters 4 and 5, respectively. Overall, as mentioned above the disputed author A1 and the actual author A15 are designed as the corresponding authors of this scenario in the following three sub-tests. (Section 6.2.1, 6.2.2 and 6.2.3.)

6.2.1. Lexical Features

In this test, 38 lexical features were tested across 225 text files from 14 different authors. A total of sixteen two-word grams in the 38 lexical features were obtained from the word list. The entire list of lexical features can be seen in Section 5.2.1.; however, the word n-grams, initialism and abbreviation features are not presented separately instead of with their general names. According to the statistical procedure presented above first, the smallest mean Jaccard distances are presented, and it is followed with the shared features and finally the author vs author comparison. The smallest mean distances between authors for lexical features are shown in Table 6-1. When the smallest values were investigated after calculating the means A1- the disputed author showed similarities with two authors which are A9 and A15. As it is mentioned above, when at least 5 of 15 values have the smallest values between another author, it is selected as a potential author. The overall performance between texts is reported mostly above 0.80 in the distance matrix. However, the best-averaged accuracy performance remains lower than maximum 0.80 mean distance score in the entire dataset. Even though the lowest score is hit 0.65 between A1 and A15, this does not highlight the possibility of being the actual author of the texts.

Table 6-1: Smallest mean distances between texts for lexical features.

AUTHORX	AUTHOR15	DISTANCE VALUES	AUTHORX	AUTHOR9	DISTANCE VALUES
A1T1	A15T1	0.719185856	**A1T1**	A9T1	0.795571912
A1T2	A15T2	0.693171641	**A1T2**	A9T2	0.793338402
A1T3	A15T3	0.667958561	**A1T3**	A9T3	0.814443277
A1T4	A15T4	0.673263415	**A1T4**	A9T4	0.821838671
A1T5	A15T5	0.716895993	**A1T5**	A9T5	0.711762368
A1T6	A15T6	0.72417843	**A1T6**	A9T6	0.744968366
A1T7	A15T7	0.719621759	**A1T7**	A9T7	0.682514521
A1T8	A15T8	0.717708018	**A1T8**	A9T8	0.67355131
A1T9	A15T9	0.695428845	**A1T9**	A9T9	0.674737168

A1T10	A15T10	0.657813003	**A1T10**	A9T10	0.663197768
A1T11	A15T11	0.722969367	**A1T11**	A9T11	0.732146837
A1T12	A15T12	0.767603087	**A1T12**	A9T12	0.782583266
A1T13	A15T13	0.69116388	**A1T13**	A9T13	0.765478435
A1T14	A15T14	0.732074122	**A1T14**	A9T14	0.686688691
A1T15	A15T15	0.82141784	**A1T15**	A9T15	0.689392167

Furthermore, with an inadequate number of features, classifying the authors incorrectly as belonging to an author than other than the actual author is expected, however, in this case, there are two candidate authors for the disputed author. Therefore, for further investigation, shared features between both potential authors and the disputed author is observed and presented in Table 6-2 to show the distribution of shared features. Because of the low occurrences of some features, if the value is 10 or more than 10 instances, it is excluded from the matrix for further analysis. Between 38 lexical features, only 10 features which are 26 % were highly shared between the selected authors. Moreover, rest of the 28 lexical features were used at least twice by A1 and A15, on the other side, A9 did not use 9 features between 28 features which are not presented due to the frequency threshold in Table 6-2.

Table 6-2: Shared features between A1 and A9&A15

FEATURES	**AUTHOR1**	**AUTHOR9**	**AUTHOR15**
F5 Lex - bi sey	10	4	11
F18 Lex- bi	15	14	15
F21 Lex- falan	8	6	10
F22 Lex- Int. akp	0	0	0
F23 Lex- Int. tl	0	1	0
F24 Lex- Interjection ay	0	0	0
F30 Lex- lan	14	11	13
F34 Lex- Omission R Clip	0	0	0
F37 Lex- Reduplication/ Same Form	12	7	3
F38 Lex- tabiki	0	0	0

However, the high occurrences and absent shared features between the first and the second-best averaged authors A9 and A15 are included in this table. Among the selected features which were used at least 10 times, F18 and F30 for Author9 are rendered the approximate results with A1 and A15. However, F5 is not used even a single time by A9 while A1 and A15 have the highest scores in this feature. That is to say that, A1 and A15 used F5 in 73% percent of their texts. It can also be seen in Table 6-2 is the unused features which show a high degree of

agreement among three authors. A9 uses only F23 while A15 and A1 are not used that feature particularly.

Furthermore, the 2-word gram F5 *bi sey* is used 25 times within 225 texts from 14 different authors; however, A1 and A15 used this feature 21 times alone. This use of the feature is found in some texts that belong to A9. Also, this can explain the situation lowest mean distances in the corpus.

A similar situation occurred in F18 *bi;* this feature is used 78 times by different authors in the entire corpus. However, the use of F18 is mostly shared between A1, A9 and A15 which is equal to 58% of this feature in the corpus.

The first two steps of interpretation in the context of the results reported. However, it is still not evident to attribute the authorship correctly. The second approach of the analysis of authorship attribution is to concatenate the individual texts and treat them as a single chunk. Since the disputed author is the A1, the potential author should be aligned with the A1. The distance between each author using the raw Jaccard values based on the presence and absence of each feature disregarding the occurrences is presented in Table 6-3 and showed that the smallest distance values belong to A15 when it is compared to all authors.

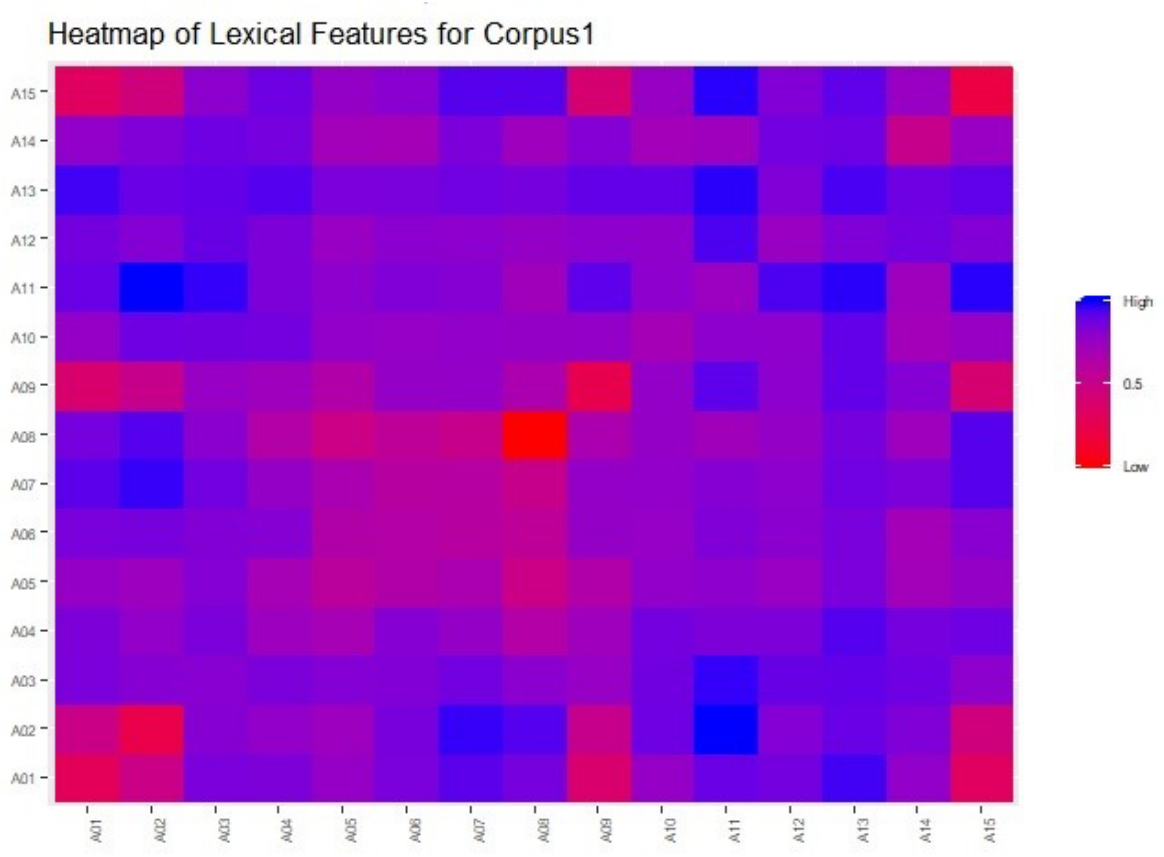
Table 6-3: Distances between authors for lexical features.

	row.names	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15
1	A01	0.7027688	0.7743854	0.8996382	0.8973812	0.8691884	0.9006482	0.9245676	0.9046488	0.7351260	0.8695955	0.9150907	0.9062028	0.9377297	0.8747764	0.7123426
2	A02	0.7743854	0.6829782	0.8877193	0.8743225	0.8589986	0.9029584	0.9424243	0.9286845	0.7810409	0.9107441	0.9513580	0.8905255	0.9146925	0.8944879	0.7587747
3	A03	0.8996382	0.8877193	0.8867007	0.8989575	0.8899496	0.8927407	0.9079436	0.8825254	0.8668905	0.9090664	0.9428765	0.9163401	0.9192734	0.9116660	0.8788932
4	A04	0.8973812	0.8743225	0.8989575	0.8584948	0.8424829	0.8879521	0.8704021	0.8222184	0.8561078	0.9066638	0.8965891	0.8966125	0.9287690	0.9034567	0.9108075
5	A05	0.8691884	0.8589986	0.8899496	0.8424829	0.8069659	0.8246618	0.8378801	0.7726849	0.8263724	0.8749476	0.8798924	0.8645107	0.8993383	0.8510258	0.8711951
6	A06	0.9006482	0.9029584	0.8927407	0.8879521	0.8246618	0.8230574	0.8144444	0.8001446	0.8703269	0.8685079	0.8935079	0.8823228	0.9009206	0.8470006	0.8858532
7	A07	0.9245676	0.9424243	0.9079436	0.8704021	0.8378801	0.8144444	0.8147619	0.7812663	0.8717776	0.8744039	0.8881376	0.8791623	0.9092063	0.8980104	0.9273249
8	A08	0.9046488	0.9286845	0.8825254	0.8222184	0.7726849	0.8001446	0.7812663	0.6024301	0.8367110	0.8702389	0.8535097	0.8699416	0.9035520	0.8559967	0.9273538
9	A09	0.7351260	0.7810409	0.8668905	0.8561078	0.8263724	0.8703269	0.8717776	0.8367110	0.6876913	0.8723108	0.9231717	0.8798154	0.9200768	0.8911930	0.7406739
10	A10	0.8695955	0.9107441	0.9090664	0.9066638	0.8749476	0.8685079	0.8744039	0.8702389	0.8723108	0.8460084	0.8801164	0.8785719	0.9192769	0.8493460	0.8661804
11	A11	0.9150907	0.9513580	0.9428765	0.8965891	0.8798924	0.8935079	0.8881376	0.8535097	0.9231717	0.8801164	0.8611791	0.9321058	0.9453333	0.8572540	0.9458116
12	A12	0.9062028	0.8905255	0.9163401	0.8966125	0.8645107	0.8823228	0.8791623	0.8699416	0.8798154	0.8785719	0.9321058	0.8635711	0.8941640	0.9073288	0.8929573
13	A13	0.9377297	0.9146925	0.9192734	0.9287690	0.8993383	0.9009206	0.9092063	0.9035520	0.9200768	0.9192769	0.9453333	0.8941640	0.9344444	0.9113845	0.9212849
14	A14	0.8747764	0.8944879	0.9116660	0.9034567	0.8510258	0.8470006	0.8980104	0.8559967	0.8911930	0.8493460	0.8572540	0.9073288	0.9113845	0.7808125	0.8655111
15	A15	0.7123426	0.7587747	0.8788932	0.9108075	0.8711951	0.8858532	0.9273249	0.9273538	0.7406739	0.8661804	0.9458116	0.8929573	0.9212849	0.8655111	0.6715114

In Table 6.3. there is a slightly different pattern when it is compared to the provided information in the two previous tables. As it is mentioned above, this table shows the results of the

concatenated texts using all lexical texts against all authors. Between 225 values the lowest raw Jaccard distance value is hit by A15 which is 0.71 when it is crossed with another other apart from itself, and the result is plotted with a red line around. This value is showing the distance between A1 and A15. However, when the results from A9 is investigated the lowest score is found 0.73 when the texts are compared with the disputed texts from A1. In such a case, 0.71 has the smallest distance rather than 0.73. The same tendency between A9 and A1 is not observed in this table; instead of A15 is provided evidence to show a better performance in lexical features. Figure 6-1 heatmap has also provided the match between disputed and potential authors to improve the accuracy which was previously obtained with the results.

Figure 6-1: Heatmap of Lexical Features for Corpus1.



It is possible to observe that there are three distinct patterns which illustrated as inter-author and intra-author distances in this heat map. On one side, the similarities between intra-authors are showed a tendency when they are compared with each other. However, intra-author similarities are not regarded as an issue in this study. On the other side, the lowest raw Jaccard values are presented from the distance matrix. After ignoring the intra-author similarities, there are only two most visible traces of some authors in the leftmost of the figure supporting the

values presented above. A1, A9 and A15 have consistently smaller distances when compared to the other authors, however, putting together the tables produced and shared features between the authors A15 produced the smallest distance when it is compared to the disputed author. Overall, based on these results, the disputed author's style is consistent with the A15 which the corresponding author is selected at the beginning of the tests, while both authors are distinctive from the rest of the authors and the method adopted here leads to classification in authorship in other words it is a confident attribution as Love (2002) stated.

6.2.2. Syntactic Features

In the second test, syntactic features were assessed with authorship attribution in Turkish text. 40 syntactic features were derived from the data set including punctuation features. The entire list of syntactic features can be seen in Section 5.2.2. After the distance test was performed, the texts more similar to each other were found from the entire list. Table 6-3 reveals that the smallest mean distances between texts for syntactic features between the disputed author and the A15.

Table 6-3: Smallest mean distances between texts for syntactic features.

AUTHORX	AUTHOR15	DISTANCE VALUES
A1T1	A15T1	0.78365079
A1T2	A15T2	0.74226884
A1T3	A15T3	0.80335979
A1T4	A15T4	0.81639795
A1T5	A15T5	0.85180061
A1T6	A15T6	0.82677249
A1T7	A15T7	0.76851611
A1T8	A15T8	0.82988937
A1T9	A15T9	0.83364043
A1T10	A15T10	0.82713324
A1T11	A15T11	0.89661228
A1T12	A15T12	0.83780423
A1T13	A15T13	0.95481481
A1T14	A15T14	0.80386003
A1T15	A15T15	0.82596922

The overall performance between texts is generally over 0.80 in this dataset. It shows that some of these mean distances in Table 6-3 are high; only three values are lower than 0.80 which are indicated by lines. In this case, A15 is not qualified as a candidate author since it has only three smallest values when it is compared with the rest. For that reason, a second graph is produced

to indicate the other lower values between A1 and the selected authors who have a degree of similarity in their texts in Table 6-4.

Table 6-4: Smallest mean distances between all texts for syntactic features.

AUTHORS	DISTANCE VALUES
A2T10	0.791056721
A3T1	0.79042328
A9T4	0.796296296
A10T4	0.796236171
A10T9	0.798809524
A10T10	0.793434343
A10T12	0.799259259
A14T3	0.791741222
A14T4	0.760916306
A14T7	0.799227994
A14T11	0.762063492
A15T1	0.783650794
A15T10	0.742268842
A15T15	0.768516114

As Table 6-4 presents that the mean score is close to 0 in six of the authors; however, A2, A3 and A9 have a little linguistic link between the A1 since they only have one value. Three other authors including A10, A14 and A15 have a mean distance lower than 0.8 and among these A14 and A10 are more striking as they have smaller values than A15. However, when all of the distance matrices investigated the lowest value is found 0.74 between A1 and A15. Although it can be possible to think that A15 is responsible for all the texts by A1, there is no strong link between the authors.

From this analysis in Table 6-4, it is evident that the there are three candidate authors, and there is no strong link between A1 and A15 although they have one of the lowest mean distance among all texts. Thus, shared linguistic features are presented in order to find the linguistic link between A1 and A15 in Table 6.5. It should be noted that because of the low occurrences of some features if the value is 10 or more than 10 instances, it is not stated in this table. That is to say; there are some other features between 40 syntactic features which occurred less than 10 and not presented here. However, they affect the overall performance which can be seen in author vs author comparison.

Table 6-5: Shared linguistic features between A1 and A15.

FEATURES	AUTHOR1	AUTHOR15
F45 Syn- ConConj ama	0	0
F46 Syn- ConConj ancak	0	0
F51 Syn- ConConj mesela	0	0
F54 Syn- EndConj ancak	0	0
F56 Syn- EndConj hatta	0	0
F59 Syn- MixedTypo. Punctuation	0	0
F60 Syn- MultiTypo. Punc. Exclamation	0	0
F61 Syn-MultiTypo.Punc. Question Mark	0	0
F62 Syn- NonStdPunc. Question Mark	0	0
F66 Syn- Punc.Empahis Exclamation	0	0
F68 Syn-Punc. Absence Apostrophe	0	0
F69 Syn-Punc.Absence Fullstop	0	0
F70 Syn- Sentence quote	9	13
F73 Syn- StrConj ama	0	0
F74 Syn- StrConj ancak	0	0

As it is seen in Table 6-5, only F70- Sentence quotation used by both authors while the rest did not use at all. This feature is a quotation which is set off from the rest of the sentence by a comma and used 122 times among 225 texts from 14 distinctive authors. Thus, the presence of F70 does not support the hypothesis that these two authors are the same person. Even though there are 14 absent features which are 35% of entire syntactic features and are not-used by both authors, it is not enough to decide the author of these texts are the same person. The results of the syntactic features were challenging when compared to the lexical features. Although 40 syntactic features, including punctuation marks and syntactically structured conjunctions, were found, their performance was not as good as expected.

One reason for this observation may be because this study did not focus on POS tagging but rather individual syntactical features, such as conjunctions and punctuation marks, thus all features in a sentence were not analysed. However, using elements of speech annotation for online text is still undeveloped for the Turkish language. Moreover, it is impossible to determine the superiority of distinctive linguistic features considering there are generally no

accepted suitable features on authorship attribution (Rudman, 1998), mainly because of the varying conditions in different settings.

The second reason was to have standard punctuation marks as features. When it is considering the frequency of such feature between 14 different authors and 225 texts the discriminative power of the feature decreases.

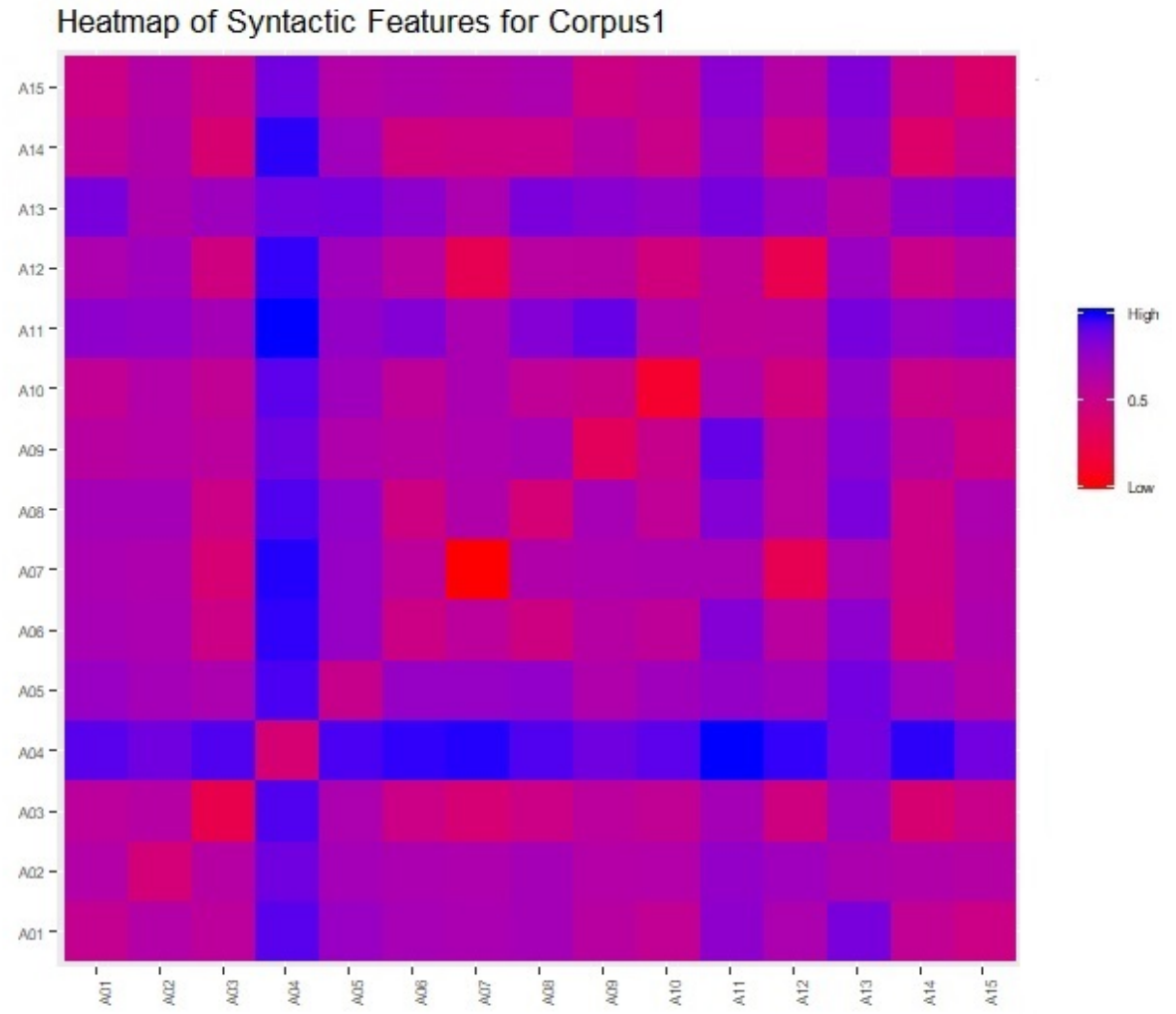
This analysis also points out the comparison between authors when their texts are accepted as a single piece. The previous results, however, found little linguistic evidence that supports the common authorship; it can be possible to find a similarity when the texts are combined similar to lexical features.

	row.names	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15
1	A01	0.8407703	0.8697691	0.8539648	0.9601318	0.9065903	0.8870750	0.8835645	0.8898256	0.8627874	0.8447353	0.9188239	0.8795565	0.9387718	0.8431352	0.8268327
2	A02	0.8697691	0.8070126	0.8646293	0.9448217	0.8919468	0.8813997	0.8783789	0.8902274	0.8684124	0.8707789	0.9130725	0.8982399	0.8811940	0.8719708	0.8668261
3	A03	0.8539648	0.8646293	0.7493500	0.9640829	0.8801744	0.8262326	0.8017789	0.8261445	0.8564828	0.8454528	0.8899591	0.8176931	0.9004674	0.7984797	0.8307308
4	A04	0.9601318	0.9448217	0.9640829	0.8002154	0.9668519	0.9757919	0.9786014	0.9640071	0.9451004	0.9580141	0.9821711	0.9745538	0.9408360	0.9765395	0.9443505
5	A05	0.9065903	0.8919468	0.8801744	0.9668519	0.8349093	0.9089295	0.9092536	0.9152520	0.8752166	0.8976891	0.9116080	0.8978764	0.9436808	0.8975499	0.8706051
6	A06	0.8870750	0.8813997	0.8262326	0.9757919	0.9089295	0.8231746	0.8538937	0.8209875	0.8657298	0.8504334	0.9290436	0.8593827	0.9208448	0.8175926	0.8784491
7	A07	0.8835645	0.8783789	0.8017789	0.9786014	0.9092536	0.8538937	0.6777273	0.8730627	0.8788679	0.8839702	0.8829933	0.7570984	0.8801725	0.8244501	0.8717012
8	A08	0.8898256	0.8902274	0.8261445	0.9640071	0.9152520	0.8209875	0.8730627	0.8042215	0.8875007	0.8482961	0.9291194	0.8604214	0.9368113	0.8264260	0.8798043
9	A09	0.8627874	0.8684124	0.8564828	0.9451004	0.8752166	0.8657298	0.8788679	0.8875007	0.7677362	0.8332976	0.9525851	0.8627177	0.9243704	0.8658340	0.8218691
10	A10	0.8447353	0.8707789	0.8454528	0.9580141	0.8976891	0.8504334	0.8839702	0.8482961	0.8332976	0.7135110	0.8697255	0.8141552	0.9113739	0.8294105	0.8405845
11	A11	0.9188239	0.9130725	0.8899591	0.9821711	0.9116080	0.9290436	0.8829933	0.9291194	0.9525851	0.8697255	0.8509335	0.8524004	0.9396614	0.9092374	0.9229987
12	A12	0.8795565	0.8982399	0.8176931	0.9745538	0.8978764	0.8593827	0.7570984	0.8604214	0.8627177	0.8141552	0.8524004	0.7509599	0.9047372	0.8311757	0.8666101
13	A13	0.9387718	0.8811940	0.9004674	0.9408360	0.9436808	0.9208448	0.8801725	0.9368113	0.9243704	0.9113739	0.9396614	0.9047372	0.8664853	0.9173704	0.9320649
14	A14	0.8431352	0.8719708	0.7984797	0.9765395	0.8975499	0.8175926	0.8244501	0.8264260	0.8658340	0.8294105	0.9092374	0.8311757	0.9173704	0.7858644	0.8385737
15	A15	0.8268327	0.8668261	0.8307308	0.9443505	0.8706051	0.8784491	0.8717012	0.8798043	0.8218691	0.8405845	0.9229987	0.8666101	0.9320649	0.8385737	0.7877196

Table 6-6: Distances between authors for syntactic features.

In Table 6-6, there are many values which are above 0.8 while only a few have the closest values to 0. The lowest raw Jaccard score hit by A12 and A7 that is relevant neither to the disputed author, not the corresponding author. Although there are some lowest intra-author values, this is ignored since it does not contribute to the aim of the book. It is evident that the disputed author is not successfully managed to indicate a consistent linguistic distinctiveness between the rest of the authors. Additional evidence of misclassification is presented in Figure 6-2.

Figure 6-2: Heatmap of syntactic features for Corpus 1.



The heat map in Figure 6-4 supports the general results based on the syntactic features. There are some patterns apparent in this figure; however, they mostly indicate the same author range. Furthermore, there are some other unexplained patches distributed across the data. The linguistic distinctiveness is unlikely to shows that there is an authorial link between the disputed author and the other authors.

In summary, the results in this study are not consistent with previous studies that claim syntactic features are an ideal discriminative feature (e.g. Chaski, 2002). It is important to note that syntactic features in this study are different from other studies. For instance, this study did not focus on POS tagging or function words but rather the conjunctions in a sentence, their

position and the standard and non-standard usage of punctuation marks. Moreover, the text files used to represent the author's syntactic tendencies were significantly short compared to other studies. This test was not an appropriate method to determine consistency and distinctiveness with limited syntactic features selected for this study and to attribute the correct author.

6.2.3. Structural Features

In the final test of this study, structural features were used to establish their role in authorship attribution cases. However, only five structural features were identified including format-related features, for instance, *bknz* (see also), *hyperlink, itemise* (organising the text with bullet points), *spoiler* and *swh*. For example, *bknz* is considered a potential discriminative feature; however, it occurs infrequently within the data set, unlike the frequently occurring lexical and syntactic features.

Structural features reveal more clues about the author depending on the organisation of the text, for instance, emails in other genres (e.g. de Vel, 2000). Furthermore, Wright (2014) used greetings and farewells in identifying individuals' idiolects in 126 Enron authors' email corpora. Similar to greetings and signatures in emails, an ideal Eksi Sozluk entry should include the *'-dir/-tir'* suffix in describing the terms that refer to the 'it is' structure in English. However, such an approach requires a morphologically syntactic tagging, and this was not the focus of this study.

In light of this information, structural features are considered ineffective in this particular genre; however, they are still included in the general reference list and the following tests to improve the overall accuracy.

As it is not possible to run the analysis with only structural features included, their performance is unknown in the current data set.

6.3. Section Conclusion

In this section, how feature types affect the accuracy of authorship was investigated. When comparing the results of the three feature sets, it was clear that lexical features have more success than syntactic and structural features. The results of this test have shown that lexical features have a positive effect on performance, while syntactic and structural features were unable to attribute authorship to the Turkish text in the current data set. Despite this, syntactic

features include some valid markers that have been tested in previous studies, such as punctuation that was not found to correctly attribute the disputed author (e.g. Chaski, 2000). Due to the limited number of structural features used in this study, no test was used for this set. However, the results may have varied if different text types with more structural features were used, such as emails. In conclusion, lexical features showed the most consistent performance in the data set; however, a better result may have been evident for structural features if a different genre was chosen for analysis.

Under these conditions, it is impossible to claim superiority of any features used in authorship attribution in Turkish text. Lexical features tend to function better than syntactic and structural features. Rather than using individual feature sets, combining all features in authorship attribution could not only lead to higher performance, but the reliability of the test may improve. However, this section does not explore whether the combination of linguistic levels shows a discriminative high score or not; this is assessed in the next chapter.

6.4. The Role of Size in Attributing Authorship

In authorship problems, there are many challenges regarding the text size. As Coulthard and Johnson (2007, p.9) pointed out 'they [forensic text types] are often limited in size and availability' on the other hand, the notion of 'more data, better data' (Moore, 2001) is accepted is in computational authorship attribution methods since the success rate is quite low. Although it is hard to obtain meaningful results from small texts since longer texts include more linguistic features which reveal the structure of a text (McMenamin, 2002), in real-world forensic linguistics cases the short text size is the new challenge in authorship studies.

In order to examine the effect of different combinations of various situations such as text size, a limited number of texts per author and the number of candidate author size are performed authorship attribution in Turkish. For the first task the text size Corpus1 – long texts, Corpus2 – medium texts, Corpus3 – short texts are trained, after this, the number of candidate authors is increased to 30 from 15 authors. For the final test in this section, the number of texts per author is limited 5 to 10 respectively. The same feature coding and data analysis methods are conducted for all tasks. Contrary to the previous tests in Section 6.2. all features are combined to contribute the solution effectively.

6.4.1. Text Size

As it is presented above three corpora is compiled with the aim of establishing the role of text sizes in Turkish texts. First Corpus1 – Long Texts is and balanced in terms of the number of texts per author, and the average size of texts is 347 words per text. Second, Corpus 2 is balanced in terms of text numbers – 15 texts per author and text size with an average of 119 words per text. Finally, Corpus 3 is balanced in terms of text numbers – 15 texts per author and with an average of 43 words per text. This set up allows straightforward comparison of the data sets with different texts sizes.

6.4.1.1. Corpus1 - Long Texts

In this test, a combined set of 83 linguistic features are tested across 225 texts from 14 different authors. The evaluation procedure presented above is applied in order to observe the closest values to 0. The disputed author and the corresponding author are determined as A1 and A15 in this corpus. The smallest mean Jaccard distances which hit the lowest score at least five times is selected as a potential author as a first step of the analysis. Thus, Table 6-7 below indicates good results with an analysis of the combined feature set that includes 83 linguistic features.

Table 6-7: Smallest mean distances between texts in Corpus1.

AUTHORX	AUTHOR15	DISTANCE VALUES
A1T1	A15T1	0.753788717
A1T2	A15T2	0.758477224
A1T3	A15T3	0.754784411
A1T4	A15T4	0.739969484
A1T5	A15T5	0.813718057
A1T6	A15T6	0.801976012
A1T7	A15T7	0.777241031
A1T8	A15T8	0.767973078
A1T9	A15T9	0.847718939
A1T10	A15T10	0.733126537
A1T11	A15T11	0.721330434
A1T12	A15T12	0.733744085
A1T13	A15T13	0.784011209
A1T14	A15T14	0.752790557
A1T15	A15T15	0.75267991

In Table 6-7 the smallest mean distances between A1 and A15 are presented that are more similar to each other between entire dataset although the general results are above 0.80 in the entire distance matrix. The amount of similarity is striking considering the indefinite matching in syntactic features.

A closer examination that shows the shared features between the two authors is presented in Table 6-8 — 21 out of 83 features shared highly which is about 25% of the features in the corpus. Because of the low instances of some features, when the frequency of a feature is less than 10 is not demonstrated in the table. As it is mentioned above (in Section 6.2.1) there are some distinctive features which are used mostly by the disputed and the candidate author.

Table 6-8: Shared linguistic features between A1 and A15.

FEATURES	**AUTHOR1**	**AUTHOR15**
F5 Lex - bi sey	10	11
F18 Lex- bi	15	15
F21 Lex- falan	8	10
F22 Lex- Int. akp	0	0
F23 Lex- Int. tl	0	0
F24 Lex- Interjection ay	0	0
F30 Lex- lan	14	13
F34 Lex- Omission R Clip	0	0
F37 Lex- Reduplication/ Same Form	12	3
F38 Lex- tabiki	0	0
F45 Syn- ConConj ama	0	0
F46 Syn- ConConj ancak	0	0
F51 Syn- ConConj mesela	0	0
F54 Syn- EndConj ancak	0	0
F56 Syn- EndConj hatta	0	0
F59 Syn- MixedTypo. Punctuation	0	0
F60 Syn- MultiTypo. Punc. Exclamation	0	0
F61 Syn-MultiTypo.Punc. Question Mark	0	0
F62 Syn- NonStdPunc. Question Mark	0	0
F66 Syn- Punc.Empahis Exclamation	0	0
F68 Syn-Punc. Absence Apostrophe	0	0
F69 Syn-Punc.Absence Fullstop	0	0
F70 Syn- Sentence quote	9	13
F73 Syn- StrConj ama	0	0
F74 Syn- StrConj ancak	0	0

The use of F5 and F18 have higher frequency among the other authors while F21 is used 61 times in different texts across the authors. Even though it is used 8 times by the disputed author, it is not enough to indicate a high level of distinctiveness in this feature. Among the shared features, F30 *lan* is one of the most substantial feature that shows the consistency between A1 and A15. This feature is used 59 times in general, and the disputed and the corresponding authors use 27 of them that is more than 45%. The syntactic feature F70 is already discussed above, and the use of a sentence quotation does appear many times in the other texts from the other authors also.

Furthermore, the analysis of absent features is also necessary to show the distinctive linguistic similarities. For instance, syntactic feature F45 the use of *ama* as connecting conjunction is used 19 times in general while it is not used even a single time by A1 and A15. The F69 feature is related to the absence of a full stop at the end of the sentence and is used 12 times while it does not appear in A1 and A15 texts. The absence of F74 - *StrConj ancak* and it is used 15 times in different texts also supports the hypothesis that A1 and A15 are consistent in their writings. After the analysis that points out the linguistic link between A1 and A15, the author vs author is done with concatenating all single texts by the authors in Table 6-9.

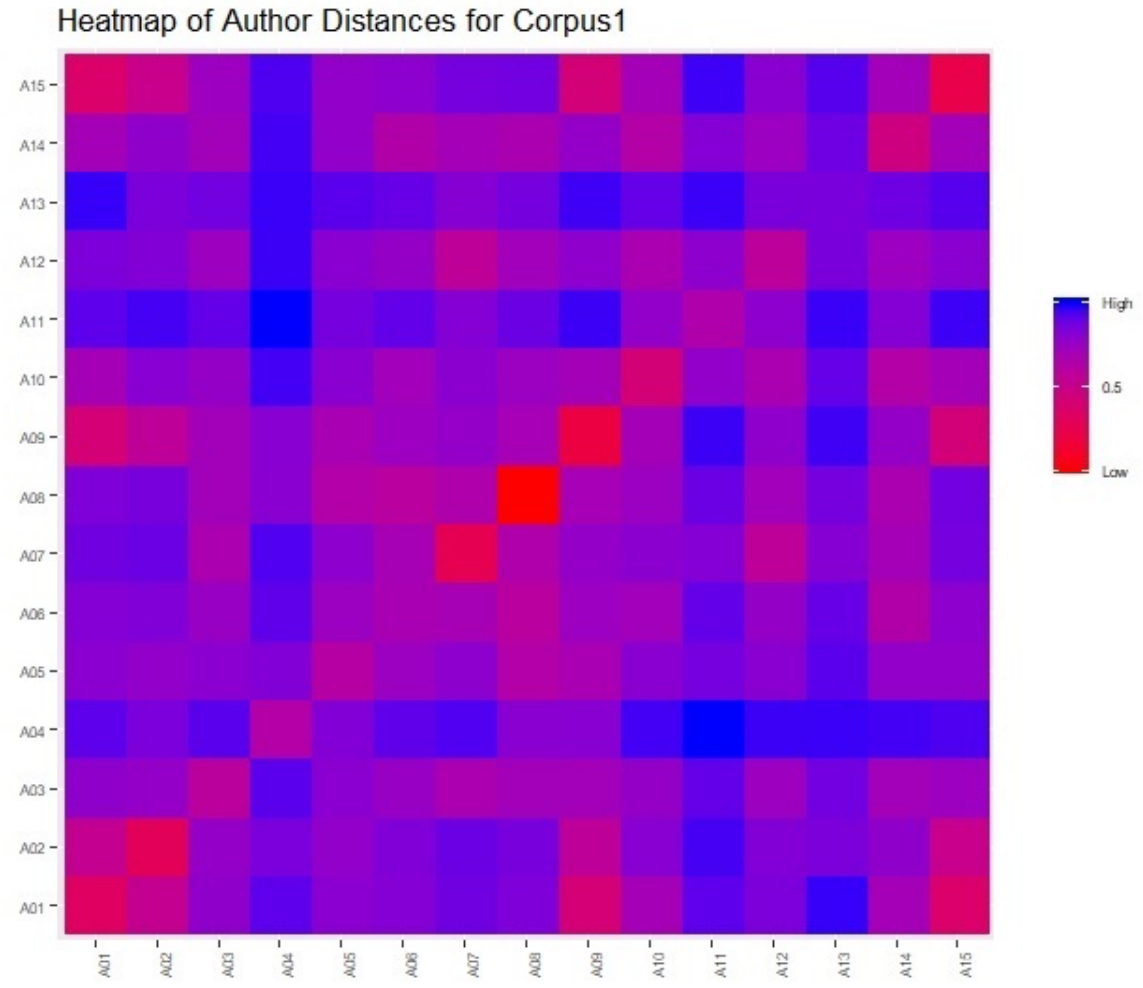
Table 6-9: Similarities between authors in Corpus 1.

	row.names	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15
1	A01	0.7557600	0.8164073	0.8896829	0.9281658	0.8948848	0.9004586	0.9163418	0.9058759	0.7808596	0.8641297	0.9281819	0.9076195	0.9440388	0.8646365	0.7662220
2	A02	0.8164073	0.7451764	0.8854689	0.9088964	0.8869319	0.9044425	0.9199905	0.9117007	0.8263958	0.8972614	0.9390968	0.9031876	0.9078211	0.8897718	0.8086520
3	A03	0.8896829	0.8854689	0.8299983	0.9303545	0.8951335	0.8811491	0.8557944	0.8680068	0.8681427	0.8848326	0.9254405	0.8743946	0.9155516	0.8678997	0.8756391
4	A04	0.9281658	0.9088964	0.9303545	0.8433799	0.9029697	0.9270614	0.9348345	0.8966151	0.8973729	0.9404232	0.9518968	0.9425783	0.9429711	0.9398706	0.9353845
5	A05	0.8948848	0.8869319	0.8951335	0.9029697	0.8390983	0.8767300	0.8921795	0.8456377	0.8600972	0.8953296	0.9122074	0.8960715	0.9297665	0.8877226	0.8879378
6	A06	0.9004586	0.9044425	0.8811491	0.9270614	0.8767300	0.8597781	0.8609563	0.8339913	0.8752612	0.8685956	0.9252014	0.8843950	0.9233466	0.8458629	0.8923652
7	A07	0.9163418	0.9199905	0.8557944	0.9348345	0.8921795	0.8609563	0.7354508	0.8497704	0.8858909	0.8947506	0.9013755	0.8273606	0.8997852	0.8657573	0.9120492
8	A08	0.9058759	0.9117007	0.8680068	0.8966151	0.8456377	0.8339913	0.8497704	0.6603880	0.8629381	0.8764735	0.9196026	0.8685995	0.9123161	0.8562556	0.9159356
9	A09	0.7808596	0.8263958	0.8681427	0.8973729	0.8600972	0.8752612	0.8858909	0.8629381	0.7167267	0.8662547	0.9423231	0.8902126	0.9412757	0.8853717	0.7832457
10	A10	0.8641297	0.8972614	0.8848326	0.9404232	0.8953296	0.8685956	0.8947506	0.8764735	0.8662547	0.7826936	0.8875874	0.8585815	0.9244089	0.8449590	0.8656476
11	A11	0.9281819	0.9390968	0.9254405	0.9518968	0.9122074	0.9252014	0.9013755	0.9196026	0.9423231	0.8875874	0.8492640	0.8928202	0.9430965	0.9001807	0.9419219
12	A12	0.9076195	0.9031876	0.8743946	0.9425783	0.8960715	0.8843950	0.8273606	0.8685995	0.8902126	0.8585815	0.8928202	0.8286274	0.9093638	0.8754698	0.8965042
13	A13	0.9440388	0.9078211	0.9155516	0.9429711	0.9297665	0.9233466	0.8997852	0.9123161	0.9412757	0.9244089	0.9430965	0.9093638	0.9106283	0.9185863	0.9316934
14	A14	0.8646365	0.8897718	0.8678997	0.9398706	0.8877226	0.8458629	0.8657573	0.8562556	0.8853717	0.8449590	0.9001807	0.8754698	0.9185863	0.7942886	0.8667181
15	A15	0.7662220	0.8086520	0.8756391	0.9353845	0.8879378	0.8923652	0.9120492	0.9159356	0.7832457	0.8656476	0.9419219	0.8965042	0.9316934	0.8667181	0.7275188

As seen in Table 6-9 there is only one distinctive raw Jaccard score which shows the smallest distance to A1 apart from the intra-author similarities when it is compared with the rest. Despite the presence of a few high numbers of shared features, the combining texts as a single chunk and the raw Jaccard scores in this showed quite an efficiency in filtering the rest of the authors and focusing the disputed and the most probable candidate author. Between 225 raw Jaccard

values there is no other value is below 0.76 in inter author comparison; thus 83 linguistic features are discriminative between authors and it can be taken as a departure point to show the linguistic link. Additionally, the pattern for the raw Jaccard distance values is visualised in Figure 6-3.

Figure 6-3: HeatMap of author distances for Corpus1.



As it is expected there are red patterns which show intra author similarities in the diagonal line. The present figure visualisation supports the shared authorship between A1 and A15 on the comparison of their texts. The first box on the bottom refers to the A1, and the last one is A15, it is found that red boxes in these places demonstrate the smallest distances between two authors. In the light of the other linguistic evidence presented above, a heat map like this provides evidence and significantly contributes the ease to present it in the courtroom.

6.4.1.2. Corpus2 - Medium Size Texts

The second corpus with an average of 119 words in length is investigated in order to decide how successful the approach that applied to various conditions in the study. A combination of 85 features including lexical, syntactic and structural features is applied in order to assess the performance in authorship attribution scenario. It should be noted that the feature sets are different for each corpus for that reason F1 in Corpus1 do not refer the same F1 in Corpus2. The disputed author and the actual author are decided as A12 and A13 in this training test. After examining the lowest mean Jaccard scores in the entire distance matrix, the author who had the lowest scores at least five times were selected as a potential author. Table 6-10 below is presented the scores between A12- disputed author and the potential actual author A13.

Table 6-10: Smallest mean distances between texts in Corpus2.

AUTHORX	AUTHOR13	DISTANCE VALUES
A12T1	A13T1	0.800302024
A12T2	A13T2	0.78230361
A12T3	A13T3	0.805711992
A12T4	A13T4	0.759309895
A12T5	A13T5	0.747951742
A12T6	A13T6	0.76215565
A12T7	A13T7	0.753667721
A12T8	A13T8	0.766560847
A12T9	A13T9	0.787962695
A12T10	A13T10	0.833244902
A12T11	A13T11	0.913739118
A12T12	A13T12	0.779618715
A12T13	A13T13	0.855104922
A12T14	A13T14	0.775679032
A12T15	A13T15	0.845058769

The general results are above 0.80 in the Corpus2 distance matrix while some are closer to 0.90. Accordingly, there are some mean Jaccard values which are around 0.80 and 0.90 and underlined in the table. There is also an apparent amount of similarity considering the rest nine values which are the smallest distance means between texts. However, it is essential to observe the shared linguistic features between the disputed and the potential author.

Table 6-11: Shared linguistic features between A12 and A13.

FEATURES	AUTHOR12	AUTHOR13
F1 Lex- Abb.	0	0
F9 Lex-Intj. lan	0	0
F11 Lex-Reduplication Mdoublet	0	0
F17 Lex-bi	8	6
F19 Lex- falan	0	0
F21 Lex-Intj. e	0	0
F23 Lex-Keyboard Emoticon	13	14
F28 Lex-Reduplication SameForm	1	8
F29 Str- bkz	0	0
F32 Str- Itemise	0	0
F34 Syn- ConConj ama	7	8
F37 Syn- ConConj bile	6	4
F39 Syn- ConConj fakat	0	0
F41 Syn- ConConj hele	0	0
F43 Syn- ConConj ve	9	4
F47 Syn-Emp1stPerson	3	7
F48 Syn- EndConj ama	0	0
F50 Syn- EndConj ayrica	0	0
F53 Syn- EndConj mesela	0	0
F55 Syn- Mixed TypoPunc	0	0
F56 Syn- MultiTypoPunc Excl.	0	0
F58 Syn-NonStdPunc Fullstop	0	0
F59 Syn- NonStdPunc DoubleFullStop	6	7
F60 Syn- NonStdPunc QuestionMark	0	0
F61 Syn- NonStdPunc Semicolon	0	0
F62 Syn- NonStdSpace before Punc	0	0
F64 Syn- PuncAbs Apost.	4	5
F65 Syn- PuncAbs Fullstop	11	13
F67 Syn-Punc. Ellipsis	5	0
F68 Syn- Punc EmphasisExclm.	0	0
F69 Syn- SenQuotation	6	8
F70 Syn- StrPunc Colon	2	5
F74 Syn- StrConj ancak	0	0
F75 Syn- StrConj. Artik	0	0
F76 Syn- StrConj ayrica	0	0
F77 Syn- SynConj cunku	0	0
F78 Syn- StrConj fakat	0	0
F79 Syn- StrConj hatta	0	0
F80 Syn- StrConj hele	0	0
F81 Syn- StrConj mesela	0	0
F82 Syn- StrConj ve	0	0
F83 Syn- StrConj veya	0	0

In Table 6-11 the shared features between the disputed author and the A13 are represented, however, due to the size of Corpus2, shared features are demonstrated only when they have at least five instances among all texts. Even though small instances are ignored in the table, 42 out of 85 features that is almost 49.5% of the entire feature list is shared between the authors. One of the most striking examples is F23 – *Keyboard Emoticon* which are the textual portrayals of the authors' mood is used 27 times by both authors; moreover, this feature used 48 times in the entire 225 texts. That is to say that, more than 56% of the usage belongs to only these authors. In the same vein, F65- *Punctuation Absence Full stop* is used 38 times in general, and the disputed and the potential authors use 24 of them that is more than 63%.

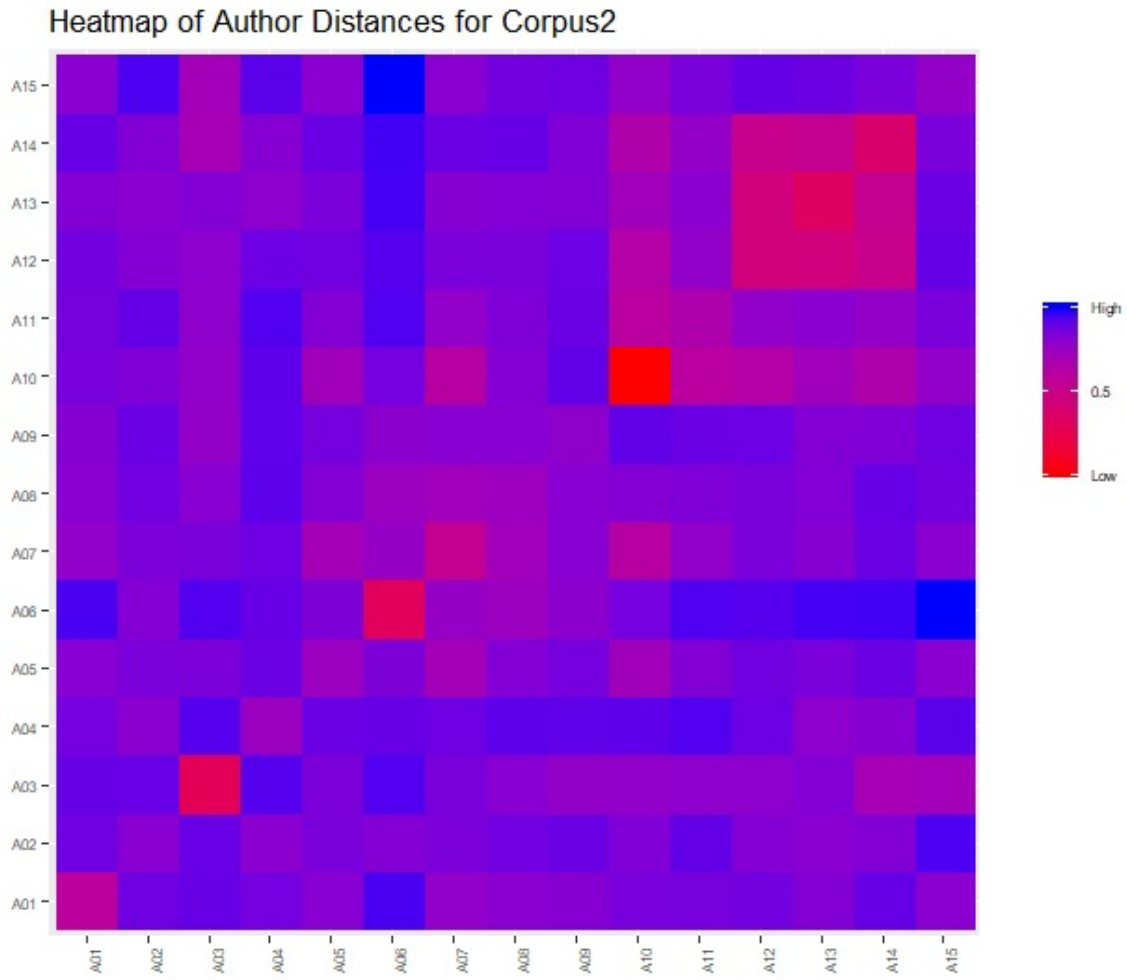
Furthermore, the absence of a feature is essential when both authors share it. For instance, F19 -*falan* appears 25 times in some of the 195 texts from the rest of the authors while A12 and A13 have not used this feature even a single time. After the points have made it can be said that there is a linguistic relation between A12 and A13 and this requires further analysis in order to establish a reliable ground in attributing the authorship. The author vs author raw Jaccard distance measures is presented in Table 6-12.

Table 6-12: Distances between authors in Corpus2.

	row.names	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15
1	A01	0.8450265	0.9379309	0.9431788	0.9329998	0.9166401	0.9592202	0.9049257	0.9153883	0.9173325	0.9303652	0.9307766	0.9334685	0.9209405	0.9437797	0.9154006
2	A02	0.9379309	0.9139254	0.9424186	0.9128496	0.9285667	0.9196197	0.9271832	0.9359494	0.9410126	0.9233792	0.9458930	0.9192585	0.9138295	0.9222621	0.9585983
3	A03	0.9431788	0.9424186	0.7459577	0.9541656	0.9265758	0.9558791	0.9291055	0.9163814	0.9036805	0.9041671	0.9080756	0.9097930	0.9213755	0.8783745	0.8825819
4	A04	0.9329998	0.9128496	0.9541656	0.8928918	0.9404592	0.9437923	0.9391650	0.9495883	0.9480104	0.9482109	0.9558428	0.9399690	0.9108997	0.9171603	0.9506496
5	A05	0.9166401	0.9285667	0.9265758	0.9404592	0.8925420	0.9263202	0.8820644	0.9212414	0.9319439	0.8855994	0.9220887	0.9369258	0.9285573	0.9406786	0.9135236
6	A06	0.9592202	0.9196197	0.9558791	0.9437923	0.9263202	0.7541332	0.9007315	0.8929356	0.9116965	0.9330173	0.9565029	0.9541906	0.9618164	0.9625696	0.9755109
7	A07	0.9049257	0.9271832	0.9291055	0.9391650	0.8820644	0.9007315	0.8287875	0.8865302	0.9162073	0.8518589	0.9040154	0.9282033	0.9174647	0.9406515	0.9137988
8	A08	0.9153883	0.9359494	0.9163814	0.9495883	0.9212414	0.8929356	0.8865302	0.8918172	0.9163987	0.9204320	0.9254187	0.9296976	0.9225598	0.9437636	0.9338767
9	A09	0.9173325	0.9410126	0.9036805	0.9480104	0.9319439	0.9116965	0.9162073	0.9163987	0.9077054	0.9471664	0.9402424	0.9400060	0.9222783	0.9227974	0.9381467
10	A10	0.9303652	0.9233792	0.9041671	0.9482109	0.8855994	0.9330173	0.8518589	0.9204320	0.9471664	0.6567683	0.8465274	0.8565376	0.8871830	0.8630335	0.9051096
11	A11	0.9307766	0.9458930	0.9080756	0.9558428	0.9220887	0.9565029	0.9040154	0.9254187	0.9402424	0.8465274	0.8667950	0.9054205	0.9131491	0.9025358	0.9298153
12	A12	0.9334685	0.9192585	0.9097930	0.9399690	0.9369258	0.9541906	0.9282033	0.9296976	0.9400060	0.8565376	0.9054205	0.7980726	0.7978914	0.8193185	0.9447603
13	A13	0.9209405	0.9138295	0.9213755	0.9108997	0.9285573	0.9618164	0.9174647	0.9225598	0.9222783	0.8871830	0.9131491	0.7978914	0.7610460	0.8263596	0.9394789
14	A14	0.9437797	0.9222621	0.8783745	0.9171603	0.9406786	0.9625696	0.9406515	0.9437636	0.9227974	0.8630335	0.9025358	0.8193185	0.8263596	0.7754160	0.9286770
15	A15	0.9154006	0.9585983	0.8825819	0.9506496	0.9135236	0.9755109	0.9137988	0.9338767	0.9381467	0.9051096	0.9298153	0.9447603	0.9394789	0.9286770	0.9032385

It is presented in Table 6-12 that there only one distinctive raw Jaccard score which hit the smallest distance between A12 and A13 while most of the authors have values above 0.90. These results are signed with a red line in the table. Even though there were some mean distance values over 0.80 in Table 6-10 which leads an uncertainty in authorship attribution, the raw scores showed promising results in selecting the actual author. Furthermore, between 225 values there are no other scores lower than 0.79 unless the author is compared with itself. It can be said that 85 linguistic features used in Corpus2 are discriminative between authors and they are efficient in showing the linguistic link between the disputed and the actual authors. Additionally, a basic visualisation method is required in presenting the results to the court. For that reason, Figure 6-4 is generated to show the author distances for Corpus2.

Figure 6-4: HeatMap of author distances for Corpus2.



In Figure 6-4 the four tendencies that can be detected easily. The most remarkable point on the right side that shows a red pattern between the authors A12, A13 and A14. It is important to note that, two of those red boxes represent the similarities between A14 and A14 thus it is related to neither with the disputed author nor the candidate author. In a similar manner with the linguistic relation presented above, it is possible to demonstrate the distinctive manner of A12 and A13 when it is compared with the rest of the authors.

Overall, even though this corpus is smaller than Corpus1, the performance is similar to that. In this case, is it possible to report that corpus size does not affect authorship attribution? It is impossible to answer this question straightforward since there are other factors to affect this. For instance, the number of feature sets is more related to the overall accuracy. In Figure 5-5 (see Chapter 5) the number of syntactic features is high in Corpus2 than Corpus1. Despite Section 6.2.2. did not provide best results it can be concluded that when we combine syntactic features with the other features, it is likely to get good performance even in the medium size texts which lead to correct classification between the disputed and the actual authors.

6.4.1.3. Corpus3 - Short Size Texts

Authorship application studies on short messages have been a debatable issue in both approaches. Traditional stylometric approaches show a significant level of decrease in accuracy when the data size decreases. However, it is also controversial in stylistic studies since it is difficult to obtain a wide range of feature sets. For instance, only 30 syntactic features were found between 225 texts while there are 51 syntactic features Corpus2.

Moreover, the average of texts is 47 words in this data set that is a relatively small text to find discriminative features. As it is mentioned above, when an author has scored the lowest mean distance values at least five times is selected as a potential author. The disputed author is A8, and the actual author is A6 in this corpus. In this case, the potential author is selected by looking at the smallest mean distance values between tests and it is presented in Table 6-13.

Table 6-13: Smallest mean distances between texts in Corpus3.

AUTHORX	AUTHOR6	DISTANCE VALUES
A8T1	A6T1	0.774280164
A8T2	A6T2	0.860711881
A8T3	A6T3	0.758778259
A8T4	A6T4	0.754247419
A8T5	A6T5	0.785586266
A8T6	A6T6	0.753677434
A8T7	A6T7	0.80773754

A8T8	A6T8	0.694986495
A8T9	A6T9	0.725732786
A8T10	A6T10	0.78989122
A8T11	A6T11	0.719317534
A8T12	A6T12	0.774546565
A8T13	A6T13	0.773023088
A8T14	A6T14	0.879189699
A8T15	A6T15	0.775121882

Different than the previous corpora, the overall performance between texts are found mostly around 1 – which indicates the absolute difference between the texts. However, there are also some values around 0.33 which never found before, but it does not reflect the general. Additionally, there are no values below 0.69 in Table 6-13, and there are three scores above 0.80, but they are not considered less efficient due to the common distribution of 1- the absolute difference in the entire list. There are 53 linguistic features in Corpus3, while 19 of them are 19, 30 syntactic and the rest 4 is structural. The distribution of the shared features between A6 and A8 is presented in Table 6-14. Due to the limited size of Corpus3, any shared features are demonstrated without looking at the frequencies.

Table 6-14: Shared features between A6 and A8 in Corpus3.

FEATURES	AUTHOR6	AUTHOR8
F1 Lex bir	11	12
F2 Lex bu kadar	0	0
F3 Lex filan	2	3
F4 Lex gibi bir	0	0
F5 Lex Intj lan	3	3
F6 Lex Intj ya	1	4
F7 Lex ne kadar	0	2
F8 Lex o kadar	1	0
F9 Lex Redup MDoubl	1	0
F10 Lex ya da	1	3
F11 Lex Emotext	0	1
F12 Lex falan	0	0
F13 Lex filan	2	3
F14 Lex Keyboard Emoticon	3	3
F15 Lex Numbers (Numeral)	3	5
F16 Lex Numbers (Word)	0	3
F17 Lex Omission RClip	0	1
F18 Lex ProsodicEmp	2	2

F19 Lex Redup SameForm	6	4
F20 Str edit	0	0
F21 Str bkz	2	2
F22 Str Hyperlink	0	0
F23 Str Itemise	0	0
F24 Syn ConConj ama	5	5
F25 Syn ConConj bile	1	2
F26 Syn ConConj cunku	0	1
F27 Syn ConConj ve	6	8
F28 Syn ConConj zaten	2	0
F29 Syn Emp1st Person	0	0
F30 Syn EndConj cunku	0	1
F31 Syn EndConj zaten	1	0
F32 Syn MixedTypoPunc	8	6
F33 Syn MultiTypoPunc Exclamation	1	0
F34 Syn MultiTypo QuestionMark	0	1
F35 Syn NonStnPunc Colon	0	0
F36 Syn NonStnPunc Double Full Stop	0	1
F37 Syn NonStnPunc Fullstop	0	0
F38 Syn NonStnPunc SemiColon	0	0
F39 Syn NonStnSpace BeforePunc	0	0
F40 Syn ParanthCla	3	4
F41 Syn Punc AbsApost	0	0
F42 Syn Punc AbsFullstop	0	0
F43 Syn Punc AbsSpa	0	1
F44 Syn SenQuote	12	8
F45 Syn StnPunc Colon	2	1
F46 Syn StnPunc Ellipsis	8	5
F47 Syn StnPunc Exclm	0	0
F48 Syn StnPunc QuestionMa	0	0
F49 Syn StnPunc Semicolon	1	0
F50 Syn StnPunc Slash	0	0
F51 Syn StrConj ama	0	0
F52 Syn StrConj ve	0	0
F53 Syn StrConj zaten	0	0

In Table 6-14,19 out of 53 features that are 35% of the features used at least once between the authors while there is a 19 times agreement in the absence of some features. In this case, A6 and A8 shared 38 features in total which is 71% of the entire features in the corpus. There are also some distinctive features that are used often than the others. For instance, F1 *bi* has the highest frequency between the other features, when the entire presence/absence matrix is

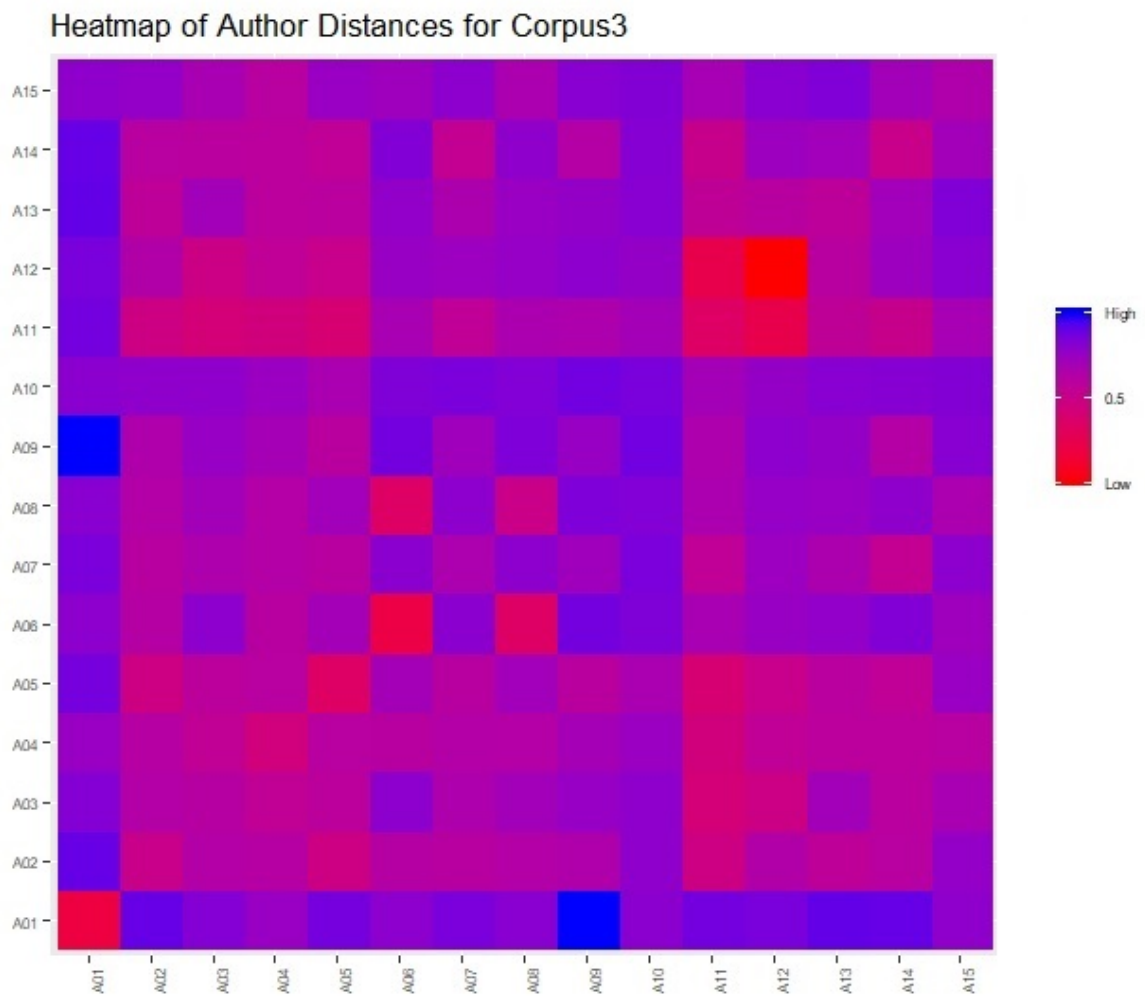
investigated it is found that this feature is used 157 times, thus it is one the most frequent feature in this corpus for that reason this feature is not enough to indicate a high level of distinctiveness. On the one side, F32 *Mixed Typo Punc* is used 14 times by the authors; accordingly, the frequency of this feature is 14 in the entire list. This is a strong feature which shows the consistency and distinctiveness between A6 and A8. On the other side, F44 *SentenceQuotation* and F46 *StndPunc Ellipsis* are used 20 and 13 times by the authors respectively, and these features are used 42 and 24 by the all authors including the disputed and the candidate author. In this case, more than fifty percent of the usage belongs to only A6 and A8. It is evident that there is some linguistic connection between the disputed author and the potential author A8. In Table 6-15 this connection is presented in author vs author raw distance scores comparison.

Table 6-15: Distances between authors in Corpus3.

	row.names	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15
1	A01	0.7464436	0.9079504	0.8901081	0.8739558	0.9001706	0.8838896	0.8972409	0.8866700	0.9303731	0.8854596	0.9006719	0.8979323	0.9093311	0.9079123	0.8823585
2	A02	0.9079504	0.8184467	0.8465637	0.8440245	0.8112434	0.8443515	0.8414797	0.8471189	0.8508693	0.8830110	0.8116032	0.8487452	0.8334173	0.8397584	0.8787789
3	A03	0.8901081	0.8465637	0.8426070	0.8292227	0.8363280	0.8835237	0.8528623	0.8652658	0.8756223	0.8831970	0.8007725	0.8136041	0.8652585	0.8385115	0.8589456
4	A04	0.8739558	0.8440245	0.8292227	0.8041568	0.8400932	0.8410718	0.8468110	0.8462364	0.8621002	0.8727669	0.8056503	0.8311749	0.8374935	0.8375624	0.8400171
5	A05	0.9001706	0.8112434	0.8363280	0.8400932	0.7804837	0.8637195	0.8409224	0.8656208	0.8393457	0.8581526	0.7945238	0.8193086	0.8385394	0.8307354	0.8742937
6	A06	0.8838896	0.8443515	0.8835237	0.8410718	0.8637195	0.7498549	0.8850974	0.7805873	0.9006925	0.8938505	0.8583540	0.8751904	0.8801425	0.8924368	0.8685586
7	A07	0.8972409	0.8414797	0.8528623	0.8468110	0.8409224	0.8850974	0.8544974	0.8844266	0.8677213	0.8969937	0.8304268	0.8709344	0.8551265	0.8258430	0.8833340
8	A08	0.8866700	0.8471189	0.8652658	0.8462364	0.8656208	0.7805873	0.8844266	0.8164933	0.8941980	0.8919320	0.8557582	0.8766019	0.8738781	0.8830041	0.8553492
9	A09	0.9303731	0.8508693	0.8756223	0.8621002	0.8393457	0.9006925	0.8677213	0.8941980	0.8753515	0.9020454	0.8527443	0.8840914	0.8775572	0.8451605	0.8881090
10	A10	0.8854596	0.8830110	0.8831970	0.8727669	0.8581526	0.8938505	0.8969937	0.8919320	0.9020454	0.8974506	0.8637645	0.8781984	0.8879243	0.8894316	0.8916015
11	A11	0.9006719	0.8116032	0.8007725	0.8056503	0.7945238	0.8583540	0.8304268	0.8557582	0.8527443	0.8637645	0.7803855	0.7572257	0.8314566	0.8198607	0.8595724
12	A12	0.8979323	0.8487452	0.8136041	0.8311749	0.8193086	0.8751904	0.8709344	0.8766019	0.8840914	0.8781984	0.7572257	0.7040023	0.8410079	0.8699065	0.8874175
13	A13	0.9093311	0.8334173	0.8652585	0.8374935	0.8385394	0.8801425	0.8551265	0.8738781	0.8775572	0.8879243	0.8314566	0.8410079	0.8341414	0.8659681	0.8928647
14	A14	0.9079123	0.8397584	0.8385115	0.8375624	0.8307354	0.8924368	0.8258430	0.8830041	0.8451605	0.8894316	0.8198607	0.8699065	0.8659681	0.8184429	0.8651140
15	A15	0.8823585	0.8787789	0.8589456	0.8400171	0.8742937	0.8685586	0.8833340	0.8553492	0.8881090	0.8916015	0.8595724	0.8874175	0.8928647	0.8651140	0.8512082

As seen in Table 6-15 the overall raw Jaccard scores are above 0.80; however A8 show the smallest distance to A8 apart from the intra-author similarities like A12. The disputed and the candidate author showed a high rate agreement on the presence and the absence of the features. Moreover, combining the texts a single text demonstrated the linguistic connection between A6 and the A8 aligning with the shared features and the mean Jaccard distance scores presented above. Additionally, these values are visualised in a heatmap in Figure 6-5.

Figure 6-5: HeatMap of author distances in Corpus3.



In furtherance the raw score, the heatmap presents a square shape red pattern in corresponding to the authors A6 and A8. There are also some other visible patterns apart from A6; however, these boxes demonstrate the distances between themselves like in A12. Overall, it is likely to assign the correct authorship in short size texts with the current feature set. Even though previous studies on short size texts such as Twitter messages (e.g. Layton et al. 2010) are included genre specific structural features into their datasets, Eksi Sozluk is a monotype website which does not allow different structural features. However, it still provides evidence and signally contributes the authorship attribution in short size Turkish texts.

6.4.2. Candidate Author Size

In this section, the effect of the candidate author size in authorship studies when working on a forensic text. Although it is an essential factor in the performance, it received less attention. As it is discussed in Section 2.4.2. the stylometric studies which focus on candidate author size establish accurate results when the candidate author set size is large while the number of candidate author size is a few in stylistic based studies. By increasing the number of candidates,

stylistics approaches need a robust method while stylometric approaches demonstrate a positive effect on performance. On the other hand, a small number of candidate authors may lead an overestimation in results. However, in this study, it is aimed to increase the number of authors from fifteen (default author set size in the present study) to thirty. This section presents the performance of the authorship attribution approaches in a controlled number of candidate authors in order to assess the success of the method on different sets of candidate authors.

As it is mentioned previously, most studies are focused on two/three or a few authors in authorship attribution. In this section, the results of depending on the candidate author set size are presented which increased gradually from 15 to 30. The risks of using a few authors discussed before which leads to overestimating the accuracy rates.

6.4.2.1. 30 Authors

This set of authors are different from previous datasets where it is used 15 texts from 15 authors. Corpus 1 and Corpus 2 are combined in order to demonstrate the effect of candidate author set size in authorship attribution studies. This has the sample with the largest number of words and candidate authors. However, the approach remains the same as the other studies. When the data sets combined for the study, the common features were listed from both corpora. As it is mentioned above, there are small nuances in feature selection between corpora. Therefore, a set of 51 common linguistic features for this test. The disputed author and the actual author are settled as A1 and A15 however; they do not refer the same A1 and A15 in Corpus1. The smallest mean distances investigated between authors and selected when it hit one of the lowest scores at least five times. In this training test, as it is presented in Table 6-16, A1 and A15 indicated lower results when they compared with the rest of the authors.

Table 6-16: Smallest mean distances between texts for 30 authors.

AUTHORX	AUTHOR15	DISTANCE VALUES
A1T1	A15T1	0.687707848
A1T2	A15T2	0.670243057
A1T3	A15T3	0.707371039
A1T4	A15T4	0.72725517
A1T5	A15T5	0.752207101
A1T6	A15T6	0.65480945
A1T7	A15T7	0.674223184

A1T8	A15T8	0.641413848
A1T9	A15T9	0.713496955
A1T10	A15T10	0.67743593
A1T11	A15T11	0.8049398
A1T12	A15T12	0.762421443
A1T13	A15T13	0.745269126
A1T14	A15T14	0.686477167
A1T15	A15T15	0.85039204

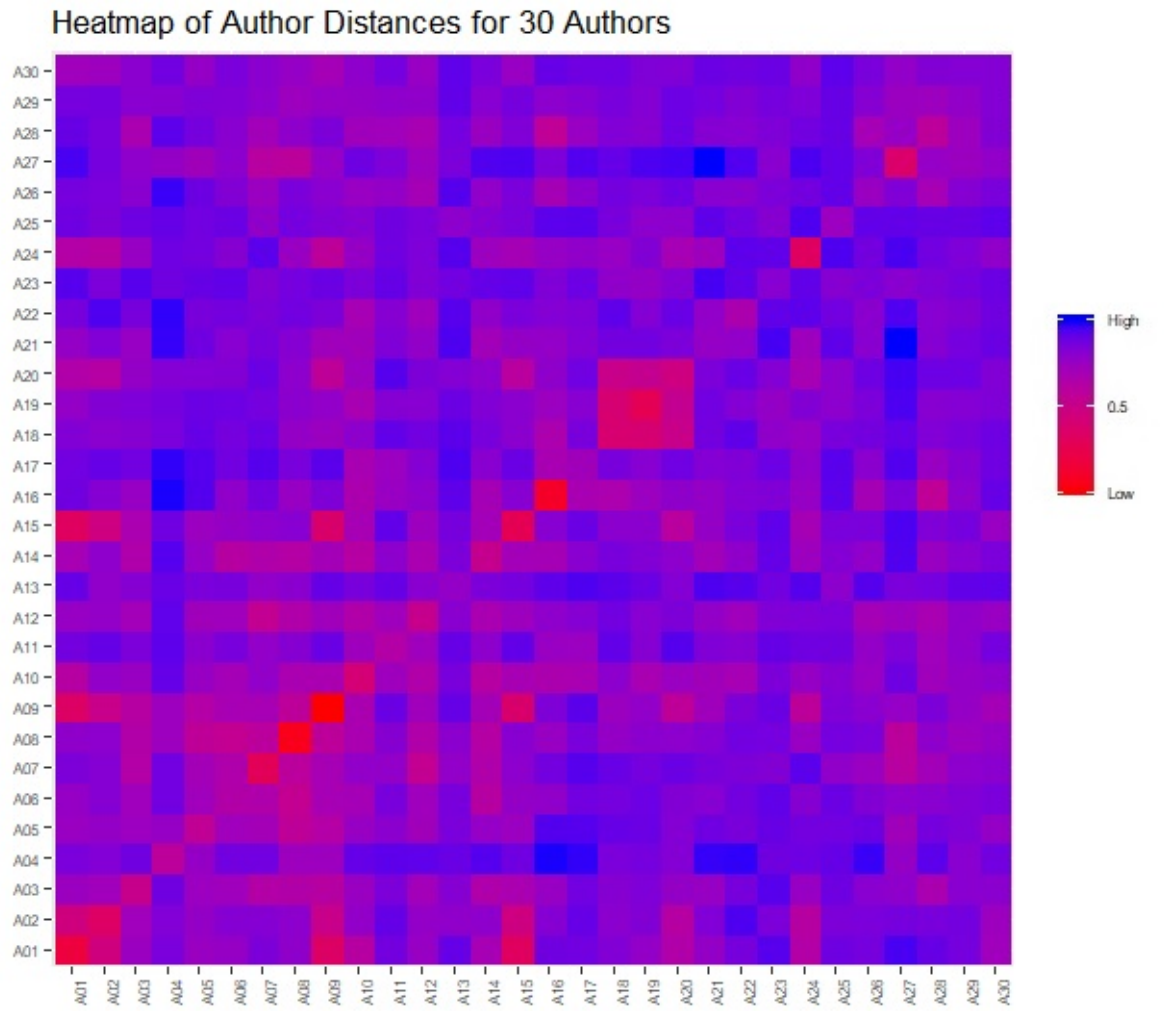
Although the general results are between 0.80 and 0.99 in this dataset, it is possible to find at least seven values which are around 0.60 that shows a significant level of agreement on the linguistic features between the disputed author and the candidate author. This leads a further investigation on the linguistic connection between A1 and A15. In Table 6-17 shared features between A1 and A15 is presented; however, due to the number of texts, linguistic feature frequency threshold is determined as 10 similar to Corpus1. Thus, the shared features demonstrated unless they used at least 10 times by both authors.

Table 6-17: Shared features between A1 and A15.

FEATURES	**AUTHOR1**	**AUTHOR15**
F1 Lex bir de	15	15
F3 Lex ben de	10	10
F8 Lex bu kadar	0	0
F10 Lex en iyi	10	3
F16 Lex amina koyim	0	0
F17 Lex amk	0	0
F21 Lex falan	0	0
F22 Lex Init.akp	10	4
F25 Lex Intrj. e	0	0
F26 Lex Intrj. Ha	0	0
F27 Lex Intrj. Ya	0	0
F28 Lex Intrj. yuh	0	0
F32 Lex Numbers (Word)	0	0
F34 Lex OmissionRClip	0	0
F35 Lex ProsodicEmp.	0	0
F36 Lex Redup.Mdoublet.	8	13
F39 Str bkz	0	0
F40 Str. Edit	0	0
F49 Syn ConConj hatta	15	13

There are six features are present in high occurrences both in A1 and A15; however, when the entire list is examined, it is found that F1 *bir de* occurred 128 across the texts, although 23% usage belongs to them, this feature is not enough to show the distinctiveness of the authors. Moreover, F49 *ConConj hatta* is used 28 times in total by both authors, and in general, this feature is used 70 times only, that is 40% of this feature shared by the A1 and A15. On the other side, there are 13 different features that are not used even a single time by the authors, when the attitude of the rest of the authors examined, it is found that F16 *amina koyim* feature is so popular between the rest of the 28 authors which is 99 but not in the texts of A1 and A15. The second example is F35 *ProsodicEmph* used 50 times by the other authors while A1 and A15 did not prefer it in their writings. The results are promising even though 51 common linguistic features are used only for this training test. It is evident that there is some linguistic connection between A1 and A15. This analysis included 30 authors who mean 900 texts in total, even though the concatenating process is applied in author vs author comparison. The output presented 450 values in the table which is not possible to present in here like in the previous tests. However, the output is used to produce a heatmap to visualise the 450 values in one place in Figure 6-6.

Figure 6-6: HeatMap of author distances for 30 authors.



It is seen that similar patterns are demonstrated in Figure 6-6 using the raw Jaccard scores between authors. At first sight, there are some patterns which include the traces from the same authors in the diagonal line. Moreover, there is a visible red square in the middle of the figure which represents the A18, A19 and A20. As it is mentioned above, the heatmaps indicate all similarities between the entire author list; however, in this dataset, the main concern is the A1 since it is the disputed author. When the raw distance scores are compared with A1 none of the author above (A18, A19 and A20) represents any similarities. For instance, between A1 and A19 the value is over 0.86 in a similar manner, A20 has is over 0.94. On the other side, A1 and A15 have 0.74 distance with concatenating all single texts. This similarity score can be in the boxes correspond the A1 and A15; they have the same level of red tinted pattern. In this case, it is likely to say that 51 features are enough to find the linguistic link between the disputed author and the actual author. This linguistic link between A1 and A15 lead a consistent and

distinctive attitude than the rest of the authors. In the end, it may be concluded that with supportive data visualisation methods, it is possible to find similar patterns between author.

6.4.3. Limited Texts per Author

As it is discussed above, there are some cases where there is only one disputed text and a couple of candidate authors such as Mr Gogging case which analysed from Coulthard (2013) and the task to find the author of an email between four candidate authors. Since this situation is anticipated in real life cases, it is decided to examine the number of texts lower than the default number (15) in this study. With this aim, two different data sets are derived from Corpus1 – which is the case Author1 and Author15 matching authors. First, the performance of the five texts per author and the second, the performance of the ten texts is tested.

6.4.3.1. Five Texts

In this test 59 features were tested across 75 texts from 14 different authors. The feature set is the same one that was used in Corpus1 (Section 6.4.1.1.). It is worth to note that, it was not possible to use all features from Corpus1 since some of them occurred in the following texts and these features were not used in the five texts that were randomly selected between fifteen texts. As it is mentioned above, each author has only five texts for training purposes. A1 and A15 are selected as corresponding authors in this dataset. Due to the limitation in available texts per author, the figures are simple to follow. Contrary to the previous procedure applied in the other tests, at least five smallest values were not searched instead all values investigated and the one who had the smallest value chosen for the further investigation. Table 6-18 indicates the smallest mean distances between texts for five texts per author.

Table 6-18: Smallest mean distances between texts for 5 texts per author.

AUTHORX	AUTHOR15	DISTANCE VALUES
A1T1	A15T1	0.707218396
A1T2	A15T2	0.725362591
A1T3	A15T3	0.68397878
A1T4	A15T4	0.712074251
A1T5	A15T5	0.819493761

The average performance between texts are above 0.85; however, the smallest distance value belongs to A15 that is 0.68 in the entire dataset for that reason it is selected as a candidate author. Since the number of texts is not adequate, it is essential to reveal the shared features between the authors. These shared features are presented in Table 6-19 without considering the occurrences of them within the texts.

Table 6-19: Shared features between A1 and A15.

FEATURES	AUTHOR1	AUTHOR15
F1 Lex bir de	2	3
F2 Lex o kadar	2	1
F5 Lex - bi sey	2	5
F6 Lex bi sure	0	0
F7 Lex boyle bi	1	1
F8 Lex bu kadar	1	1
F9 Lex bugune kadar	0	0
F11 Lex gibi bi	1	2
F14 Lex ya da	3	3
F16 Lex amina koyim	0	0
F18 Lex- bi	5	5
F20 Lex- en az bir kez	0	0
F21 Lex- falan	2	3
F22 Lex- Int. akp	0	0
F23 Lex- Int. tl	0	0
F24 Lex- Interjection ay	0	0
F30 Lex- lan	4	5
F31 Lex Numbers (Numeral)	2	1
F32 Lex Numbers (Word)	3	2
F33 Lex olm	1	2
F34 Lex- Omission R Clip	0	0
F35 Lex ProsodoEmph	3	2
F36 Lex- Redup Mdoublet	0	0
F37 Lex- Reduplication/ Same Form	5	1
F38 Lex- tabiki	0	0
F41 Str - Hyperlink	0	0
F42 Str- Itemise	0	0
F44 Syn- NonStndPunc DoubleDot	0	0
F45 Syn- ConConj ama	0	0
F46 Syn- ConConj ancak	0	0
F47 Syn- ConConj cunku	0	0
F48 Syn- ConConj fakat	0	0
F49 Syn- ConConj hatta	0	0
F51 Syn- ConConj mesela	0	0
F52 Syn- ConConj yada	2	3

F53 Syn- Emp1stperson	2	2
F54 Syn- EndConj ancak	0	0
F56 Syn- EndConj hatta	0	0
F59 Syn- MixedTypo. Punctuation	0	0
F60 Syn- MultiTypo. Punc. Exclamation	0	0
F61 Syn-MultiTypo.Punc. Question Mark	0	0
F62 Syn- NonStdPunc. Question Mark	0	0
F63 Syn- NonStdPunc Semicolon	2	1
F64 Syn- Parant.Clause	1	3
F66 Syn- Punc.Empahis Exclamation	0	0
F67 Syn- Punc.NonStdSpa	0	0
F68 Syn-Punc. Absence Apostrophe	0	0
F69 Syn-Punc.Absence Fullstop	0	0
F70 Syn- Sentence quote	4	5
F71 Syn- StnPunc Ellipsis	3	3
F72 Syn- StndPunc Semicolon	0	0
F73 Syn- StrConj ama	0	0
F74 Syn- StrConj ancak	0	0
F76 Syn- StrConj cunku	1	2
F77 Syn- StrConj fakat	2	1
F78 Syn- StrConj hatta	0	0
F79 Syn- StrConj herhalde	0	0
F82 Syn- StnPunc Colon	1	1
F83 Syn- StnPunc. Excl.	3	3

There are not many widely used features due to the limited number of texts per author; however there are still some patterns show the agreement between the authors. For instance, F5 *bi sey* 2-word gram is used 7 times in total by the disputed and the candidate author. Moreover, this feature is used 8 times in the entire dataset. A similar situation occurred in F30 *lan;* this feature used 20 times by different authors which are mostly shared between A1 and A15 which is equal to 45% of this feature in the corpus. Additionally, F18 *bi* and F83 *StnPunc Exclamation* are shared 10 and 6 times by both authors respectively while the general usage is 25 and 15. That is to say, more than 40% of these features are used by A1 and A15 only while 13 authors shared the rest. Moreover, non-used features show the agreement between the authors; for instance, F72 *StnPunc Semicolon* was used 12 times between the 75 texts however neither A1 nor A15 employed this feature in their texts. Depending on the linguistic connection found between A1 and A15, author vs author comparison is employed in Table 6-20.

Table 6-20: Distances between authors for 5 texts per author.

	row.names	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15
1	A01	0.7218745	0.7945013	0.8470534	0.8850133	0.8923471	0.8969378	0.8350882	0.8103375	0.7786583	0.8275679	0.9026232	0.9231445	0.9227018	0.8420875	0.7296256
2	A02	0.7945013	0.7538865	0.8975413	0.8896720	0.9085011	0.9167236	0.8932608	0.9042043	0.8335289	0.8877856	0.9433000	0.9048762	0.9082739	0.9052785	0.7838250
3	A03	0.8470534	0.8975413	0.8553102	0.9028336	0.9020464	0.8769065	0.8312546	0.8392960	0.8444377	0.8764702	0.9189893	0.8782756	0.8942555	0.8994166	0.8856295
4	A04	0.8850133	0.8896720	0.9028336	0.8534848	0.8971189	0.9210556	0.9339089	0.8901485	0.8664653	0.9343749	0.9393570	0.9618886	0.9535859	0.9456094	0.9185626
5	A05	0.8923471	0.9085011	0.9020464	0.8971189	0.9024608	0.9147143	0.9035605	0.9053746	0.8616889	0.8927887	0.9114607	0.9597772	0.9493355	0.8938969	0.8935832
6	A06	0.8969378	0.9167236	0.8769065	0.9210556	0.9147143	0.8702564	0.8626116	0.8404472	0.8696683	0.9033338	0.9490463	0.9217509	0.9299738	0.8962630	0.9163588
7	A07	0.8350882	0.8932608	0.8312546	0.9339089	0.9035605	0.8626116	0.7535681	0.8562871	0.8183539	0.8601501	0.8896967	0.8749167	0.9217852	0.8839392	0.8638384
8	A08	0.8103375	0.9042043	0.8392960	0.8901485	0.9053746	0.8404472	0.8562871	0.5958535	0.8616342	0.8588778	0.9102208	0.9003902	0.9237483	0.8159931	0.8943268
9	A09	0.7786583	0.8335289	0.8444377	0.8664653	0.8616889	0.8696683	0.8183539	0.8616342	0.7461541	0.8916982	0.9501661	0.9392876	0.9256016	0.8997084	0.7880954
10	A10	0.8275679	0.8877856	0.8764702	0.9343749	0.8927887	0.9033338	0.8601501	0.8588778	0.8916982	0.7495009	0.8680270	0.9181663	0.9002061	0.8424180	0.8476541
11	A11	0.9026232	0.9433000	0.9189893	0.9393570	0.9114607	0.9490463	0.8896967	0.9102208	0.9501661	0.8680270	0.9240476	0.9157573	0.9535472	0.9043308	0.9310269
12	A12	0.9231445	0.9048762	0.8782756	0.9618886	0.9597772	0.9217509	0.8749167	0.9003902	0.9392876	0.9181663	0.9157573	0.8985931	0.9127976	0.9035515	0.9126606
13	A13	0.9227018	0.9082739	0.8942555	0.9535859	0.9493355	0.9299738	0.9217852	0.9237483	0.9256016	0.9002061	0.9535472	0.9127976	0.8919289	0.9599385	0.9206140
14	A14	0.8420875	0.9052785	0.8994166	0.9456094	0.8938969	0.8962630	0.8839392	0.8159931	0.8997084	0.8424180	0.9043308	0.9035515	0.9599385	0.7686890	0.8662689
15	A15	0.7296256	0.7838250	0.8856295	0.9185626	0.8935832	0.9163588	0.8638384	0.8943268	0.7880954	0.8476541	0.9310269	0.9126606	0.9206140	0.8662689	0.7047594

In accord with the previous findings, Table 6-20 provides evidence to show the lowest raw Jaccard values for Author1 and Author15 in the entire distance matrix. Moreover, there is clear consistency between Author1 and Author15 which leads to correct classification in authorship attribution. Although there are some lowest intra-author values (e.g. A8 vs A8 0.59) these are ignored since it does not contribute to the general discussion in attributing the disputed author. It is evident that even there are a few texts by the disputed author it is still possible to indicate a consistent linguistic distinctiveness with the authors and consistency with the candidate author with the aid of an extensive linguistic features list. Additional evidence is presented in Figure 6-7 to demonstrate the similarities between A1 and A15.

Figure 6-7: HeatMap of author distances for 5 texts per author.

As it is presented in Figure 6-7, there are some red boxes which indicate the similarity between the authors. In the bottom-left corner, there is a square red pattern refers to A1 and A2; the reason of this density is that each author comes cross with each other in this part. Moreover, there are some red boxes in the diagonal line as it is expected. However, the patterns in the corners display a similar shade according to the colour threshold in the heatmap.

Overall, the results in this test are consistent with the previous tests that used the same linguistic features. It is verified the reliability of the features in different conditions even with a limited text per author.

6.4.3.2. Ten Texts

The second test in this section is related to ten texts. Similar to the previous scenario, there are ten texts available per author, and the number of authors stays the same. Ten texts are not selected in order instead randomly similar to five tests training test, for that reason the features do not follow the sequence either, which is possible to find F1 and F83 at the same time. Moreover, in this test 61 features were tested across 100 texts. When the smallest values were investigated after calculating the mean distances between texts for texts per author, one candidate author is found for further investigation. The best-averaged performance of between texts is reported below.

Table 6-21: Smallest mean distances between texts for 10 texts per author.

AUTHORX	AUTHOR15	DISTANCE VALUES
A1T1	A15T1	0.7300322
A1T2	A15T2	0.729789095
A1T3	A15T3	0.718508925
A1T4	A15T4	0.726807915
A1T5	A15T5	0.800291673
A1T6	A15T6	0.759678011
A1T7	A15T7	0.746072951
A1T8	A15T8	0.760264598
A1T9	A15T9	0.754642405
A1T10	A15T10	0.738149525

In the distance matrix, the average of the performance is around 0.80, however as it is presented in Table 6-21, A15 has only one value that is the same distance while rest of them hit the lowest score at least a few times. For that reason, A15 is selected as a candidate author in this dataset. When we recall the test between 14 distinct authors with 5 texts in Section 6.4.3.1, it confirms the same situation, although the texts are almost different from the previous test. In order to investigate the reason behind, shared features between the selected authors are revealed in Table 6-22. All shared features are presented without any limitations and thresholds.

Table 6-22: Shared features between A1 and A15.

FEATURES	AUTHOR1	AUTHOR15
F1 Lex- bir de	2	6
F2 Lex- o kadar	4	2
F3 Lex- ben de	1	3
F5 Lex - bi sey	5	8
F6 Lex- bi sure	0	0
F7 Lex- boyle bi	2	1
F8 Lex- bu kadar	3	2
F11 Lex- gibi bi	4	3
F13 Lex- o da	1	3
F14 Lex- ya da	4	7
F18 Lex- bi	10	10
F19 Lex- Emotext	2	2
F20 Lex- en az bir kez	0	0
F21 Lex- falan	3	7
F22 Lex- Int. akp	0	0
F23 Lex- Int. tl	0	0
F24 Lex- Interjection ay	0	0
F26 Lex Intrj. Ha	1	3
F27 Lex Intrj. Ya	3	4
F28 Lex Intrj. yuh	1	1
F30 Lex- lan	9	10
F31 Lex Numbers (Numeral)	3	1
F32 Lex Numbers (Word)	6	5
F33 Lex olm	5	4
F34 Lex- Omission R Clip	0	0
F35 Lex ProsodoEmph	5	2
F36 Lex- Redup Mdoublet	0	0
F37 Lex- Reduplication/ Same Form	8	2
F38 Lex- tabiki	0	0
F41 Str - Hyperlink	0	0
F44 Syn- NonStndPunc DoubleDot	1	1
F45 Syn- ConConj ama	0	0
F46 Syn- ConConj ancak	0	0
F47 Syn- ConConj cunku	0	0
F48 Syn- ConConj fakat	0	0
F51 Syn- ConConj mesela	0	0
F52 Syn- ConConj yada	3	6
F53 Syn- Emp1stperson	5	3
F54 Syn- EndConj ancak	0	0
F59 Syn- MixedTypo. Punctuation	0	0
F60 Syn- MultiTypo. Punc. Exclamation	0	0

F61 Syn-MultiTypo.Punc. Question Mark	0	0
F62 Syn- NonStdPunc. Question Mark	0	0
F63 Syn- NonStdPunc Semicolon	3	1
F64 Syn- Parant.Clause	3	6
F65 Syn- Punc AbsSpa	1	5
F67 Syn- Punc.NonStdSpa	0	0
F68 Syn-Punc. Absence Apostrophe	0	0
F69 Syn-Punc.Absence Fullstop	0	0
F70 Syn- Sentence quote	7	10
F71 Syn- StnPunc Ellipsis	5	3
F72 Syn- StndPunc Semicolon	0	0
F73 Syn- StrConj ama	0	0
F74 Syn- StrConj ancak	0	0
F76 Syn- StrConj cunku	2	2
F77 Syn- StrConj fakat	4	2
F78 Syn- StrConj hatta	2	1
F80 Syn- StrConj mesela	1	2
F81 Syn- StrConj ya da	2	1
F82 Syn- StndPunc Colon	2	1
F83 Syn- StnPunc. Excl.	3	4

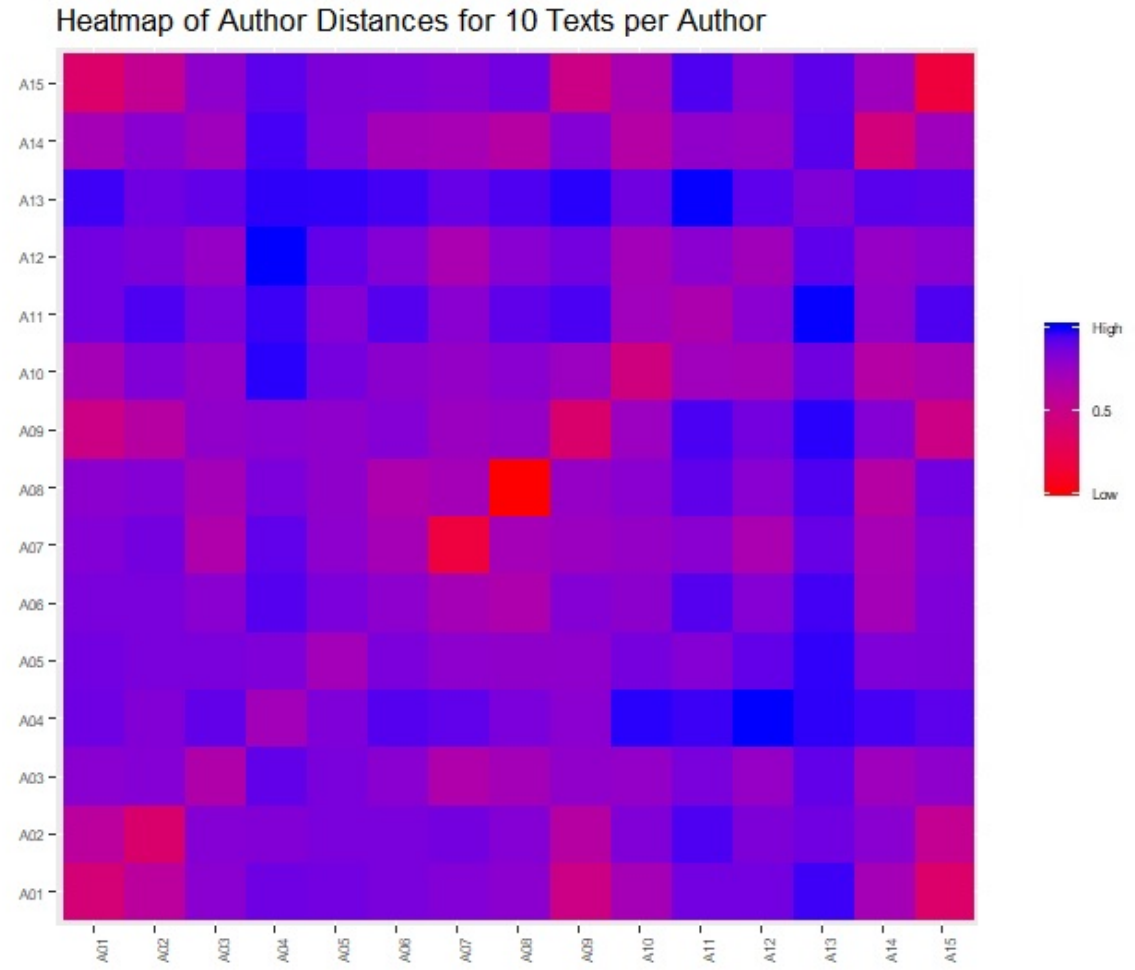
Even though there are not many high-frequency features because of the text size, there are still some usages to investigate the linguistic link between A1 and A15. For example, F5 *bi sey* 2-word gram is used by the selected authors 13 times, as it is used 15 times, in general, this rate makes 86% of this usage. In a similar manner, F18 *bi* is used 20 times in total by the disputed author and the candidate author while this feature is used 49 times only in all 100 texts. Moreover, F30 *lan* is used 38 times in the entire dataset while 19 of them already used by A1 and A15. These rates are significant in establishing the linguistic similarity between the disputed and the candidate authors. On the other side, as it is mentioned above absent features are meaningful in this context either. For instance, neither A1 nor A15 used F72 *StnPunctuation Semicolon* feature in their texts; however, this feature is found in 26 different texts that are more than a quarter in the entire dataset. Depending on the linguistic similarities between the A1 and A15 author vs author comparison is employed in the following Table 6-23.

Table 6-23: Distances between authors for 10 texts per author.

	row.names	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15
1	A01	0.7633800	0.8190994	0.8885807	0.9132089	0.9099056	0.9035612	0.8955703	0.8872044	0.7831800	0.8534200	0.9103445	0.9109229	0.9392354	0.8534563	0.7464237
2	A02	0.8190994	0.7481319	0.8950448	0.8962104	0.9041298	0.9038061	0.9081762	0.8937251	0.8259135	0.8980515	0.9327895	0.9007202	0.9123534	0.8895617	0.8026606
3	A03	0.8885807	0.8950448	0.8370326	0.9217764	0.9041727	0.8890320	0.8386675	0.8559993	0.8802637	0.8777277	0.9039857	0.8745431	0.9219201	0.8644425	0.8824768
4	A04	0.9132089	0.8962104	0.9217764	0.8582609	0.8993134	0.9290997	0.9227504	0.9026070	0.8879952	0.9455468	0.9400434	0.9502679	0.9440222	0.9368725	0.9246319
5	A05	0.9099056	0.9041298	0.9041727	0.8993134	0.8569331	0.9029031	0.8855962	0.8820719	0.8834362	0.9064991	0.8942201	0.9212706	0.9433118	0.8990788	0.9006062
6	A06	0.9035612	0.9038061	0.8890320	0.9290997	0.9029031	0.8856961	0.8534941	0.8404682	0.8947214	0.8860682	0.9295135	0.8932176	0.9373670	0.8560551	0.8999126
7	A07	0.8955703	0.9081762	0.8386675	0.9227504	0.8855962	0.8534941	0.6854106	0.8550156	0.8683047	0.8760118	0.8897187	0.8458965	0.9182659	0.8519042	0.8944129
8	A08	0.8872044	0.8937251	0.8559993	0.9026070	0.8820719	0.8404682	0.8550156	0.6298230	0.8748865	0.8890959	0.9232923	0.8907250	0.9318815	0.8275769	0.9107844
9	A09	0.7831800	0.8259135	0.8802637	0.8879952	0.8834362	0.8947214	0.8683047	0.8748865	0.7499500	0.8671603	0.9341984	0.9083389	0.9453958	0.8935631	0.7839336
10	A10	0.8534200	0.8980515	0.8777277	0.9455468	0.9064991	0.8860682	0.8760118	0.8890959	0.8671603	0.7756973	0.8609416	0.8576349	0.9111306	0.8298490	0.8459567
11	A11	0.9103445	0.9327895	0.9039857	0.9400434	0.8942201	0.9295135	0.8897187	0.9232923	0.9341984	0.8609416	0.8438846	0.8878855	0.9498299	0.8810360	0.9322420
12	A12	0.9109229	0.9007202	0.8745431	0.9502679	0.9212706	0.8932176	0.8458965	0.8907250	0.9083389	0.8576349	0.8878855	0.8601525	0.9242276	0.8746250	0.8888708
13	A13	0.9392354	0.9123534	0.9219201	0.9440222	0.9433118	0.9373670	0.9182659	0.9318815	0.9453958	0.9111306	0.9498299	0.9242276	0.8986714	0.9271663	0.9237838
14	A14	0.8534563	0.8895617	0.8644425	0.9368725	0.8990788	0.8560551	0.8519042	0.8275769	0.8935631	0.8298490	0.8810360	0.8746250	0.9271663	0.7691099	0.8635877
15	A15	0.7464237	0.8026606	0.8824768	0.9246319	0.9006062	0.8999126	0.8944129	0.9107844	0.7839336	0.8459567	0.9322420	0.8888708	0.9237838	0.8635877	0.6850929

In a similar manner with the previous findings, Table 6-23 demonstrates the distances between authors. The lowest raw Jaccard values are indicated with a red line in the distance matrix. When A1 is crossed with A15 in the matrix, they both render the lowest inter author scores. It is apparent that 61 features were enough to obtain the linguistic distinctiveness and the consistency between authors even with a limited number of texts. Moreover, combining the texts a single chunk provided to demonstrate the linguistic connection between A1 and A15 clearly in this distance matrix. No other value is lower than 0.74 when the intra-author distances are disregarded. In order to visualise the similarities between authors, a heatmap is provided in Figure 6-8. It should be stated that at the visualisation stage, heatmap uses the entire information on the raw Jaccard distance matrix, for that reason, it is possible to see any other similarities between the other authors rather than the disputed author.

Figure 6-8: HeatMap of author distances for 10 texts per author.



As it is mentioned above, there is a red line in the diagonal line that is equated to lower degrees in the colour thresholding on the right side in Figure 6-8. Apart from the boxes with less red colour, when it is focussed on the A1 box, it is hit the same colour range as A15.

In summary, the results in this section are consistent with 5 texts test (Section 6.4.3.1.) and Corpus1-Long Texts (Section 6.4.1.2.) that presented the similarity between A1 and A15. Moreover, it is supported by the shared linguistic connection between the authors. Thus, it is possible to determine consistency and distinctiveness with limited available texts per author with a comprehensive combined feature list.

6.5. Section Conclusion

In this section, the way in which the size affects the accuracy of authorship was investigated. First Corpus1-long size texts were investigated, and there were found some common traces between Author1 and Author15, in the same way, Corpus2- medium size texts presented an attribution between the questioned and corresponding author. Despite the size of the texts, the

linguistic link between the disputed author the corresponding author was also found the author vs author comparison distance values and heat map were confirmed that there is a featured matching between Author12 and Author13. Furthermore, in Corpus3 even though there were average 47 words in this dataset that is relatively small forensic linguistics data. The smallest mean distances referred to the corresponding author when it was compared with the A6. For this data set, 34 features out of 53 features were shared between authors. Furthermore, candidate author set size is tested in this section and found that the method is still applicable to 30 authors. Finally, in parallel to real-world cases, five texts and ten texts taken from Corpus1 and tested the applicability of the method. As a result, both tests performed well and led to correct attribution between disputed and the corresponding author.

6.6. Cross-Genre Authorship Analysis

In this test, Eksi Sozluk corpora are compared with the texts from Twitter. The structure of Twitter is very different from other online mediums like blogs, Facebook status messages or Wiki-like collaborative online encyclopaedias. Although Grant (2013) mentioned the necessity of relevant comparison corpus from the same genre, this study is applied in cross-genre texts; however, Twitter and Eksi Sozluk are counted as allied genre depending on the similarities between text productions in this study. Moreover, entries in Eksi Sozluk include irregular Internet language usage, keyboard emoticons or some Internet-specific features which do not pose a big difference when it is compared with Twitter. Characteristically Eksi Sozluk is not full of Internet language features such as emoticons while authors have this feature on Twitter. However such type of features comprises a small amount of data when it is compared to all. For that reason, such genre-specific features were not taken to the list. One of the recent research results on cross-genre stylometric analysis is found challenging and applying standard methods to these problems results in misclassification due to the “features that separate the authors in the source domain are distorted in the target domain” (Overdorf and Greenstadt 2016 p.169). Moreover, since the tweets are short, it is required more than 15 texts per author. It should be stated that in cross-genre comparison, the distribution of the texts is not necessarily balanced in terms of the size of the texts because of the various variables in a real-world dataset, i.e. an author may have more texts in one genre and less in the other one. In this test, A8 from Eksi Sozluk is the disputed author and the A9 from Twitter is the actual author. Due to the limited size of the texts only 42 linguistic features were found in this dataset. After applying the Jaccard distance test, mean distances were calculated in order to find the smallest values between authors. Unlike the previous tests presented above, this training test demonstrated

different results. In the distance matrix, there were three potential authors in comparison with the disputed author. It was difficult to decide which one was the candidate author for further analysis; thus it decided to continue from the author vs author comparison to see the distances between authors. The raw Jaccard distances between authors are presented in Table 6-23.

Table 6-23: Distances between authors for Corpus4.

	row.names	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15
1	A01	0.7232427	0.9394718	0.9086319	0.8890361	0.9119488	0.8970115	0.9198519	0.8420504	0.8391422	0.9303651	0.8882274	0.8403284	0.8790804	0.9039757	0.8040339
2	A02	0.9394718	0.9203401	0.9188113	0.9058765	0.9339365	0.8815185	0.9114004	0.9264900	0.9582628	0.9319647	0.9573016	0.9684339	0.9205732	0.9357053	0.9839788
3	A03	0.9086319	0.9188113	0.8766327	0.8572116	0.9335115	0.9268530	0.9034709	0.9279436	0.9681198	0.9495679	0.9470035	0.9543510	0.9409683	0.9551408	0.9683580
4	A04	0.8890361	0.9058765	0.8572116	0.8252608	0.9172011	0.8645305	0.8911993	0.9049678	0.9632379	0.9366171	0.9205009	0.9600194	0.9353175	0.9368814	0.9582663
5	A05	0.9119488	0.9339365	0.9335115	0.9172011	0.9576190	0.9214868	0.9336984	0.9419751	0.9551958	0.9320265	0.9500000	0.9457460	0.9438307	0.9377143	0.9442240
6	A06	0.8970115	0.8815185	0.9268530	0.8645305	0.9214868	0.7694331	0.9194903	0.8873449	0.9610770	0.9227669	0.9208413	0.9664002	0.8744954	0.8868559	0.9665658
7	A07	0.9198519	0.9114004	0.9034709	0.8911993	0.9336984	0.9194903	0.9108390	0.9066965	0.9373263	0.9480141	0.9436085	0.9487866	0.9192998	0.9534339	0.9566120
8	A08	0.8420504	0.9264900	0.9279436	0.9049678	0.9419751	0.8873449	0.9066965	0.7298839	0.8483561	0.9451417	0.8842976	0.8944228	0.8713094	0.8763570	0.8488044
9	A09	0.8391422	0.9582628	0.9681198	0.9632379	0.9551958	0.9610770	0.9373263	0.8483561	0.8349206	0.9614602	0.8978342	0.8859295	0.8830723	0.9175799	0.8379750
10	A10	0.9303651	0.9319647	0.9495679	0.9366171	0.9320265	0.9227669	0.9480141	0.9451417	0.9614602	0.8982993	0.9545485	0.9528695	0.9397443	0.9367279	0.9644390
11	A11	0.8882274	0.9573016	0.9470035	0.9205009	0.9500000	0.9208413	0.9436085	0.8842976	0.8978342	0.9545485	0.8865533	0.9126561	0.9047725	0.9340247	0.8866208
12	A12	0.8403284	0.9684339	0.9543510	0.9600194	0.9457460	0.9664002	0.9487866	0.8944228	0.8859295	0.9528695	0.9126561	0.8542819	0.8968889	0.9232944	0.8104356
13	A13	0.8790804	0.9205732	0.9409683	0.9353175	0.9438307	0.8744954	0.9192998	0.8713094	0.8830723	0.9397443	0.9047725	0.8968889	0.8343953	0.8988483	0.8855961
14	A14	0.9039757	0.9357053	0.9551408	0.9368814	0.9377143	0.8868559	0.9534339	0.8763570	0.9175799	0.9367279	0.9340247	0.9232944	0.8988483	0.8857998	0.9016575
15	A15	0.8040339	0.9839788	0.9683580	0.9582663	0.9442240	0.9665658	0.9566120	0.8488044	0.8379750	0.9644390	0.8866208	0.8104356	0.8855961	0.9016575	0.7536508

In a similar manner to the text vs text comparison, this distance matrix is demonstrated that there are meaningful similarities between A8 and A1, A8 and A9 and A1 and A15. Even though, A8 and A9 pairing correspond the disputed and the actual author for this analysis, under these conditions it is not possible to attribute the authorship correctly. In order to examine the distinctive linguistic similarities between A8 and A1, A9 and A15, shared features are presented in Table 6-24.

Table 6-24: Shared features between A8 and A1, A9, A15

FEATURES	AUTHOR8	AUTHOR1	AUTHOR9	AUTHOR15
F1 Lex Intj. Ya	2	1	3	3
F2 Lex bi	3	0	3	2
F4 Lex falan	0	4	0	0
F5 Lex filan	3	1	2	1
F7 Lex Keyboard Emoticon	12	0	2	2
F8 Lex Numbers (Numeral)	6	6	0	1
F9 Lex Omission Rclip	0	4	0	2
F10 Lex Prosodic Emph.	0	2	0	2
F11 Lex RedupSameform	1	4	0	3
F12 Lex sey	0	0	0	1
F14 Str edit	0	0	0	0
F15 Str Itemise	0	0	0	0
F17 Syn ConConj ama	9	1	0	1

F22 Syn ConConj zaten	0	0	1	1
F24 Syn EndConj ile	1	4	1	0
F26 Syn MultiTypoPunc. Excl.	0	0	1	0
F28 Syn NonStdPunc. DoubleFullstop	3	1	0	0
F30 Syn Parenth.Clause	0	0	1	1
F31 Syn PuncAbs Apost.	4	0	1	1
F32 Syn PuncAbs. Fullstop	12	8	12	0
F33 Syn Punc. AbscSpace	0	1	0	0
F34 Syn Punc Slash	0	1	0	7
F35 Syn SenQuotation	3	15	0	14
F36 Syn StndPunc. Colon	0	9	2	6
F39 Syn StndPunc. QuestionMark	2	1	4	0
F42 Syn StrConj cunku	0	0	1	1

In Table 6-24, the features shared with A8 is filled with yellow colour and when it is not shared, it is left blank. As it is seen in the table, 42 combined linguistic features were used by the disputed author, 17 of them is shared with A1 while 16 of them is common with A9. On the other side, A15 has shared 14 features with A8. In this case, it is difficult for the correct author while there are three potential authors with similar distances. The raw Jaccard distance values are so similar to each other in Table 6-23. The distances are 0.8420 for A8-A1, 0,8483 for A8-A9 and 0.8488 for A8 and A15, the number of shared linguistic features are supporting the minimal differences between the values.

Overall, the cross-genre comparison with 42 combined linguistic features did not attribute the authorship correctly. It can be explained with the unbalanced text sizes in the corpus since a standard Twitter message can contain a maximum of 280 characters. However, this method is reduced the number of candidates from 14 to 3, for that reason it needs a further analysis with a developed comprehensive linguistic feature and a balanced dataset from both genres. It is evident that there are more tweets needed for further analysis.

Chapter 7: Conclusion

In recent years even though there are many studies in authorship attribution, most studies, especially in Turkish are focused on only computational analysis (e.g. Amasyali and Diri, 2006) without considering linguistics. Hence, it is impossible to have a reliable output only focusing on computer-based analysis regardless of linguistic analysis. In contrast, the traditional stylistic analysis includes more intuitive knowledge. Stylistic analysis is open to criticism regarding reliability, and for that reason developing techniques which combine stylistic and stylometric approaches (e.g. Grant, 2013; Wright, 2014) is more valid than the traditional authorship attribution methods.

As a result, this study was organised to propose various approaches to Turkish authorship attribution studies which combine stylistic and stylometric approaches. It is carried out various experiments and compared the performance of the results in the simulated scenario with Eksi Sozluk data. Although explaining the main approaches would be enough to explain the research aims, additional tests run in order to gain more insight regarding authorship attribution in Turkish. The summary of the results is presented in the following sections.

7.1. Summary of the Results

Authorship attribution aims to find the disputed author of a text based on the texts from the known authors. For authorship attribution in Turkish, some studies are done from the beginning of the 21st century although the field has already been popular in the rest of the world since the 60s. The studies in Turkish are entirely based on stylometric approaches and non-realistic cases with a small set of authors, big size data, literary texts or features which are not explainable linguistically. Most of these studies are found their method worked on a few authors with big data size and returned between 90-95% accuracy rate. However, this assumption becomes vague when the method is faced with an actual forensic case. A method which includes forensic linguistics authorship analysis was presented in this study.

The methodology proposed by Grant (2013) used and expanded in this study which was initially developed for short text messages. However, Grant (2013) had only two possible authors for his study, unlike this research. Considering the number of possible authors and cross-genre application this data set was differentiated from his data set. Moreover, Grant (2013) focused on the features which were predominantly used by one author; however, in this study, all authors and features treated equally. In terms of feature selection, three ways of decision method were followed. First, the features used for SMS texts and micro-blogging texts

(MacLeod and Grant 2012; Grant 2013), second, the internet language features and the finally data-driven features were collected with both bottom-up and top-down approaches.

Moreover, considering the success of word n-grams in previous studies (e.g. Wright 2014; Nini 2018), word strings used to identify the authors however this was not the centre of the research for that reason other features were considered along with such features. A multilevel coding system was used for each corpus. In the end, each feature was converted into presence and absence values before applying the statistical tests.

The statistical design was based on Grant's (2013) approach in scoring the distance values between authors while using presence and absence values rather than frequencies. In the previous Turkish authorship attribution studies using frequency values did not cause to excellent results, yet none of them used computer-mediated medium data which is relatively short than the other data types. However, it is indicated that binary feature input was also a remarkable aspect in authorship attribution studies in Turkish. A number of experiments were conducted to examine the impact of different parameters in authorship attribution in Turkish including the candidate author size and feature type.

In the first approach, the lexical features showed the best performance while syntactic features did not show a success. When the features are divided into the sub-categories, it is found that the performance drops whilst combining feature sets lead to significant improvements. Even so, combining all the features regardless of their function improves the reliability of the results rather than depending on the single feature set usage. In the second approach, size is tested in various conditions including text size, the number of candidate authors size, the number of texts per author. This is certainly a favourable conclusion with the size factors tested in this research. Since these approaches had good performance in results, thus it is possible to apply the same method to different text types regardless of their text size. Assuming that small texts are the most challenging text group when compared with longer ones since it produced promising results for the current research. It is essential to note that results are not pointing out longer texts are better, because when it is said *long texts* in the previous studies, it is understood as at least 1000 words long texts. However, within this research context, *long texts* are the ones with maximum of 550 words. Besides, it is found that increasing the number of authors from 15 to 30 made a slight effect the results within the scope of this research.

These results are showed that the proposed approaches in feature selection, coding and analysing the distances between texts propose satisfactory results in Turkish. Fundamentally,

contrary to the previous studies in Turkish, correct classification is possible even only with lexical features and word n-grams rather than combining all features elicited from the texts. Eventually, the results are presented in two ways as text vs text comparison by indicating the smallest mean Jaccard distances between the texts and author vs author comparison that is presented raw Jaccard scores in tables and heatmaps which showed the accuracy of classification in different settings. The texts that were hit the lowest at least five times were selected for further investigation. The author of these texts was selected as a candidate author, and the shared features between the author and the disputed author were investigated. Observing such features is needed to establish the consistency and the distinctiveness in attributing authorship. After this method, the author vs author comparison was applied, and the raw Jaccard scores were presented to demonstrate the distances between the authors. 7 out of ten training tests demonstrated good results in general which is a good a good indication of the success of the current method in authorship attribution in Turkish. However, in all training tests, there is still considerable space for developing the most successful authorship attribution results in forensic linguistics context. However, these results have reliable evidence standards to present in a courtroom when considering the most successful results (see Chapter 6) under the conditions of Daubert criteria.

According to Daubert criteria having an error rate is an expected output of any method is applied in presenting to the court in the USA, although the situation is different in other countries, researchers aim to follow Daubert criteria in order to establish a standard in forensic linguistics evidence. Although Turkish courts do not require a known error rate in the expert testimony, the scientific criteria established for Daubert can help the acceptable of authorship attribution results in criminal investigations. As it is mentioned above, to the researcher best knowledge although there are cases related to forensic linguistics, there is no forensic linguist to apply reliable and valid approaches.

In regards to the reliability of the approaches which were applied in this research, the corpus linguistics methods are used to extract the important features and word n-grams from the data set, later on, a semi-automated coding system applied. Moreover, it also included a number of data-driven features in that case in order to avoid any subjectivity, and an intercoder reliability test can be applied for this. However, it is still necessary to decrease the level of the manual coding process. Additionally, the statistical part of the study depends on automated processes that can be replicated in future studies. Also, the statistical part is based on Grant's (2013) method and achieved the same results.

Furthermore, mean distance values between texts, raw Jaccard scores between authors, shared features among the candidate authors and finally, heatmaps revealed a clue to the similarity of the texts. This study proposed that visualisation method has significant importance in classifying the authors and explaining to the non-experts via heatmaps such as in a courtroom setting. Achieving correct attribution with such a method in another language apart from English showed that this method would be appropriate for the other languages.

7.2. Revisiting Research Questions

A comprehensive analysis of the authorship attribution from different aspects such as the performance of feature type, the role of text size and candidate size, and finally the effects of cross-genre applications are made in Turkish texts. Three research questions were asked at the beginning of the research, and a list of training tests has given the answers to each along with several contributions to the literature especially to Turkish authorship attribution studies.
In line with the discussion above, it is necessary to revisit the research questions in Chapter 1.4.:

1) What is the role of feature type in authorship attribution research in Turkish texts?
    i) Which feature set achieves the accurate attribution of authorship? Is it possible to increase authorship attribution performance by selecting the appropriate feature set or combinations in the current data set?

After designing three feature sets and ten different authorship attribution simulated cases, the feature sets showed good results in some cases; however, in the small size texts, the discriminatory potential of these features decreased. There is two possible reasons in this; one there is not many linguistic features in the texts since the structure of the text does not allow, second the features selected for this study has potential only in the texts which are longer than 300 words.

Lexical features were provided best results in this study while the syntactical feature which included a number of grammatical structure features had a lower rate than this. Finally, structural features did not provide separate results due to its limited existence in the data set. However, when all feature types combined in the rest of the training tests, the accuracy is

improved in the corpora used. This indicates that rather than dividing the features into the sections it is indeed necessary to combine features to improve the accuracy rate

2) Does the text size and candidate author size affect the attribution of authorship correctly in Turkish texts?

    i. What is the effect of text size in authorship attribution?
    ii. What is the effect of candidate author set size in authorship attribution?
    iii. How many texts are needed in assigning the disputed texts to the correct author?

The effect of text size was tested with three corpora. Although Corpus1, Corpus2 and Corpus3 had performed better with, Corpus4 (43 words on average) misattributed. Thus, it can be concluded that decreasing the text size leads a poor performance in authorship attribution in Turkish. Increasing the number of candidate authors to thirty authors were provided good results by looking at the linguistic connection between the disputed author and the potential author even though there were 900 texts in comparison. Finally, in parallel to real-world cases, five texts and ten texts taken from Corpus1 and tested the applicability of the method. As a result, both tests performed well and led to correct attribution between disputed and the corresponding author.

3) To what extent can the authorship attribution method be applied in cross-genre comparison?

Since the data comparison data is collected from the microblogging web site, the size of the texts was not comparable with Eksi Sozluk data. The number of texts was balanced, but the corpus was not balanced in terms of size. Therefore, the test was not performed well.

7.3. Limitations

In Eksi Sozluk generally, authors do not share passwords and usernames with their family or friends. Although it was asked before than the data collection period, there was still a possibility of a joint account between data sets. Although the data belongs to the linguistically and demographically diverse web site, there was still a high level of similarities between unrelated

author pairs. Moreover, in order to balance the cross-genre data with Eksi Sozluk, only 15 texts collected per author from Twitter, however, some Twitter messages had a few words; therefore, the accuracy rate was dropped in this test.

7.4. Future Studies

In this study, an authorship attribution example in Turkish texts particularly online texts is proposed to evaluate the efficiency under the different conditions. The results showed that this method could assign the correct authors in long size and medium size texts. However, short text size which is less than 50 words has always been problematic for authorship studies. When the performance of the lexical features and the failure of the short texts are considered together, for the future studies it should be a focus on more lexical features especially word n-grams.

With these initial results, future research will include larger datasets from a different genre with various linguistic features and variables to work on it. Although this study runs a cross-genre analysis between Twitter and Eksi Sozluk, the amount of data was not sufficient for that reason it is needed to improve the study from these aspects.

References

Aarts, J. and Meijs, W. (1990). *Theory and Practice in Corpus Linguistics*. Amsterdam: Rodopi.

Abbasi, A. and Chen, H. (2005). Applying Authorship Analysis to Extremist-Group Web Forum Messages. *IEEE Intelligent Systems*, 20(5), pp.67-75.

Abbasi, A. and Chen, H. (2008). Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection. *ACM Transactions on Information Systems*, [online] 26(2), pp.1-29. Available at: https://dl.acm.org/citation.cfm?doid=1344411.1344413 [Accessed 16 Apr. 2018].

Agun, H., Yilmazel, S. and Yilmazel, O. (2017). Effects of language processing in Turkish authorship attribution. *2017 IEEE International Conference on Big Data (Big Data)*, [online] pp.1876-1881. Available at: http://doi.org/10.1109/BigData.2017.8258132 [Accessed 6 May 2018].

Akkoyunlu, B. and Soylu, M. (2011). Sosyal İletişim Ağları ve Dilin Yanlış Kullanımı Üzerine Nitel Bir Çalışma. *İlköğretim Online*, 10(2), pp.441-453.

Aksut, M., Batur, Z. and Avsar, T. (2006). Sanalca, Sanal Odalarda (İnternet) İletişim ve Türkçe. In: *Akademik Bilişim Konferansı*. [online] Pamukkale University. Available at: https://ab.org.tr/ab06/bildiri/23.doc [Accessed 14 Dec. 2018].

Alexa (2018). *Keyword Research, Competitor Analysis, & Website Ranking*. [online] Alexa.com. Available at: https://www.alexa.com/ [Accessed 7 Jan. 2018].

Amasyalı, M. and Diri, B. (2006). Automatic Turkish Text Categorization in Terms of Author, Genre and Gender. In: *Kop C., Fliedl G., Mayr H.C., Métais E. (eds) Natural Language Processing and Information Systems. NLDB 2006. Lecture Notes in Computer Science, vol 3999*. [online] Berlin, Heidelberg: Springer, pp.221-226. Available at: https://doi.org/10.1007/11765448_22 [Accessed 8 Feb. 2017].

Androutsopoulos, J. (2006). Introduction: Sociolinguistics and Computer-mediated Communication. *Journal of Sociolinguistics*, 10(4), pp.419-438.

Argamon, S. and Koppel, M. (2013). A Systemic Functional Approach to Automated Authorship Analysis. *Journal of Law and Policy*, 2(21), pp.299-316.

Argamon, S. and Levitan, S. (2005). Measuring The Usefulness of Function Words for Authorship Attribution. In: *Proceedings of ACH/ALLC Conference*. [online] University of Victoria, BC,: Association for Computing and the Humanities, pp.1-3. Available at: https://pdfs.semanticscholar.org/1b70/57378e2a300cde88e6f291e146981d338a63.pdf [Accessed 6 Jan. 2018].

Argamon, S., Koppel, M., Fine, J. and Shimoni, A. (2003). Gender, Genre, and Writing Style in Formal Written Texts. *Text - Interdisciplinary Journal for the Study of Discourse*, [online] 23(3), pp.321-346. Available at: https://doi.org/10.1515/text.2003.014 [Accessed 12 Oct. 2017].

Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. (2008). Automatically Profiling the Author of an Anonymous Text. *Communications of the ACM*, [online] 52(2), pp.119-123. Available at: http://doi.org/10.1145/1461928.1461959 [Accessed 8 Mar. 2018].

Argamon, S., Šarić, M. and Stein, S. (2003). Style Mining of Electronic Messages for Multiple Authorship Discrimination. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*. [online] New York, NY: ACM, pp.475-480. Available at: http://doi.org/10.1145/956750.956805 [Accessed 25 Feb. 2018].

Argamon, S., Whitelaw, C., Chase, P., Hota, S., Garg, N. and Levitan, S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, [online] 58(6), pp.802-822. Available at: https://doi.org/10.1002/asi.20553 [Accessed 8 Mar. 2018].

Aslan, C. (2007). Content Analysis On Language Mistakes Made By Turkish, Turkish Language and Literature Teachers in Internet. In: H. Uzunboylu and N. Çavuş, ed., *7. International Educational Technology Conference*. Near East University, pp.90-98.

Baayen, H., van Haltere, H., Neijt, A. and Tweedie, F. (2002). An Experiment in Authorship Attribution. In: *Proceedings of JADT 2002: Sixth International Conference on Textual Data Statistical Analysis*. [online] pp.29-37. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.6139&rep=rep1&type=pdf [Accessed 23 Feb. 2017].

Baayen, H., van Halteren, H. and Tweedie, F. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, [online] 11(3), pp.121-132. Available at: https://doi.org/10.1093/llc/11.3.121 [Accessed 7 Mar. 2018].

Babbie, E. (1989). *The Practice of Social Eesearch*. 5th ed. Belmont, CA: Wadsworth Publishing Company.

Baber, A. (2004). Idiolects. In: E. Zalta, ed., *Stanford Encyclopedia of Philosophy*. [online] Palo Alto, CA: CSLI, University of Stanford. Available at: http://plato.stanford.edu/entries/idiolects/ [Accessed 14 Jan. 2018].

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Barlow, M. (2010). Individual usage: a corpus-based study of idiolects. Paper presented at the 34th International LAUD Symposium, Landau, Germany. [online] Available at: http://michaelbarlow.com/barlowLAUD.pdf [Accessed 16 Oct. 2017].

Baron, N. (2003). Language and the Internet. In: A. Farghaly, ed., *The Stanford Handbook for Language Engineers*. Stanford, CA: CSLI, pp.59-127.

Barthes, R. (1977). The Death of the Author. In: *Image, Music, Text: Essays Selected and Translated by Stephen Heath*. New York: Hill and Wang, pp.142–148.

Barton, D. and Lee, C. (2013). *Language Online: Investigating Digital Texts and Practices*. Milton Park, Abingdon, Oxon: Routledge.

Bay, Y. and Çelebi, E. (2016). Feature Selection for Enhanced Author Identification of Turkish Text. In: O. Abdelrahman, E. Gelenbe, G. Gorbil and R. Lent, ed., *Information Sciences and Systems 2015. Lecture Notes in Electrical Engineering, vol 363*. [online] Cham: Springer, pp.371-379. Available at: https://doi.org/10.1007/978-3-319-22635-4_34 [Accessed 14 May 2018].

Baym, N. (2007). The new shape of online community: The example of Swedish independent music fandom. *First Monday*, [online] 12(8). Available at: https://doi.org/10.5210/fm.v12i8.1978 [Accessed 11 Jan. 2018].

Becker, A. (1984). Toward A Post-Structuralist View of Language Learning: A Short Essay. *Language Learning*, 33(5), pp.217-220.

Bell, M. (2007). *The Transformation of The Encyclopedia: A Textual Analysis and Comparison of The Encyclopaedia Britannica and Wikipedia*. Master's thesis. Ball State University.

Beyond Microblogging: Conversation and Collaboration via Twitter. (2009). *2009 42nd Hawaii International Conference on System Sciences*.

Bhargava, M., Mehndiratta, P. and Asawa, K. (2013). Stylometric Analysis for Authorship Attribution on Twitter. In: *Big Data Analytics. BDA 2013. Lecture Notes in Computer Science, vol 8302*. [online] Cham: Springer, pp.37-47. Available at: https://doi.org/10.1007/978-3-319-03689-2_3 [Accessed 14 May 2018].

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

Blanchard, A. (2004). Blogs as Virtual Communities: Identifying a Sense of Community in the Julie/Julia Project. *Into the Blogosphere Articles*, [online] Retrieved from the University of Minnesota Digital Conservancy. Available at: http://hdl.handle.net/11299/172837 [Accessed 10 Apr. 2017].

Bloch, B. (1948). A Set of Postulates for Phonemic Analysis. *Language*, 24(1), pp.3-46.

Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart, and Winston.

Boukhaled, M., Frontini, F., Bourgne, G. and Ganascia, J. (2015). Computational Study of Stylistics: A Clustering-based Interestingness Measure for Extracting Relevant

Syntactic Patterns. *International Journal of Computational Linguistics and Applications*, 6(1), pp.45–62.

Boutwell, S. (2011). *Authorship Attribution of Short Messages Using Multimodal Features*. Master's Thesis. Naval Postgraduate School.

boyd, d., Golder, S. and Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In: *43rd Hawaii International Conference on System Sciences*. pp.1-10.

Bozkurt, I., Baglioglu, O. and Uyar, E. (2007). Authorship Attribution: Performance of Various Features and Classification Methods. *22nd International Symposium on Computer and Information Sciences, ISCIS 2007 - Proceedings*. [online] Available at: http://repository.bilkent.edu.tr/handle/11693/27016 [Accessed 23 Feb. 2018].

Britannica.com. (2014). *Encyclopedia Britannica*. [online] Available at: https://www.britannica.com/ [Accessed 13 Jan. 2017].

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3), pp.267-287.

Cakir, H. and Topcu, H. (2006). Bir İletişim Dili Olarak İnternet. *Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 19(2), pp.71-96.

Can, F. and Patton, J. (2004). Change of Writing Style with Time. *Computers and the Humanities*, 38(1), pp.61-82.

Chaski, C. (2001). Empirical Evaluations of Language-based Author Identification Techniques. *Forensic Linguistics*, [online] 8(1), pp.1-65. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary;jsessionid=C31C2DA74445ECBDC779719E20AF5359?doi=10.1.1.465.5651 [Accessed 2 Jan. 2018].

Chaski, C. (2005). Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence*, 1(4), pp.1-14.

Chaski, C. (2007). The Keyboard Dilemma and Authorship Identification. In: P. Craiger and S. Shenoi, ed., *Advances in Digital Forensics III*. New York, NY: Springer, pp.133-146.

Chen, X., Hao, P., Chandramouli, R. and Subbalakshmi, K. (2011). Authorship Similarity Detection from Email Messages. In: P. Perner, ed., *Machine Learning and Data Mining (MLDM) in Pattern Recognition*. [online] Berlin, Heidelberg: Springer, pp.375-386. Available at: https://doi.org/10.1007/978-3-642-23199-5_28 [Accessed 7 Mar. 2018].

Cheng, E. (2013). Being Pragmatic About Forensic Linguistics. *Journal of Law and Policy*, 21(2), pp.541-550.

Coleman, S. (2006). E-mail, Terrorism, and the Right to Privacy. *Ethics and Information Technology*, [online] 8(1), pp.17-27. Available at: https://link.springer.com/article/10.1007%2Fs10676-006-9103-5 [Accessed 3 Feb. 2018].

Cotterill, J. (2010). How to use corpus linguistics in Forensic Linguistics?. In: A. O'Keeffe and M. McCarthy, ed., *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp.578–590.

Coulthard, M. (1994). On the use of Corpora in the Analysis of Forensic Texts. *International Journal of Speech Language and the Law*, [online] 1(1), pp.27-43. Available at: https://journals.equinoxpub.com/index.php/IJSLL/article/view/16584 [Accessed 16 Feb. 2018].

Coulthard, M. (2004). Author Identification, Idiolect, and Linguistic Uniqueness. *Applied Linguistics*, [online] 25(4), pp.431-447. Available at: https://academic.oup.com/applij/article/25/4/431/193364 [Accessed 29 Apr. 2018].

Coulthard, M. (2005). Some Forensic Applications of Descriptive Linguistics. *VEREDAS - Rev. Est. Ling. Juiz de Fora*, [online] 9(1), pp.9-28. Available at: http://www.ufjf.br/revistaveredas/files/2009/12/artigo016.pdf [Accessed 8 Mar. 2018].

Coulthard, M. (2013). On Admissible Linguistic Evidence. *Journal of Law and Policy*, 21(2), pp.441–466.

Coulthard, M., Johnson, A. and Wright, D. (2017). *An introduction to forensic linguistics*. Abingdon, Oxon: Routledge.

Coulthard, M., Johnson, A. and Wright, D. (2017). *An Introduction to Forensic Linguistics*. Abingdon, Oxon: Routledge.

Cresswell, M. (2003). *Heaps, Prototypes and Ethics: The Consequences of Using Judgements of Student Performance to Set Examination Standards in a Time of Change*. London: Institute of Education, University of London.

Creswell, J. and Plano Clark, V. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: SAGE Publications.

Crystal, D. (2007). *How Language Works*. London: Penguin Books.

Crystal, D. (2008). *Txtng : The gr8 db8*. Oxford: Oxford University Press.

Crystal, D. (2011). *Internet Linguistics: A Student Guide*. London: Routledge.

Danet, B. (2001). *Cyberpl@y: Communicating Online*. Oxford: Berg.

de Vel, O., Anderson, A., Corney, M. and Mohay, G. (2001). Mining e-mail Content for Author Identification Forensics. *ACM SIGMOD Record*, [online] 30(4), pp.55-64. Available at: https://dl.acm.org/citation.cfm?doid=604264.604272 [Accessed 21 Jan. 2018].

Denzin, N. (1989). *Interpretive Interactionism*. Newbury Park: Sage.

Diederich, J. (2003). Authorship Attribution with Support Vector Machines. *Applied Intelligence*, 19(1/2), pp.109-123.

Diri, B. and Amasyali, M. (2003). Automatic Author Detection for Turkish Texts. In: *Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP)*. [online] pp.1-4. Available at: https://pdfs.semanticscholar.org/f142/2461024fcec79c94fe2671923ce79be0e4ef.pdf [Accessed 19 Nov. 2017].

Dogu, B., Ziraman, Z. and Ziraman, D. (2009). Web Based Authorship in the Context of User Generated Content, An Analysis of a Turkish Web Site: Eksi Sozluk. In: D. Riha and A. Maj, ed., *The Real and the Virtual*. Oxford: Inter-Disciplinary Press, pp.119-128.

Donath, J. and boyd, d. (2004). Public displays of connection. *BT Technology Journal*, 22(4), pp.71-82.

Dresner, E. and Herring, S. (2010). Functions of the Nonverbal in CMC: Emoticons and Illocutionary Force. *Communication Theory*, 20(3), pp.249-268.

Ebner, M., Lienhardt, C., Rohs, M. and Meyer, I. (2010). Microblogs in Higher Education – A chance to facilitate informal and process-oriented learning?. *Computers & Education*, 55(1), pp.92-100.

Eden, S. and Heiman, T. (2011). Computer Mediated Communication: Social Support for Students with and without Learning Disabilities. *Educational Technology & Society*, [online] 14(2), pp.89–97. Available at: https://www.semanticscholar.org/ [Accessed 12 Feb. 2018].

Eder, M. (2013). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, [online] 30(2), pp.167-182. Available at: https://doi.org/10.1093/llc/fqt066 [Accessed 7 Mar. 2018].

Eisenmann, M., O'Neil, J. and Geddes, D. (2013). Testing the Reliability of Metrics Proposed as Standards for Traditional Media Analysis. In: *Proceedings from the 16th Annual InternationalPublic Relations Research Conference*. [online] Available at: http://kdpaine.blogs.com/files/eisenmann-and-oneal_reliability0001.pdf [Accessed 10 Jan. 2018].

Ekinci, E. and Takci, H. (2012). Using Authorship Analysis Techniques in Forensic Analysis of Electronic Mails. *2012 20th Signal Processing and Communications Applications Conference (SIU 2012)*, pp.543-546.

ekşi sözlük. (2018). *ekşi sözlük - kutsal bilgi kaynağı*. [online] Available at: https://eksisozluk.com/ [Accessed 15 Jan. 2018].

Emigh, W. and Herring, S. (2005). Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias. In: *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. [online] IEEE, pp.1-11. Available at: http://doi.org/10.1109/HICSS.2005.149 [Accessed 12 Mar. 2017].

Federal Committee (2017). *Rule 702. Testimony by Expert Witnesses | Federal Rules of Evidence | LII / Legal Information Institute*. [online] Federal Committee. Available at: https://www.law.cornell.edu/rules/fre/rule_702 [Accessed 23 Mar. 2017].

Fisher, B. and Fisher, D. (2012). *Techniques of Crime Scene Investigation*. Boca Raton, Fla.: CRC Press.

Fitzgerald, J. (2004). Using a Forensic Linguistic Approach to track the Unabomber. In: J. Campbell and D. Denivi, ed., *Profilers: Leading Investigators take you Inside the Criminal Mind*. New York: Prometheus Books, pp.193–222.

Fletcher, W. (2012). Corpus Analysis of the World Wide Web. *The Encyclopedia of Applied Linguistics*. [online] Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal0254 [Accessed 8 Mar. 2017].

Forsyth, R. and Holmes, D. (1996). Feature-finding for text classification. *Digital Scholarship in the Humanities*, 11(4), pp.163-174.

Foster, D. (2001). *On the Trail of Anonymous*. New York: Henry Holt & Company.

Freelon, D. (2013). ReCal OIR: Ordinal, Interval, and Ratio Intercoder Reliability as a Web Service. *International Journal of Internet Science*, 1(8), pp.10-16.

Gamon, M. (2004). Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features. In: *COLING '04 Proceedings of the 20th international conference on Computational Linguistics*. [online] Stroudsburg, PA: Association for Computational Linguistics, pp.611–617. Available at: http://doi.org/10.3115/1220355.1220443 [Accessed 17 Feb. 2018].

Gilquin, G. (2010). *Corpus, Cognition and Causative Constructions*. Amsterdam: Benjamins.

Gilquin, G. and Gries, S. (2009). Corpora and Experimental Methods: A State-of-the-art Review. *Corpus Linguistics and Linguistic Theory*, [online] 5(1), pp.1-26. Available at: https://www.degruyter.com/view/j/cllt.2009.5.issue-1/cllt.2009.001/cllt.2009.001.xml [Accessed 13 Jan. 2018].

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J. and Smith, N. (2011). Part-of-speech tagging for Twitter:

Annotation, features, and experiments. In: *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*. [online] Portland, USA: Association for Computational Linguistics, pp.42–47. Available at: http://www.aclweb.org/anthology/P11-2008 [Accessed 10 Jan. 2018].

Goffman, E. (1981). *Forms of Talk*. Philadelphia: University of Pennsylvania Press.

Göksel, A. and Kerslake, C. (2005). *Turkish: A Comprehensive Grammar*. London: Routledge.

Goldstein-Stewart, J., Winder, R. and Sabin, R. (2009). Person Identification from Text and Speech Genre Samples. In: *Proceedings of the 12th Conference of the European Chapter of the ACL*. [online] Athens, Greece: Association for Computational Linguistics, pp.336-344. Available at: http://www.aclweb.org/anthology/E09-1039 [Accessed 7 Feb. 2018].

Grant, T. (2004). *Authorship Attribution in a Forensic Context*. PhD thesis. University of Birmingham.

Grant, T. (2007). Quantifying Evidence in Forensic Authorship Analysis. *International Journal of Speech Language and the Law*, [online] 14(1), pp.1-25. Available at: https://journals.equinoxpub.com/index.php/IJSLL/article/view/3955 [Accessed 17 Apr. 2016].

Grant, T. (2008). Approaching Questions in Forensic Authorship Analysis. In: J. Gibbons and M. Turell, ed., *Dimensions of Forensic Linguistics*. Amsterdam: John Benjamins, pp.215–229.

Grant, T. (2010). Txt 4n6: Idiolect Free Authorship Analysis?. In: M. Coulthard and A. Johnson, ed., *The Routledge Handbook of Forensic Linguistics*. London: Routledge, pp.508–522.

Grant, T. (2013). Txt 4N6: Method, Consistency and Distinctiveness in the Analysis of SMS Text Messages. *Journal of Law and Policy*, 2(21), pp.467–494.

Grant, T. and Baker, K. (2001). Identifying Reliable, Valid Markers of Authorship: A Response to Chaski. *Forensic Linguistics*, [online] 8(1), pp.66-79. Available at: https://journals.equinoxpub.com/index.php/IJSLL/article/view/1690 [Accessed 15 Jan. 2016].

Grant, T. and Macleod, N. (2016). Assuming Identities Online: Experimental Linguistics Applied to the Policing of Online Paedophile Activity. *Applied Linguistics*, 37(1), pp.50-70.

Grant, T. and Nini, A. (2013). Bridging the Gap between Stylistic and Cognitive Approaches to Authorship Analysis Using Systemic Functional Linguistics and Multidimensional Analysis. *International Journal of Speech Language and the Law*, [online] 20(2). Available at: https://journals.equinoxpub.com/index.php/IJSLL/article/view/13599 [Accessed 2 Feb. 2017].

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, [online] 22(3), pp.251-270. Available at: https://academic.oup.com/dsh/article/22/3/251/951481 [Accessed 3 Feb. 2018].

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M. and Matthiessen, C. (2004). *An Introduction to Functional Grammar*. 3rd ed. London: Routledge.

Hänlein, H. (1999). *Studies in Authorship Recognition: A Corpus-based Approach (Vol. 352)*. Frankfurt am Main: Peter Lang.

Haug, M. and Baird, E. (2011). Finding the Error in Daubert. *Hastings Law Journal*, 62(3), pp.737-756.

Haylock, C. and Muscarella, L. (1999). *Net Success: 24 Leaders in Web Commerce Show You How to Put the Internet to Work for Your Business*. Holbrook, Mass.: Adams Media Corporation.

Herring, S. (1996). Linguistic and Critical Analysis of Computer-Mediated Communication: Some Ethical and Scholarly Considerations. *The Information Society*, [online] 12(2), pp.153-168. Available at: https://doi.org/10.1080/911232343 [Accessed 16 Feb. 2018].

Herring, S. (2001). Computer-Mediated Discourse. In: D. Schiffrin, D. Tannen and H. Hamilton, ed., *The Handbook of Discourse Analysis*. Malden, MA: Blackwell Publishers Ltd, pp.612-634.

Herring, S. (2004). Computer-mediated Discourse Analysis: An Approach to Researching Online Communities. In: S. Barab, R. Kling and J. Gray, ed., *Designing for Virtual Communities in the Service of Learning*. New York: Cambridge University Press, pp.338-376.

Herring, S. (2004). Computer-Mediated Discourse Analysis. In: S. Barab, R. Kling and J. Gray, ed., *Designing for Virtual Communities in the Service of Learning*. Cambridge: Cambridge University Press, pp.338-376.

Herring, S. (2004). Slouching Toward the Ordinary: Current Trends in Computer-Mediated Communication. *New Media & Society*, [online] 6(1), pp.26-36. Available at: http://journals.sagepub.com/doi/10.1177/1461444804039906 [Accessed 10 Mar. 2018].

Herring, S. (2004). Slouching Toward the Ordinary: Current Trends in Computer-Mediated Communication. *New Media & Society*, 6(1), pp.26-36.

Herring, S. (2007). A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet*, 4(2007), pp.1-37.

Herring, S., Johnson, D. and DiBenedetto, T. (1992). Participation in Electronic Discourse in A "Feminist" Field. In: *Locating Power: Proceedings of the Second Berkeley Women and Language Conference*. pp.250-262.

Herring, S., Scheidt, L., Bonus, S. and Wright, E. (2018). Bridging The Gap: A Genre Analysis of Weblogs. In: *37th Annual Hawaii International Conference on System Sciences*. [online] IEEE. Available at: http://doi.org/10.1109/HICSS.2004.1265271 [Accessed 15 Feb. 2018].

Hillery, G. (1955). Definitions of Community: Areas of Agreement. *Rural Sociology*, 20(2), pp.111-123.

Hirst, G. and Feiguina, O. (2007). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4), pp.405-417.

Holmes, D. (1994). Authorship Attribution. *Computers and the Humanities*, 28(2), pp.87-106.

Honeycutt, C. and Herring, S. (2009). Beyond Microblogging: Conversation and Collaboration via Twitter. In: *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42)*. [online] IEEE Press, pp.1-10. Available at: http://doi.org/10.1109/HICSS.2009.89 [Accessed 8 May 2018].

Hoover, D. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4), pp.453-475.

Hornby, A. (2005). *Oxford Advanced Learner's Dictionary*. Oxford: Oxford University Press.

Howald, B. (2009). Authorship Attribution under the Rules of Evidence: Empirical Approaches in a Layperson's Legal System. *International Journal of Speech Language and the Law*, 15(2), pp.219-247.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hyland, K., Chau, M. and Handford, M. (2012). *Corpus Applications in Applied Linguistics*. 1st ed. London: Bloomsbury.

Iqbal, F., Binsalleeh, H., Fung, B. and Debbabi, M. (2010). Mining Writeprints from Anonymous e-mails for Forensic Investigation. *Digital Investigation*, 7(1-2), pp.56-64.

Jabbar, M. (2010). Overcoming Daubert's Shortcomings in Criminal Trials: Making the Error Rate the Primary Factor in Daubert's Validity Inquiry. *New York University Law Review*, 85(6), pp.2034-2064.

Jakobson, R. (1971). *Word and Language*. The Hague: Mouton.

Johnson, A. and Woolls, D. (2010). Who wrote this? The linguist as detective. In: S. Hunston and D. Oakey, ed., *Introducing Applied Linguistics: Concepts and Skills*. London: Routledge, pp.111–118.

Johnson, A. and Wright, D. (2014). Identifying İdiolect in Forensic Authorship Attribution: An n-gram Textbite Approach. *Language and Law (Linguagem e Direito)*, 1(1), pp.37-69.

Johnson, S. (1997). Theorizing Language and Masculinity: A Feminist Perspective. In: S. Johnson and U. Meinhof, ed., *Language and Masculinity*. Oxford: Blackwell, pp.8–26.

Juola, P. (2006). Authorship Attribution for Electronic Documents. In: M. Olivier and S. Shenoi, ed., *Advances in Digital Forensics II*. New York, NY: Springer, pp.119-130.

Juola, P. (2008). *Authorship attribution*. Boston: NOW Publishing.

Juola, P. (2013). Stylometry and immigration: A case study. *Journal of Law and Policy*, 2(21), pp.287-298.

Juola, P. (2015). The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digital Scholarship in the Humanities*, [online] pp.i100–i113. Available at: https://academic.oup.com/dsh/article/30/suppl_1/i100/363234 [Accessed 20 May 2018].

Kara, M. (2006). İnternet Türkçesinin Çığlığı: Türkçe Dili (!) ve Diğerleri. *Akademik Araştırmalar Dergisi*, 30, pp.157-170.

Katsuno, H. and Yano, C. (2007). Kaomoji and Expressivity in a Japanese Housewives' Chat Room. In: B. Danet and S. Herring, ed., *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford: Oxford University Press, pp.278-300.

Kaye, D. (2001). The Dynamics of Daubert: Methodology, Conclusions, and Fit in Statistical and Econometric Studies. *Virginia Law Review*, 87(8), pp.1933-2018.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Longman.

Kestemont, M., Luyckx, K., Daelemans, W. and Crombez, T. (2012). Cross-Genre Authorship Verification Using Unmasking. *English Studies*, 93(3), pp.340-356.

Kinkus, J. (2002). Science and Technology Resources on the Internet: Computer Security. *Issues in Science & Technology Librarianship*, [online] (36). Available at: http://doi.org/10.5062/F4QN64P2 [Accessed 17 Mar. 2018].

Kniffka, H. (2007). *Working in Language and Law*. Basingstoke: Palgrave Macmillan.

Komito, L. (1998). The Net as a Foraging Society: Flexible Communities. *The Information Society*, 14(2), pp.97-106.

Konda (2007). *Toplumsal Yapı Araştırması 2006: Biz Kimiz?*. Istanbul: Konda Araştırma ve Danışmanlık, pp.Available at: http://konda.com.tr/wp-content/uploads/2017/02/2006_09_KONDA_Toplumsal_Yapi.pdf [Accessed 17 Mar. 2018].

Koppel, M. and Schler, J. (2018). Exploiting Stylistic Idiosyncrasies for Authorship Attribution. In: *Proceedings of the 18th IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis*. [online] Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.3019. [Accessed 10 Feb. 2018].

Koppel, M., Schler, J. and Argamon, S. (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, [online] 60(1), pp.9-26. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20961 [Accessed 7 May 2018].

Koppel, M., Schler, J. and Argamon, S. (2011). Authorship Attribution in the Wild. *Language Resources and Evaluation*, [online] 45(1), pp.83-94. Available at: https://link.springer.com/article/10.1007%2Fs10579-009-9111-2 [Accessed 1 Mar. 2018].

Koppel, M., Schler, J. and Argamon, S. (2013). Authorship Attribution: What's easy and what's hard?. *Journal of Law and Policy*, 21(2), pp.317–331.

Koppel, M., Schler, J., Argamon, S. and Messeri, E. (2006). Authorship attribution with thousands of candidate authors. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, [online] pp.659-660. Available at: https://dl.acm.org/citation.cfm?id=1148304 [Accessed 2 Apr. 2018].

Kotzé, E. (2010). Author Identification from Opposing Perspectives in Forensic Linguistics. *Southern African Linguistics and Applied Language Studies*, [online] 28(2), pp.185-197. Available at: https://doi.org/10.2989/16073614.2010.519111 [Accessed 17 Feb. 2018].

Kredens, K. (2002). Towards a Corpus-based Methodology of Forensic Authorship Attribution: A Comparative Study of Two Idiolects. In: B. Lewandowska-Tomaszczyk, ed., *PALC'01: Practical Applications in Language Corpora*. Frankfurt am Mein: Peter Lang, pp.405–437.

Kredens, K. (2006). On the Status of Linguistic Evidence in Litigation. In: P. Nowak and P. Nowakowski, ed., *Language, Communication, Information*. Poznan: Sorus Publishers, pp.23-30.

Kredens, K. and Coulthard, M. (2012). Corpus Linguistics inAauthorship Identification. In: P. Tiersma and L. Solan, ed., *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press, pp.504–516.

Krippendorff, K. (2004). Reliability in Content Analysis : Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3), pp.411-433.

Kucukyilmaz, T., Cambazoglu, B., Aykanat, C. and Can, F. (2008). Chat Mining: Predicting User and Message Attributes In Computer-Mediated Communication. *Information Processing & Management*, [online] 44(4), pp.1448-1466. Available at: https://doi.org/10.1016/j.ipm.2007.12.009 [Accessed 16 Jan. 2018].

Larner, S. (2014). A Preliminary Investigation into the Use of Fixed Formulaic Sequences as a Marker of Authorship. *International Journal of Speech Language and the Law*, [online] 21(1), pp.1-22. Available at: https://journals.equinoxpub.com/index.php/IJSLL/article/view/15423 [Accessed 16 Nov. 2017].

Layton, R., Watters, P. and Dazeley, R. (2010). Authorship Attribution for Twitter in 140 Characters or Less. In: *2010 Second Cybercrime and Trustworthy Computing Workshop*. pp.1-8.

Leech, G. (1992). Corpora and Theories of Linguistic Performance. In: J. Svartvik, ed., *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*. Berlin: Mouton de Gruyter, pp.125–148.

Leech, G. (2006). *A Glossary of English Grammar*. Edinburgh: Edinburgh University Press.

Leonard, R. (2006). Forensic Linguistics: Applying the Scientific Principles of Language Analysis to Issues of the Law. *International Journal of the Humanities*, 3(7), pp.65-69.

Leonard, R., Ford, J. and Christensen, T. (2017). Forensic Linguistics: Applying the Science of Linguistics to Issues of the Law. *Hofstra Law Review*, 45(3), pp.881-897.

Lombard, M., Snyder-Duch, J. and Bracken, C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, [online] 28(4), pp.587-604. Available at: https://doi.org/10.1111/j.1468-2958.2002.tb00826.x [Accessed 17 Oct. 2017].

López-Escobedo, F., Méndez-Cruz, C., Sierra, G. and Solórzano-Soto, J. (2013). Analysis of Stylometric Variables in Long and Short Texts. *Procedia - Social and Behavioral Sciences*, 95, pp.604-611.

Love, H. (2002). *Authorship and Attribution*. Cambridge: Cambridge University Press.

Lüdeling, A., Evert, S. and Baroni, M. (2006). Using Web Data for Linguistic Purposes. *Language and Computers*, 1(59), pp.7-24.

Luyckx, K. (2010). *Scalability Isues in Authorship Attribution*. PhD thesis. Proefschrift Universiteit Antwerpen.

Luyckx, K. and Daelemans, W. (2008). Using Syntactic Features to Predict Author Personality from Text. In: *Proceedings of Digital Humanities 2008*. [online] Available at: http://www.cnts.ua.ac.be/papers/2008/LD08dh.pdf [Accessed 20 Sep. 2017].

Luyckx, K. and Daelemans, W. (2010). The Effect of Author Set Size and Data Size in Authorship Attribution. *Literary and Linguistic Computing*, 26(1), pp.35-55.

MacEnery, T. and Wilson, A. (2001). *Corpus Linguistics: An Introduction*. 2nd ed. Edinburgh: Edinburgh University Press.

MacLeod, N. (2010). *Police Interviews with Women Reporting Rape: A Critical Discourse Analysis*. PhD thesis. Aston University.

MacLeod, N. and Grant, T. (2012). Whose Tweet? Authorship Analysis of Micro-blogs and Other Short Form Messages. In: S. Tomblin, N. MacLeod, R. Sousa-Silva and M. Coulthard, ed., *Proceedings of the Tenth International Association of Forensic Linguists' Biennial Conference*. [online] Aston University, Birmingham, pp.210–224. Available at: http://www.forensiclinguistics.net [Accessed 8 Sep. 2016].

Madge, C. (2007). Developing a Geographers' Agenda for Online Research Ethics. *Progress in Human Geography*, [online] 31(5), pp.654-674. Available at: http://journals.sagepub.com/doi/10.1177/0309132507081496 [Accessed 2 Apr. 2017].

Madigan, D., Genkin, A., Lewis, D., Argamon, S., Fradkin, D. and Ye, L. (2005). Author Identification on the Large Scale. In: *Proc. of Classification Society of N. America, 2005*. [online] pp.1-20. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.5324&rep=rep1&type=pdf [Accessed 11 Dec. 2017].

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

McMenamin, G. (2001). Style Markers in Authorship Studies. *Forensic Linguistics*, 8(2), pp.93-97.

McMenamin, G. (2002). *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton, FLA.: CRC Press.

McMenamin, G. (2010). Forensic Stylistics. Theory and Practice of Forensic Stylistics. In: M. Coulthard and A. Johnson, ed., *The Routledge Handbook of Forensic Linguistics*. London: Routledge, pp.487–507.

Mendenhall, T. (1887). The Characteristic Curves of Composition. *Science*, [online] 9(214), pp.237-249. Available at: https://www.jstor.org/stable/1764604 [Accessed 21 Nov. 2017].

Mikros, G. and Perifanos, K. (2013). Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles. In: *AAAI Spring Symposium: Analyzing Microtext*. [online] Association for the Advancement of Artificial Intelligence. Available at: https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/download/5714/5914 [Accessed 17 May 2018].

Mingzhe, J. and Minghu, J. (2012). Text Clustering on Authorship Attribution Based on the Features of Punctuations Usage, Signal Processing (ICSP). In: *Proceedings of the 11th International IEEE Conference on Digital Object Identifiers*. IEEE, pp.217 – 2178.

Mischaud, E. (2007). *Twitter: Expressions of the Whole Self. An investigation into user appropriation of a web-based communications platform*. Master's thesis. London School of Economics and Political Science.

Mosteller, F. and Wallace, D. (1963). Inference in an Authorship Problem. *Journal of the American Statistical Association*, 58(302), pp.275-309.

Mosteller, F. and Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley Publishing Company Inc.

Nagarajan, M., Purohit, H. and Sheth, A. (2010). A Qualitative Examination of Topical Tweet and Retweet Practices. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. [online] pp.295-298. Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1484/1880 [Accessed 16 Nov. 2018].

Nazar, R. and Sánchez Pol, M. (2007). An Extremely Simple Authorship Attribution System. In: M. Turell, J. Cicres and M. Spassova, ed., *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics / Language and the Law 2006*. Barcelona: Documenta Universitaria.

Neuendorf, K. (2002). *The Content Analysis Guidebook*. London: Sage Publications.

Nini, A. (2015). *Authorship Profiling in a Forensic Context*. PhD thesis. Aston University.

Nini, A. (2018). An Authorship Analysis of the Jack the Ripper Letters. *Digital Scholarship in the Humanities*, [online] fqx065, pp.1-16. Available at: https://doi.org/10.1093/llc/fqx065 [Accessed 17 Mar. 2018].

NVivo. (2018). QSR International Pty Ltd, https://www.qsrinternational.com/nvivo/home.

Oldenburg, R. (1989). *The great good place*. New York: Da Capo Press.

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

Olsson, J. (2008). *Forensic Linguistics*. London: Continuum.

Overdorf, R. and Greenstadt, R. (2016). Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution. *Proceedings on Privacy Enhancing Technologies*, 2016(3), pp.155–171.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N. and Smith, N. (2013). Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [online] Association for Computational Linguistics, pp.390-391. Available at: http://repository.cmu.edu/cgi/viewcontent.cgi?article=1039&context=lti [Accessed 24 May 2018].

Oxburgh, G., Myklebust, T., Grant, T. and Milne, R. (2015). Communication in Investigative and Legal Settings. *Communication in Investigative and Legal Contexts*, [online] pp.1-13. Available at: https://doi.org/10.1002/9781118769133.ch1 [Accessed 8 Jan. 2018].

Polat, N. (2007). Linking Social Networks and Attainment in an L2 Accent: Kurds Acquiring Turkish. In: *Proceedings of the Fifteenth Annual Symposium About Language and Society-Austin*. [online] Texas Linguistic Forum 51, pp.144-153. Available at: http://salsa.ling.utexas.edu/proceedings/2007/Polat.pdf [Accessed 29 Jan. 2018].

Queralt Estevez, S. and Turell Julià, M. (2013). A semi-automatic Authorship Attribution Technique Applied to Real Forensic Cases Involving Judgments in Spanish. In: R. Sousa-Silva, R. Faria, N. Gavaldà and B. Maia, ed., *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., pp.10-18.

Rheingold, H. (1993). *The Virtual Community*. Reading, Mass.: Addison-Wesley.

Rico-Sulayes, A. (2011). Statistical Authorship Attribution of Mexican Drug Trafficking Online Forum Posts. *International Journal of Speech Language and the Law*, 18(1), pp.53–74.

RStudio (2011). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc.

Rudman, J. (1998). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities*, [online] 31(4), pp.351-365. Available at: https://doi.org/10.1023/A:100101862 [Accessed 10 Mar. 2017].

Sánchez Pol, M. (2005). A Stylometry-Based Method to Measure Intra and Inter-Authorial faithfulness for Forensic Applications. In: S. Argamon, J. Karlgren and J. Shanahan,

ed., *Stylistic Analysis Of Text For Information Access*. [online] Kista: Swedish Institute of Computer Science, pp.11-15. Available at: https://pdfs.semanticscholar.org/d05f/32f6e28c2e20e4903ce9b9ff5b9d53d135b0.pdf [Accessed 11 Apr. 2018].

Santini, M. (2007). Characterizing Genres of Web Pages: Genre Hybridism and Individualization. In: *Proceedings of the 40th Hawaii International Conference on System Sciences - 2007*. [online] IEEE, pp.1-10. Available at: http://doi.org/10.1109/HICSS.2007.124 [Accessed 15 Apr. 2017].

Schmied, J. (1993). Qualitative and Quantitative Research Approaches to English Relative Constructions. In: C. Souter and E. Atwell, ed., *Corpus-Based Computational Linguistics*. Amsterdam: Rodopi, pp.85-96.

Schwartz, M. (2016). An Examination of Cross-Domain Authorship Attribution Techniques. *CUNY Academic Works*. [online] Available at: https://academicworks.cuny.edu/gc_etds/1573/ [Accessed 15 May 2018].

Schwartz, R., Tsur, O., Rappoport, A. and Koppel, M. (2013). Authorship Attribution of Micro-Messages. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. [online] Seattle, Washington, USA: Association for Computational Linguistics, pp.1880–1891. Available at: http://www.aclweb.org/anthology/D13-1193 [Accessed 9 Apr. 2018].

Scott, M. (2012). *WordSmith Tools*. Stroud: Lexical Analysis Software.

Shepherd, M. and Watters, C. (1998). The Evolution of Cybergenres. *Proceedings of the Thirty-First Hawaii International Conference on System Sciences*. [online] Available at: http://doi.org/10.1109/hicss.1998.651688 [Accessed 26 Jun. 2018].

Shuy, R. (2009). Ethical Questions in Forensic Linguistics: Introduction to Papers from a Linguistic Society of America Panel Presentation, San Francisco, California, January 9, 2009. *International Journal of Speech Language and the Law*, [online] 16(2), pp.219-226. Available at: https://journals.equinoxpub.com/index.php/IJSLL/article/view/6599 [Accessed 19 Jan. 2018].

Sinclair, J. (1988). *Collins COBUILD Essential English Dictionary*. London: HarperCollins.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (2004). *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Sixsmith, J. and Murray, C. (2001). Ethical Issues in the Documentary Data Analysis of Internet Posts and Archives. *Qualitative Health Research*, [online] 11(3), pp.423-432. Available at: https://doi.org/10.1177/104973201129119109 [Accessed 21 Feb. 2018].

Sketch Engine (2018). *Sketch Engine | language corpus management and query system*. [online] Sketchengine.eu. Available at: https://www.sketchengine.eu/ [Accessed 12 May 2018].

Smith, D., Spencer, S. and Grant, T. (2009). *Authorship Analysis for Counter Terrorism*. Unpublished Research Report. QinetiQ/Aston University.

Solan, L. and Tiersma, P. (2005). *Speaking of Crime: The Language of Criminal Justice*. Chicago: University of Chicago Press.

Solan, M. (2013). Intuition versus Algorithm: The Case of Forensic Authorship Attribution. *Journal of Law and Policy*, 21(2), pp.551-576.

Sousa Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E. and Maia, B. (2011). 'twazn me!!! ;(' Automatic Authorship Analysis of Micro-Blogging Messages. In: *Muñoz R., Montoyo A., Métais E. (eds) Proceedings of the 16th international conference on Natural Language Processing and Information Systems. NLDB 2011. Lecture Notes in Computer Science, vol 6716*. [online] Berlin, Heidelberg: Springer, pp.161–168. Available at: https://doi.org/10.1007/978-3-642-22327-3_16 [Accessed 13 Jan. 2018].

Sousa-Silva, R., Sarmento, L., Grant, T., Oliveira, E. and Maia, B. (2010). Comparing Sentence-Level Features for Authorship Analysis in Portuguese. In: T. Pardo, A. Branco, A. Klautau, R. Vieira and V. de Lima, ed., *Computational Processing of the Portuguese Language. PROPOR 2010. Lecture Notes in Computer Science, vol 6001*. Berlin, Heidelberg: Springer, pp.51-54.

Spassova, M. and Turell,, M. (2007). The Use of Morphosyntactically Annotated Tag Sequences as Markers of Authorship. In: M. Turell, J. Cicres and M. Spassova, ed., *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics / Language and the Law 2006*. Barcelona: Documenta Universitaria.

Stamatatos, E. (2007). Author Identification Using Imbalanced and Limited Training Texts. [online] pp.237-241. Available at: http://doi.org/10.1109/dexa.2007.5 [Accessed 19 Dec. 2017].

Stamatatos, E. (2008). Author Identification: Using Text Sampling to Handle the Class Imbalance Problem. *Information Processing & Management*, 44(2), pp.790-799.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, [online] 60(3), pp.538-556. Available at: https://doi.org/10.1002/asi.21001 [Accessed 5 May 2018].

Stamatatos, E. (2013). on the Robustness of Authorship Attribution Based on Character N-Gram Features. *Journal of Law & Policy*, 21(2), pp.421-439.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2001). Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities*, [online] 35(2),

pp.193-214. Available at: https://doi.org/10.1023/A:1002681919510 [Accessed 23 May 2017].

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000). Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, [online] 26(4), pp.471-495. Available at: https://doi.org/10.1162/089120100750105920 [Accessed 8 Mar. 2018].

Stubbs, M. (2005). Conrad in the Computer: Examples of Quantitative Stylistic Methods. *Language and Literature*, [online] 14(1), pp.5-24. Available at: https://doi.org/10.1177/0963947005048873 [Accessed 29 Mar. 2018].

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Takci, H. and Ekinci, E. (2012). Character Level Authorship Attribution for Turkish Text Documents. *The Online Journal of Science and Technology-TOJSAT*, 2(3), pp.12-16.

Tas, T. and Gorur, A. (2007). Author Identification for Turkish Texts. *Cankaya University Journal of Arts and Sciences*, 1(7), pp.151-161.

Teknomo, K. (2015). *Similarity Measurement*. [online] People.revoledu.com. Available at: https://people.revoledu.com/kardi/tutorial/Similarity/ [Accessed 24 Oct. 2018].

Temur, T. and Vuruş, N. (2009). İnternet (Genel Ağ) Ortamında Türkçe'nin Kullanımına İliskin Bir Çözümleme. *Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, [online] 12(22), pp.232-244. Available at: http://sbe.balikesir.edu.tr/dergi/edergi/c12s22/makale/c12s22m16.pdf [Accessed 2 Feb. 2017].

Tereszkiewicz, A. (2013). *Genre Analysis of Online Encyclopedias: The Case of Wikipedia*. Krakow: Jagiellonian University Press.

The Law Commission (2011). *Expert Evidence in Criminal Proceedings in England and Wales*. [ebook] London: The Stationery Office. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/229043/0829.pdf [Accessed 13 Jan. 2018].

Thurlow, C., Lengel, L. and Tomic, A. (2004). *Computer Mediated Communication*. London: Sage.

Tiersma, P. (2010). The Origins of Legal Language. In: L. Solan and P. Tiersma, ed., *Oxford Handbook of Language and Law*. [online] Loyola-LA Legal Studies Paper No. 2009-45, pp.1-26. Available at: https://ssrn.com/abstract=1695226 [Accessed 9 Mar. 2018].

Tiersma, P. and Solan, L. (2002). The Linguist on the Witness Stand: Forensic Linguistics in American Courts. *Language*, 78(2), pp.221-239.

Tinsley, H. and Weiss, D. (2000). Interrater Reliability and Agreement. In: H. Tinsley and S. Brown, ed., *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, 1st ed. New York: Academic Press, pp.94-124.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tomblin, S. (2013). *To Cut a Long Story Short an Analysis of Formulaic Sequences In Short Written Narratives and Their Potential as Markers of Authorship*. PhD Thesis. Aston University.

Turell, M. (2011). The Use of Textual, Grammatical and Sociolinguistic Evidence in Forensic Text Comparison. *International Journal of Speech Language and the Law*, [online] 17(2), pp.211-250. Available at: https://journals.equinoxpub.com/index.php/IJSLL/article/view/6409 [Accessed 14 Jan. 2018].

Turell, M. and Gavaldà, N. (2013). Towards an Index of Idiolectal Similitude (or Distance) in Forensic Authorship Analysis. *Journal of Law and Policy*, 21(2), pp.495–514.

Turell, M. and Rosso, P. (2013). Computational Approaches to Plagiarism Detection and Authorship Attribution in Real Forensic Cases. In: R. Sousa-Silva, R. Faria, N. Gavaldà and B. Maia, ed., *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., pp.19-30.

Turkish Criminal Procedure Code (2009). *Criminal codes - Legislationline*. [online] Legislationline.org. Available at: http://www.legislationline.org/documents/section/criminal-codes/country/50 [Accessed 11 Mar. 2016].

Türkoğlu, F., Diri, B. and Amasyalı, M. (2007). Author Attribution of Turkish Texts by Feature Mining. In: D. Huang, L. Heutte and M. Loog, ed., *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues. ICIC 2007. Lecture Notes in Computer Science, vol 4681*. [online] Berlin, Heidelberg: Springer, pp.1086–1093. Available at: https://doi.org/10.1007/978-3-540-74171-8_110 [Accessed 16 Apr. 2018].

Twitter Blog (2017). *English (US)*. [online] Blog.twitter.com. Available at: https://blog.twitter.com/ [Accessed 6 Jan. 2017].

Ustunova, K. (2002). *Türkçede Yapı Kavramı ve Söz Dizimi İncelemeleri*. Bursa: Uludağ Üniversitesi Basımevi.

van Halteren, H. (2004). Linguistic Profiling for Author Recognition and Verification. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*. [online] Association for Computational Linguistics. Available at: http://doi.org/10.3115/1218955.1218981 [Accessed 11 Feb. 2018].

van Halteren, H. (2007). Author verification by linguistic profiling. *ACM Transactions on Speech and Language Processing*, 4(1), pp.1-17.

Wellman, B. and Gulia, M. (1999). Net Surfers Don't Ride Alone: Virtual Communities as Communities. In: B. Wellman, ed., *Networks in the Global Village*. Boulder, Colo.: Westview Press, pp.331-366.

Wikicount.net. (2018). *How many articles are there on Wikipedia? - Wikipedia article count*. [online] Available at: http://wikicount.net/ [Accessed 15 Feb. 2018].

Wikipedia. (2018). *Wikipedia, the free encyclopedia*. [online] Available at: https://www.wikipedia.org/ [Accessed 15 Feb. 2018].

Williams, K. (2000). Reproduced and Emergent Genres of Communication on the World Wide Web. *The Information Society*, 16(3), pp.201-215.

Wright, D. (2013). Stylistic Variation within Genre Conventions in the Enron Email Corpus: Developing a Textsensitive Methodology for Authorship Research. *International Journal of Speech Language and the Law*, [online] 20(1). Available at: https://journals.equinoxpub.com/index.php/IJSLL/article/view/10595 [Accessed 10 May 2018].

Wright, D. (2014). *Stylistics versus Statistics: A Corpus Linguistic Approach to Combining Techniques in Forensic Authorship Analysis Using Enron Emails*. PhD thesis. University of Leeds.

Wright, D. (2017). Using Word N-Grams to Identify Authors and Idiolects. *International Journal of Corpus Linguistics*, [online] 22(2), pp.212-241. Available at: http://www.jbe-platform.com/content/journals/10.1075/ijcl.22.2.03wri [Accessed 9 May 2018].

Yaman, H. and Erdogan, Y. (2007). İnternet Kullanımının Türkçeye Etkileri: Nitel Bir Araştırma. *Journal of Language and Linguistic Studies*, 3(2), pp.237-249.

Yannikos, Y., Graner, L., Steinebach, M. and Winter, C. (2014). Data Corpora for Digital Forensics Education and Research. In: *Peterson G., Shenoi S. (eds) Advances in Digital Forensics X. DigitalForensics 2014. IFIP Advances in Information and Communication Technology, vol 433*. [online] Berlin, Heidelberg: Springer, pp.309-325. Available at: http://doi.org/10.1007/978-3-662-44952-3_21 [Accessed 9 Dec. 2017].

Yates, J. and Orlikowski, W. (1992). Genres of Organizational Communication: A Structurational Approach to Studying Communication and Media. *Academy of Management Review*, 17(2), pp.299-326.

Yates, J. and Orlikowski, W. (2002). Genre Systems: Structuring Interaction Through Communicative Norms. *Journal of Business Communication*, 39(1), pp.13-35.

Yule, G. (1939). On Sentence- Length as A Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship. *Biometrika*, [online] 30(3-4), pp.363-390. Available at: https://doi.org/10.1093/biomet/30.3-4.363 [Accessed 20 Mar. 2017].

Zappavigna, M. (2012). *The Discourse of Twitter and Social Media: How we Use Language to Create Affiliation on the Web*. London: Continuum.

Zheng, R., Li, J., Chen, H. and 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*, [online] 57(3), pp.378-393. Available at: https://doi.org/10.1002/asi.20316 [Accessed 16 Nov. 2017].

Zheng, R., Qin, Y., Huang, Z. and Chen, H. (2003). Authorship Analysis in Cybercrime Investigation. In: H. Chen, R. Miranda, D. Zeng, C. Demchak, J. Schroeder and T. Madhusudan, ed., *Intelligence and Security Informatics. ISI 2003*. Berlin, Heidelberg: Springer, pp.59-73.

Zipf, G. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press.

Appendix List

A. Ethics

B. NVivo Projects

C. Inter-Codes Tests

D. Jaccard Outputs

Authorship Attribution in Turkish Texts:

- introduces useful linguistic concepts and tools
- outlines the methods forensic linguists employ to analyse written texts in legal situations
- introduces topics such as: forensic authorship attribution in Turkish language, feature selection, the role of feature types in attributing authorship and the work linguists do to help solve language crimes.

Dr. Hülya Kocagül Yüzer is a lecturer in Linguistics at İstanbul Medeniyet University, İstanbul, Türkiye

ARTSÜREM