INTERNATIONAL PREMIER JOURNAL OF LANGUAGES & LITERATURE (IPJLL)
VOLUME 3, ISSUE 3, 2025 p-ISSN: 3007-2336 e-ISSN: 3007-2344

THE ROLE OF CORPUS-BASED FORENSIC LINGUISTIC APPROACHES IN PROFILING THE SUSPECTS OF TAX EVASION

Adeel Ahmad adeel.ahmed@gmail.com	PhD Scholar, Department of English, The University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan.
Shahida Khalique shahida.khalique@ajku.edu.pk	Assistant Professor, Department of English, University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan.

Abstract

Tax evasion has been a serious issue in the financial based administrative system of Pakistan which has played its adverse role in declining the economic stability of the country since decades. Though many managerial efforts have been made to address the issue yet it has remained there obstinately. In this regard the role of language is imperative which is effectively used in the process of financial deception by manipulating it strategically. The focus of this study is on the prospects of integrating the corpus based forensic linguistic techniques in profiling the suspects of tax evasion in Pakistan. This is done by considering the selected corpus related to taxation and taking into consideration the approach of analyzing it through the consideration of forensic linguistic approach using corpus software LancsBox. This method covers the resources for the language based outlining of the tax evasion issue on the foundation of linguistic analyses of the selected corpora which is used as linguistic markers in the legal spheres. This study shows that the corpus based forensic linguistic approach is a significant framework of not only detecting the patterns of language used in the tax evasion but it can also be useful for the linguistic profiling of suspects involved in tax evasion through analysis of the communication patterns.

Keywords: *Corpora*, *Corpus-Based,* *Forensic Linguistics*, *Deception*, *Linguistic Profiling*, *Tax Evasion*

Corresponding Author: Shahida Khalique (Assistant Professor, Department of English, University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan).
Email: shahida.khalique@ajku.edu.pk

1.Introduction

Pakistan's ranking in terms of tax payments is amongst the lowest as compared with the rest of the world, which shows the state of affairs in relation to the tax avoidance in the country. As per results of Census (2023), the total population of the country comprises of about 241.5 Million. FBR (2023), reports that out of such huge population, only 2.45 % are the filers of tax. This clearly indicates an extremely small proportion of population involved in the process of following the legal responsibility of tax payments in the country. The aspect of tax evasion postures a significant challenge to the financial stability and the public objectivity in Pakistan. Despite so many years of the country's independence, the tax culture has not prevailed effectively in the country, which has resulted in keeping it underdeveloped even after decades despite having potential to grow with a great population and resources. The key reasons related to the negligible proportion of population worried about the payment of tax liability include the system flaws, the procedural loop holes, deficiency of public awareness and other social and cultural issues besides the lack of responsiveness regarding the importance of tax payments.

It also demonstrates that most people in Pakistan are keen to save themselves from the tax payments by using sophisticated tactics. Language being the key tool used for the intentional deception of financial facts has been overlooked rather the traditional methods of the financial audits and investigations have usually been carried out to address the issue in Pakistan. The procedure of confined financial investigation has been proved insufficient that required more advanced methods of exploration and especially the aspect of language based study approach. In this regard, the execution of corpus linguistics methodology is imperative because it makes use of the real world data such as tax discourse for the authentic linguistic analysis. Corpus linguistics has also been used historically to investigate effectively the language aspect in the subject matter. Sinclair (2004) is of the view that corpus linguistics considers the large collection of real world data sets for the linguistic research. Corpus is the large set of real world data that is used the linguistic research so as to make the analysis more generalizable and authentic. Therefore the corpus based approach of linguistic research is significant in relation to study the tax evasion issue in Pakistan.

On the other hand, forensic linguistics deals with the aspect of financial crimes such as tax evasion or tax avoidance. Coulthard (1994) considers that corpus analysis of data can be helpful in the financial crime detection of tax evasion taking into account the forensic side with the corpora. On the other hand Shuy (2005) also think that the use of forensic linguistics approaches in financial matters is of great significance with respect to its application. Forensic linguistics is comparatively new with respect to its application in different fields and more specifically in the financial matters in Pakistan. Also, it has been ascertained that the forensic linguistics methodologies can contribute significantly in Pakistan's taxation system. This is due to the fact that all the narrated language oriented strategies of tax evasion including creation of vagueness, concealment, obfuscation and the equivocation are practiced in Pakistan, to mislead the tax authorities (Ahmed, 2021). The role of forensic linguistics thus is very crucial with respect to the exploration of tax evasion in Pakistan and considering the connection of language and tax law. The persistence of the problem of tax evasion in the Pakistan calls for the consideration of the issue on the line of the interdisciplinary approach of study.

This study presents the general ideologies concerning the use of corpus linguistics approach in discovering the forensic linguistic evidence in relation to the issue of tax evasion in Pakistan. As, McVay (2006) proposes the effectiveness of corpus based analysis in detection of tax evasion by identifying the patterns of manipulation and financial deception through analysis of large texts. This is due to the fact that it helps in exploring the unique communication patterns showing linguistic deception found in the large legal tax discourse under different settings considering the taxation system of the country. The focus of this research is to uncover the role of the corpus based forensic linguistic techniques to expose the tax evasion and likewise profiling the suspects of the tax evasion in Pakistan through the analysis.

1.1. Research Questions

1. How can the corpus-based forensic linguistic methods be used to identify linguistic patterns associated with tax evasion in Pakistan?

2. To what extent can corpus-based forensic linguistic profiling assist in identifying the suspects involved in tax evasion cases in Pakistan?

2.Literature Review

Criminalities such as tax evasion are considered in the arena of criminology

through the support of law and other sciences. According to Turell (2001), the forensic linguistics has an imperative role to play for the identification of the language based deceptions and also it can play its key role to highlight the vague and ambiguous linguistic forms which are intended to deceit somebody for the purpose of attainment of some vested aims. Language crimes are based on the idea of the use of language for the purpose of deceit, trickery, offense and hate speech, misrepresentation of details, thoughtful miscommunication or deceiving the legal authorities in the financial matters. All these language based tricks are counted as the language crimes. Gibbons (2003) points out that criminalities are achieved by numerous illegitimate speech acts such as kickback offer, bribe offers, threats etc. In the investigation of these languages based crimes, the forensic linguistics offers a critical support through its variant approaches.

Coulthard (2010) holds that every speaker or writer has his distinct style of language in spoken or written form which is termed as an '*idiolect*' having unique vocabulary, lexical choices and grammatical choices used for the purpose of communication. This is same in relation to the fact that as everybody has the unique style of writing like someone's unique finger prints and note matching with others. Same is applied through the introduction of the computer software that checks the plagiarism out of the texts which are considered suspicious in terms of their real authorship (MacMenamin, 2002).

Language is the means of communication that connects the legal authorities with the suspects during the process of court oriented proceedings to come to the legal findings (Shuy, 1993). Forensic linguistics considers the connection of the language with the law and provides with the requisite tools to be used to analyze the deceptive discourse and uncover the possible tactics used by the individuals in a legal sphere to disguise. In the words of Coulthard & Johnson (2007), forensic linguistics can expose the linguistic strategies employed in the texts to reach to the possible intentions of the writer through the approach of linguistic profiling where tactics like the vagueness, indirectness and the use of hedging beside many other techniques determine the possibility of the deception through language use. Similarly, law is thoroughly a language based institution which is translated and made simple by the language and the task of a forensic linguist is to make it simple for the concerned persons to identify anything mysterious (Gibbson, 2003).

In case of taxation process, analyzing the language of financial statements and tax related documents can help in the identification of evasive approaches to gain monetary benefits. Some other pragmatic concepts including speech acts by Austin

(1962) and contextual background as per view of Levinson (1983) are also very relevant during the forensic linguistic analysis. This is due to the fact that these approaches define how the individuals use language signs to avoid responsibility and accountability or to mislead the authorities in certain legal matters by considering the non-cooperation in communication.

A corpus is the assortment of text written or spoken that has been collected in order to be analyzed linguistically. Sinclair (2004), states that a large collection of data that makes the empirical foundation of linguistics research is corpus. Since the advent of the modern-day corpus linguistics, many fields have benefitted from its ability to identify patterns in texts; some of the earliest and most seminal work in forensic linguistics is corpus-based analysis of the language used. Svartvik (1968) used a corpus-based methodology to analyze the witness statements in the murder case.

Coulthard (1994) used specific corpora of witness statements and police statements. Though the corpus-based approaches in the stated cases have been found successful but the utility of corpus-based linguistics research approach found its foot through its combination with the forensic linguistics research. For this Kredens (2002) is the foremost who has used the approach of corpus based analysis with the forensic linguistics, in order to compare the idiolect of two musicians while Grant (2013) used the corpus approach to identify lexical variations while investigating a murder case.

Likewise, Johnson and Wright (2014) made use of this corpus based forensic approach for the analysis of authorship identification in the business emails. Similarly, Sinclair (1991) presented the '*idiom principle*' with a view that speakers have the option of using some pre-constructed phrases.

Likewise, Nattinger and DeCarrio (1992) presented '*principle of pragmatic competence*' which operated lexical phrases. Based on brief findings of the studies presented above, it is determined that corpus-based approach is the most appropriate way of forensic linguistics analyses in legal matters.

3.Research Methodology

This study adopts a mixed method of linguistic research that amalgamates both quantitative and qualitative methodologies. These approaches are also a combination of a primary and the secondary procedures of the study respectively based on their linguistic considerations. The primary method is the corpus based linguistic study method, which includes the collection, compilation, analysis and the interpretation of the specific corpus

of the legal text pertaining to the taxation process in Pakistan. This corpus oriented approach on one hand allows a thorough exploration of the linguistic features and patterns that cause the aspects of possible financial deception and evasion during the process of tax related communications.

In addition to the corpus oriented approach, a qualitative element is integrated by the minute investigation of the tax evasion text and the relevant linguistic occurrences acquired from the tax discourse; the morpho-syntactic apparatus is used for the qualitative study of the evasion instances at the word and phrase levels.

In order to categorize the instances of tax evasion tactics and the evasion strategies employed by the individuals during the taxation process, the pragmatic consideration and contextualization of the identified linguistic patterns from the discourse is also done. Through the combination of these quantitative and the qualitative methods of study, the understanding of the tax evasion practices incorporated by the evaders in the deceiving discourse through the language use has been tried to be comprehended.

3.1. Data Collection

The fundamental of the study is the accumulation of the specific tax related corpus, and the primary step in the process is data collection. Thirty significant tax relevant documents available online as well as the tax evasion cases files obtained from the tax establishments and legal proceedings undertaken in Pakistan are the sources of the data collection for the purpose of this study. These files contain the communication documents such as audit reports, financial statements, assessment orders and the legal correspondence on tax matters. Also court reported judgements on taxation cases are the important sources of data. The data is text oriented and was placed in separate files on the basis of the categorization of tax case. The data has been collected in such a way that it may be relevant.

3.2. Corpus Compilation

The text has been selected cautiously from different sources, so as to make it represented of the comprehensive collection of discourse pertaining to the tax evasion issue of Pakistan. The data is made anonymized and unnecessary details are omitted from it to make a represented corpus. The selected texts have been separated into different sub-corpora based on the case or the context of use. These linguistic contexts include the income underreporting, asset concealments or fraudulent deductions, non-declaration or under declarations, other financial claims etc. The total word count of the corpus is more

than half million, which is sufficient for the frequency centered quantitative analysis and also for the onward qualitative analysis. To make it compatible with the software the word files are transformed in to txt files of corpus. The corpus pertaining to the tax related documents in the pdf format files have been imported into the LancsBox for analysis.

3.3. Meta Data

Meta data of each text is also maintained within the separate excel sheets wherein the information about the text has been compiled which include the category of tax case, the entity involved in the process, the nature of tax case, the forum where it has been under consideration and the kind of tax or head of income taken up for consideration

4. Data Analysis And Findings

The corpus for this study consists of the 30 legal texts that are pertaining to the taxation cases of different sectors and have been under consideration at different forums acquired for the purpose of this study. These texts have been randomly selected from the sources that include both the manual and online sources comprising tax documents, courts cases and other tax related financial reports available online with respect to taxation. The description of the corpus for this study is presented as follows;

Table 01: Corpus for the Study

File Name	Language	No. of texts	No. of Tokens	Supplementary Information
Corpus	English	30	6,94,125	Types: 6510 Lemmas: 5332

Table shows the total size of the corpus selected for the study as 6,94,125 words containing the tax relevant selected documents. Also, there are 6510 types of items with 20 vocabulary types that appear in all 30 texts that form total tax corpus of text considered for the analysis. Likewise, the marginal text items are 2322 which appear in these categories of text comprising tax documents, courts cases and financial reports that are used for the onward analysis through the help of the corpus software LancsBox. The list of top ten keywords occurring in the corpus of this study is presented as follows:

Table 02: List of Top Ten Lexical Items

Sr No.	Keyword	Absolute frequency (Relative frequency)	Dispersion (CV)
1	*Tax*	16312 (106.151)	1.374
2	*Income*	15550 (102.983)	1.287
3	*Taxpayer*	11147 (81.431)	1.229
4	*Ordinance*	10590 (76.033)	1.032
5	*Notice*	9152 (65.365)	0.738
6	*Return*	8953 (55.206)	1.594
7	*Business*	7961 (50.934)	1.421
8	*Record*	6788 (49.861)	1.771
9	*Withholding*	5545 (45.582)	1.651
10	*Penalty*	4642 (39.693)	1.142

The top ten tax-related lexical items existing in the corpus and extracted through the software LancsBox are shown in the table. It contains the terms that frequently occur in the text and the table shows their absolute frequency along with their relative frequency within the text. Also, it shows the CV dispersion of these lexical items that indicates their parallel distribution in the corpus.

The keywords include terms like *tax, return, taxpayer, income and withholding* which are specific to the taxation register. Moreover terms like *notice, ordinance and penalty* are also specific to the legal language that indicates the use of legal terminologies. Also, terms like *business* and *record* are the referring expressions that refer to the business activity and its relevant record maintenance as a prerequisite for the tax related investigations. Overall the above table presents a comprehensive listing of keyword items found in the corpus of this study along with their frequencies across the whole text.

4.1. Classification of Corpus-Based Forensic Linguistic Roles in Tax Evasion Profiling

The arrangement of the corpus based forensic linguistic approach has been classified in to five important categories which consider the role of corpus based forensic linguistic analysis in tax evasion profiling. It also presents the aspect of tax evasion through linguistic trickeries. On the basis of this classification of different linguistic roles, the

framework of corpus based forensic linguistic relevance has been interpreted which is the key to reach to the targeted conclusions of this study.

The table presents this classification of categories which have been categorized for the consideration of corpus based forensic linguistic role with the description of each with respect to its part and onward application in the tax evasion domains as follows;

Table 03: Corpus-Based Forensic Linguistic Tax Evasion Profiling

Classification	Purpose / Role	Methods / Corpus Tools	Application in Tax Evasion Cases
Authorship Atribution	Identification of likely authors of suspicious financial text	Word frequency, Stylometry, POS tagging (e.g., *LancsBox*)	Connecting unidentified tax documents or fake profiles to suspects of tax
Evasive Language Detection	Exposing vague, deceptive, or ambiguous language	Key Word in Context (KWIC)analysis, modality markers, syntactic features	Recognizing linguistic approaches used intentionally to hide financial details
Thematic Keyword Profiling	Detecting frequent themes and doubtful collocation	Keywords extractions, collocation networks (*GraphColl*)	Tracing language associated to tax evading companies, offshore accounts
Sociolinguistic Profiling	Understanding regional, social, or professional backgrounds of suspects	Lexical variations, code-switching, writing distinctions, register analysis	Profiling suspects demographics or professional networks of suspects

Classification	Purpose / Role	Methods / Corpus Tools	Application in Tax Evasion Cases
Legal Discourse Interpretation	Support legal argumentation using linguistic proof	Statistical charts, corpus meditations (*LancsBox*)	Presenting linguistic conclusions in courts of law or tax tribunal

4.1.1. Authorship Attribution

The foremost role in the corpus based analysis in the process of profiling the suspects of tax evasion is the authorship attribution through the investigation of text based on corpus approach. It is an important step in the profiling of suspects involved in the deceptions that considers the frequency of certain words, the phrases constructions, sentence length, the style of writing and other aspects to be able to find out the author of the text. The authorship attribution is the process of identification of the specificity of the style of an author through the investigation of a given text, for instance the tax based communications. The writing style is unique to every writer, so the patterns of writings unique to writers have been considered to identify the writer in the tax related communications. Further in case of taxation process, the authorship attribution helps in the identification of the arrangements of writings that may indicate towards the author of the writing. These arrangements may include how often and in what context a specific term has been used, how constantly repetitions have been made and on what aspects and many other ways to look at which is helpful in profiling the suspects of tax evasion through the use of language tricks. Also, the frequency of words and the phrases contribute to the identification of the author of the text through the analysis of the language used in the tax related communications.

Table 04: Keywords indicating Authorship Attribution

File Name	Keyword	Frequency	Context of Use
Tax file 06	*Exemption*	420	Claim of exemptions from taxes
Tax file 22	*Expense*	368	Overstating expenses to avoid taxability
Tax file 18	*Loss*	326	Showing loss to lessen the taxable income

File Name	Keyword	Frequency	Context of Use
Tax file 16	*Inheritance*	319	Claiming inheritance to avoid legal questioning
Tax file 11	*Gifted*	303	Showing gifted assets to justify investments
Tax file 19	*Asset*	298	Concealing the assets through linguistic tricks
Tax file 14	*Shareholder*	295	Lower the tax liability by showing partnerships
Tax file 15	*Ordinary*	288	Declaring the business activity as of limited nature
Tax file 25	*Mistake*	254	Trying to justify the wrongdoings or defaults
Tax file 20	*Miscellaneous*	222	Creating ambiguousness in the communications

The above table shows some specific keywords used in the tax communications and extracted from corpus giving their frequencies and the context of their use in the corpus. Table indicates the use of those keywords that apparently favor the taxpayer in his stance. This is because the words like *exemption, loss, expense, gifted, mistake* clearly show the aspect of linguistic strategy employed by the taxpayers to get away from the legal actions pertaining to the imposition of tax. Also, the consideration of anonymous texts gives the indication of the shrewdness of the writer that lack the objective clarity and purpose of use rather strategically designed. The investigation of these linguistic aspects can be helpful in the identification of the writer and the possible intent behind its use. Also, the phrase constructions and the stylometry of the text can add to the identification of the writer through the consideration of the recurring patterns of text using the corpus software LancsBox which presents the concordance lines indicating the left and right nodes of the keyword. Also, the part of speech used in the text and the sentence length can also contribute to the authorship identification. The concordance lines extracted through the help of LancsBox present the examples of these keywords with their background use in the corpus;

- On receipt of the legal notice, I sought ***exemption*** from tax under relevant provisions of law (File 02)
- The bifurcation of the ***expenses*** incurred during the year has already been given which show that taxable income has not been acquired (File 11)
- The business has been under huge ***loss*** in the period under consideration, that is

why the tax was not payable by us (File 24)

- The property owned by myself is the ***inheritance*** that I received from my parents and I have not made any investment in the property (File 16)
- The jewelry was ***gifted*** by my father and after selling it to the jeweler I received the amount through which I constructed the plaza (File 19)

The concordance line indicates the use of specific keywords to downplay the aspect of tax payments. Also, the sentence constructions with the background of the use of keywords show the authorship attributes which after investigation linguistically may help in the identification of the author of the text. Thus the corpus based approach can be a tool towards the authorship recognition at a first step in profiling suspects of tax evasion

4.1.2. Evasive Language Detection

The second important step in profiling the suspects of tax evasion is the detection of the evasive language in tax corpus. Corpus based forensic linguistic analysis plays important role of evasive language detection that may also help in the profiling the suspects of the tax evasion in many ways. The language is considered through the help of corpus linguistics software LancsBox which identifies the structures of the language that indicate the aspect of evasive language. The linguistic structures are framed in such a way that show the features of vagueness and deception which can be identified through linguistic analysis. LancsBox can highlight such structures through its features of KWIC (Keyword in Context). For instance words like, *initial investments, unreconciled assets, gifted* etc. apparently convey the meanings that are used in a routinely basis in the tax related communications. On the other hand such terms can be studied in their context of use which may show the aspects of creation of linguistic vagueness and uncertainty through the use of such types of terms in the corpus. In addition to it, the use of modals like *could, might, may* etc. are also employed in the tax discourse to create the kind of ambiguity and vagueness in the communication by not giving the clear picture of events and keeping the other party guessing on the actual state of affairs. Likewise many keywords are also specific to the taxation process which creates evasiveness in the communications. So the application of the corpus based approach through the tool of corpus software LancsBox helps in the identification of the evasive language use in the corpus and this feature may also help in profiling the suspects of tax evasion. Table indicates the use of some specific keywords that indicate the use of vagueness and evasiveness in the tax corpus;

Table 05: Keywords indicating Evasiveness

File Name	Keywords	Frequency	Context of Use
Tax file 20	*Partnership*	478	Strategically shifting responsibility of income
Tax file 11	*Nominal*	401	Creating vagueness about income or asset
Tax file 12	*Miscellaneous*	388	Unexplained income, asset or expenses
Tax file 17	*Discrepancy*	381	Minimizing the default caused
Tax file 13	*Perhaps*	365	Use of hedging device to create uncertainty
Tax file 9	*Depreciation*	315	Claim of tax relief through depreciation
Tax file 14	*Rebate*	302	Lowering the tax liabilities by rebate claim
Tax file 18	*Might*	295	Use of modal to create the tactical ambiguity
Tax file 22	*Deductions*	284	Claim of already paid tax through deductions
Tax file 19	*No, Not, Never*	281	Use of negation markers to disown liabilities

The above table represents the occurrences of the keywords which indicate the aspect of creation of ambiguousness and vagueness in the corpus. These keywords have been extracted using corpus assisted software LancsBox, which also presents the contextual use of these keywords in the text showing element of linguistic deception. In relation to the taxation process, these terms have been employed by the taxpayers strategically in order to lower down the impact of tax liabilities. The contextual and the background study of the use of these terms indicate that their usage in the tax corpus is for the purpose of evasiveness through the use of certain linguistic tricks consisting of specifically vague terms like *partnership, nominal, discrepancy* etc. use of hedging devices such as *perhaps*, the use of modals *might* and also the use of negation markers such as *no, not, never* etc. The concordance lines for some of these terms indicate the context of use in the corpus.

- On the receipt of notice, I had presented my reply that this property is made through *partnership* and I *may be* able to provide the requisitioned record also (File 13)

- I have the *nominal* business stock comprising steel items of worth one million rupees and the per day sale is *approximately* twenty thousand rupees (File 18)
- In response to the legal notice served upon me for previous tax year, I *perhaps* had already provided the *miscellaneous* record in the office (File 09)
- The tax office *might* have misunderstood my business capacity so they served me a legal notice to attend the office (File 14)
- I have *no* other sources of business except this also I have *not* travelled to any foreign country and *never* invested anywhere except this business (File 20)

The concordance lines extracted from the corpus of this study present with the instances of the use of keywords indicating the aspect of linguistic manipulation and evasion. It shows that terms like *partnership* show responsibility shifting and terms like *miscellaneous* and others show ambiguity.

4.1.3. Thematic Profiling of Key Terms

Another key area where the corpus based linguistic profiling can be helpful in the legal domains such as taxation process is the thematic profiling consisting key phrases. This is due to the fact that it represents the recurring patterns and the specific themes consisting key phrases that show the aspect of language based deception and trickeries incorporated by taxpayers in the communications. The corpus software LancsBox identifies these patterns and key themes from the discourse for the onward analysis. These patterns are mostly associated with the language based frauds in the taxation process and assist in the tax evasion. These patterns contain the keywords and phrases that indicate the use of vagueness and ambiguity in the tax communications. The double meanings words have been commonly used in this domain to create the sense of indecisiveness and vagueness. The table represents some of these key phrases associated with the linguistic deception in the tax corpus.

Table 06: *Phrases indicating Linguistic Trickeries*

File Name	Key Expressions	Frequency	Contextual Use
Tax file 02	*Private loans*	359	Concealing the lender of loan
Tax file 01	*Indirect expenses*	341	Undescriptive expenses incurred in business

File Name	Key Expressions	Frequency	Contextual Use
Tax file 09	*Various sources*	326	Non-specified sources of income
Tax file 08	*Certain irregularities*	317	Minimization of violations of law
Tax file 06	*Financial institution*	303	Monitory organization without naming them
Tax file 09	*Clerical mistake*	285	Responsibility shifting to other entities
Tax file 07	*Operational expenses*	246	Overstating the expenses in the business
Tax file 08	*Double taxation*	231	Claiming the levy of tax as double taxation
Tax file 12	*Such investments*	224	Unspecified investments made by someone
Tax file 10	*Foreign remittances*	211	Claiming the assistance through remittances

The table presents the instances of the use of some specific expressions that have been extracted from the corpus using the corpus software LancsBox. The frequencies and the context of their use in the corpus have also been presented in the table. These expressions are the indicators of linguistic trickeries incorporated by the evaders of taxes in their communications. It is evident that the use of these expressions is strategically designed to obscure, misrepresent, conceal, overstate, and minimization of the facts for the purpose of deception in tax related communications. For instance, phrases like *private loans* indicate the intent of concealing the person from whom the loan has been acquired. The term *indirect expenses'* is again strategically employed to obscure the actual detail of expense. The aspect of minimization as a linguistic trick has been shown in the use of term *certain irregularities* where the offense has been presented by minimizing it. Further, the aspect of the responsibility shifting is shown in the table through the use of term *clerical mistake* by presenting the default as a common mistake. Likewise, terms like *operational expenses, double taxation, foreign remittance,* etc. have been employed strategically in the corpus to misrepresent the information and giving an indication of linguistic trickeries for the purpose of evasion of taxes. These key expressions have been extracted from the corpus through the corpus software LancBox that has the ability to analyze the large data sets and is useful in finding out the specific language patterns that are used to make financial frauds through linguistic themes.

- It is stated that I have acquired the *private loans* of 10 million rupees for the business of cash and carry (File 11)
- The cost on the plaza construction was around thirty five million and the same was received from my brother as a *foreign remittance* (File 14)
- Profit earned from *financial institutions* during the year is Rs. 500,000 (File 17)
- The shortcomings in the tax return was not intentional but the *clerical mistakes* (File 19)
- Income acquired from *various sources* has also been declared in tax filings (File 22)

The concordance lines extracted from the corpus indicate the linguistic trickeries incorporated through the use of double meaning words which is a deliberate attempt on part of taxpayers to create confusion or to mislead the tax authorities regarding the actual financial potential of taxpayers. It is evident that phrases like *financial institutions* without indicating the nature of financial institutions and *various sources* without specifying the sources are the indication of linguistic trickery. Likewise, *private loans* without specifying the source are linguistic tactic. So it is one of the important aspect of tracing the tax evasion and profile the concerned involved in it.

4.1.4. Sociolinguistic Profiling

Another key aspect of the corpus based linguistic analysis is the ability of the sociolinguistic profiling of the writer of a text through the analysis of the language used. The role of sociolinguistic profiling is to ascertain the social or professional background of the suspect involved in the tax evasion. This is done through the consideration and investigation of the language patterns that have been used by the individuals. It does not indicate fault but it highlights the linguistic patterns that may suggest the points to help the forensic and investigation relevance in the tax issues. The sociolinguistic profiling is done through the examination of the word choices and lexical variations including the aspect of code switching and code mixing during communications. This sociolinguistic profiling through corpus analysis enables the analysts to identify the social background, social class, age and the level of academic and professional competency of the writer in tax related matters. These recurring themes of language use by the tax evaders is the base of the sociolinguistic profiling of the suspects of tax evasion and this is done through the help of

corpus analysis to reach to some sort of forensic evidence in tax related communications. The sociolinguistic profiling is represented in the following table with the salient features.

Table 07: Featuring Sociolinguistic Profiling

Category	Key Features	Role in Detecting Tax Evasion
Speech Acts	Denial, promise, justify	Can be beneficial in identification of evasion
Politeness	Requesting, hedging,	Often employed to gain favour from tax monitors
Referencing	Nominal, loss, gift	Used as part of making things ambiguous
Jargons	Depreciation, exemption	Making claims being part of financial jargons
euphemism	Optimization, tax planning	Intentional use of euphemism to create vagueness
Sentimental	Guilt, fear, anger, irony	Ploys used as part of linguistic trickeries
Styling	tone, repetition, emphasis	Consideration pattern is helpful in identification

The above table presents the aspects of sociolinguistic profiling of the persons involved in the tax evasion. The category wise key features include the aspects of speech acts, politeness, referencing, jargons, euphemisms, sentiments and the styling of the suspects involved in the process of tax evasion. The corpus based consideration of speech acts enables the researcher to consider the aspects of denials, promise and justifications of the deeds made by the suspects in the taxation process that determines the possibility of intent behind these speech acts. Also, the over politeness shown in the tax related communications through the help of requesting and use of hedging is also another important feature to look into while considering the sociolinguistic profiling. Further, referencing is yet another important feature in sociolinguistic profiling of suspects with the aspects of jargons use and creation of euphemism in the communications intending to mislead the tax authorities regarding the actual financial positions. The consideration of these features through the analysis of the communications can help profile the suspects and their intention behind its use. Besides these features the sentiment and the styling of the suspects can also help determine their vested intention. The frequency of these features can help identify the patterns of the language used in communications. The consideration of these features may give the indication of the linguistic deception in tax matters.

4.1.5. Legal Discourse Interpretation

One of the important categories involved in profiling the suspects through the help of corpus device is the legal discourse interpretation. Legal discourse interpretation enables to identify the ambiguities and vagueness in the language used, and the elements of language deceptions. Also, the legal discourse interprets the power of the tax authority in administering the tax affairs in the country and the responsibilities of the taxpayers in complying with the legal requirements in tax matters. Likewise, it improves the communication between the key participants of the process i.e. taxpayers and the monitoring authorities. Its role is to provide the linguistic evidence extracted from the texts in legal representations that draws the possibility of language based deceit. The linguistic evidence is acquired from the frequency graphs and collocation charts using the corpus software LancsBox.

These findings of the corpus based analysis are presented as linguistic evidences in courts settings and other legal domains. This linguistic evidence may comprise upon the details pertaining to the findings such as a specific expression used in the text and the frequency of its use in the discourse. Also, the consideration of the terminology is carried out that indicate some sort of linguistic manipulation to create the vagueness, uncertainty and deliberate confusion through the use of certain language tactics. Thus the corpus based profiling of the suspects help in legal settings through the interpretation of the legal discourse.

Table 08: Highlighting Legal Discourse Interpretation

Category	Key Features	Contextual Use
Modality	Obligation or permission	Used to seek permissions or obligations
Vagueness	Ambiguity in statements	Creation of uncertainty in communication
Nominalization	Obscure responsibility	Making Indistinctness in matters
Passive	Systematic slowness	Intentional slowing down the process
Contradictions	Explanation shifting	Creation of inconsistencies in communications

Table above shows the key features for the interpretation of legal discourse especially in relation to the taxation process. First of all is the aspect of modality that considers the aspect of possibility, permission or obligation. In relation to taxation the use of modality can often deflect accountability, responsibility or certainty which may be done through modality by softening commitment or responsibility. On the other hand,

vagueness can create indecision, or it may conceal the facts or manipulate them with the intention to avoid or evade while considering the taxation process. It is commonly use aspect of the legal discourse pertaining to the taxation process. Next in list is the feature of nominalization that hides the doer of the action by linguistic manipulation. Through the help of the nominalization the doer of the action tries to distance himself from any wrongdoings committed or in simple words the responsibility is shifted to others in order to get away from the legal proceedings. Passivization is the feature in legal discourse which is the intentional obfuscation in the legal matters such as taxation strategically employed to mislead or disguise the monitoring authorities. The last in the list is the aspect of inconsistencies in the statements and contradictions in explanations. It shows the intent of fabrication and dishonesty in the communications such as taxation. Thus it has become obvious that these features of legal discourse interpretation may be used as the linguistic evidence in courts of law and legal forums such as taxation process. Also, their frequency can point out the intentional linguistic strategical manipulation. Finally the interpretation can be helpful in identification of distinct linguist patterns that can be useful in profiling suspects of tax evasion.

5.Conclusion

Corpus linguistics is a practical approach of language study that is based on the real-world language use rather than supposed. It deals with the language in actual use or the language in work in different spheres of life, as we consider the language use in the taxation process. On the other hand, forensic linguistics is also trending with growing challenges in different fields need the forensic linguistics evidence. So, the role of forensic linguistics in the recent times has grown substantially in aiding different other fields including the field of linguistics. Thus the combination of corpus linguistics with that of forensic linguistics is a steady approach in the language research that can contribute considerably. This was the reason of taking it in to consideration the interdisciplinary approach for the purpose of this study. Corpus based forensic linguistic approach was the operating framework of this study and this has been a successful and effective approach in the linguistic analysis in number of reasons. Corpus based analysis gave more advanced methods of linguistic research in which the corpus software helped the researchers to analyze the datasets with much efficiency and with ultimate speed and accuracy to highlight the prevailing linguistic cues in tax deception in Pakistan. This study has considered the effectiveness of the corpus based forensic linguistics approach in highlighting the financial crimes such as tax evasion based on the communication patterns and features of language. The results of the study have emphasized the usefulness of this approach in which the dealing with large data and acquiring efficient and desired results

with speed can be considered. The identification of linguistic patterns of deception and evasive strategies through this approach has shown the effectiveness of this framework of study. The answer to the research questions based on the analysis of the study lies in the fact that this study approach has been successful in exploring the strategies through certain types and classes of taxpayers in their communications for their vested interests mainly to deceive the tax authorities and evade the taxes. Also, the consideration of these highlighted language based tactics help to profile the suspects involved in tax evasion in the country by considering the linguistic forms to ascertain the possible intent of the taxpayers to evade the taxes through every possible means and methods. The conclusion of the study continues to favor the fact that a taxpayer is always found involved in the execution of different kinds of tax evasion tactics through the linguistic methodology. In this regard the proposed corpus based forensic linguistic approach can be very beneficial in not only highlighting the linguistic means of deception used by the taxpayers to hide their financial positions and evade taxes but more importantly this approach can also aid in profiling the suspects of tax evasion with much speed and accuracy.

References

Ahmed, M. (2021). *Exploring forensic linguistics in the context of tax evasion in Pakistan*.

Austin, J. L. (1962). Speech acts.

Coulthard, M. (1994). *A corpus-based methodology for forensic linguistics*.

Coulthard, M. (2010). *Forensic linguistics: Advances in forensic linguistics and legal communication*. In S. M. Johnson & A. Johnson (Eds.), *The linguistics of law* (pp. 22-35). Routledge.

Coulthard, M., & Johnson, A. (2007). *The role of forensic linguistics in legal contexts: Insights and applications*. Journal of Forensic Linguistics, 14(2), 144-158.

Gibbson, G. (2003). *The role of language in legal contexts: Analysis of tax-related deception*. Journal of Forensic Linguistics, 8(1), 14-23.

Johnson, A., & Wright, A. (2014). *Authorship identification in business emails: A forensic linguistic approach*.

Kredens, K. (2002). *Corpus-based analysis in forensic linguistics: A case study of musicians*.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

MacMenamin, J. (2002). *The forensic stylistics and authorship identification*. Forensic Science International, 129(1), 1-7.

McVay, D. (2006). *Corpus-based analysis and tax evasion in corporate communications*. Journal of Business Communication, 43(1), 45-62.

Nattinger, J. R., & DeCarrio, J. S. (1992). *Pragmatic competence and lexical phrases in forensic linguistics*. *Corpus Linguistics, 9*(1), 1-18.

Shuy, R. W. (1993). *Language and the legal process: A case study approach*. Oxford University Press.

Shuy, R. W. (2005). *The language of deception in tax fraud investigations*. Forensic Linguistics Journal, 13(2), 131-147.

Sinclair, J. (1991). *The idiom principle: Linguistic foundations for corpus-based analysis*.

Sinclair, J. (2004). *Corpus linguistics and forensic studies: A new approach*.

Svartvik, J. (1968). *The Evans statements: A case for forensic linguistics*. Humanities Press.

Turell, M. T. (2001). *Forensic linguistics: An introduction to language in the legal context*. Routledge.