Półrocznik Językoznawczy Tertium. Tertium Linguistic Journal 9 (1) (2024)
www.journal.tertium.edu.pl
https://doi.org/10.7592/Tertium.2024.9.1.295

Authorship Analysis, Social Networks and Catalan. Can we Distinguish X Users by Gender and Age?

Elga Cremades
University of the Balearic Islands, Spain
elga.cremades@uib.cat

Abstract

This paper presents data that can help determine whether factors such as gender or age can become significant in authorship analysis through X in Catalan. Considering the principles of forensic linguistics (in particular, authorship analysis and idiolectal style), 500 publications have been analyzed from a stylistic point of view, focusing on three discursive aspects: specific features of X, pragmatic variables and stylistic variables. Contrary to what some authors have found for English X users (Cicres, 2015), the paper shows that, in Catalan, emoticons, exclamations or letter multiplication are not distinctive features for gender or age. However, elements such as the concatenation of hashtags, the use of links, the intensification of first-person subject pronouns, the use of capital letters or the use suspension points can be meaningful for age —but not for gender. This paper thus constitutes a first step towards finding truly distinctive elements in the use of X in Catalan, even though more studies, with larger corpora, need to be done to confirm these tendencies.

Keywords: authorship analysis, forensic linguistics, social network, Catalan, stylistics.

Streszczenie

Analiza autorstwa, sieci społecznościowe i język kataloński. Czy możemy odróżnić użytkowników X według płci i wieku?

Niniejszy artykuł przedstawia dane, które mogą pomóc w określeniu, czy czynniki takie jak płeć lub wiek mogą być istotne w analizie autorstwa postów publikowanych na X w języku katalońskim. Uwzględniając zasady lingwistyki kryminalistycznej (w szczególności analizę autorstwa i stylu idiolektalnego), przeanalizowano 500 publikacji pod kątem stylistycznym, skupiając się na trzech aspektach dyskursywnych: specyficznych cechach X, zmiennych pragmatycznych i zmiennych stylistycznych. Wbrew temu, co niektórzy autorzy odkryli w przypadku użytkowników X w języku angielskim (Cicres 2015), niniejszy artykuł pokazuje, że w języku katalońskim emotikony, wykrzyknienia czy multiplikacja liter nie są cechami charakterystycznymi dla płci ani wieku. Jednak elementy takie jak łączenie hashtagów, użycie linków, wzmocnienie zaimków osobowych w pierwszej osobie, użycie wielkich liter czy użycie

wielokropków mogą być znaczące dla wieku, ale nie dla płci. Niniejszy artykuł stanowi zatem pierwszy krok w kierunku znalezienia naprawdę charakterystycznych elementów w użyciu X w języku katalońskim, choć potrzebne są dalsze badania z większymi korpusami, aby potwierdzić te tendencje.

Słowa kluczowe: analiza autorstwa, lingwistyka kryminalistyczna, sieć społecznościowa, język kataloński, stylistyka.

1. Introduction: forensic linguistics, authorship analysis and X

1.1. Forensic linguistics and authorship analysis

Forensic linguistics is the scientific discipline at the crossroads of language studies and legal practice, i.e., it is an interdisciplinary area that applies linguistic knowledge and methodologies to legal settings and criminal investigation, providing valuable insights that can significantly impact legal outcomes (Olsson, 2008). As language serves as a primary medium for human interaction and communication, its analysis in legal contexts offers a unique window into the intricacies of criminal activities, legal proceedings, and the interpretation of law itself (Queralt, 2020*a*). The scope of forensic linguistics encompasses several key areas, each contributing to the broader understanding of language in legal contexts. As Cicres and Gavaldà (2014) explain, there are three primary focuses —the analysis of legal language, the study of judicial language, and the use of language for evidential purposes (probative language).

Therefore, one of the fields involved in forensic linguistics is the analysis of legal texts, including statutes, contracts, and other official documents (Queralt, 2020*a*). Forensic linguists work to enhance the clarity and comprehensibility of these texts, ensuring that legal language is accessible and unambiguous. This work is crucial in preventing misinterpretations that could lead to legal disputes or miscarriages of justice. Another significant area is the study of legal discourse, which involves examining spoken language in various legal settings. This includes the analysis of police interviews, courtroom proceedings, and witness testimonies (Cicres & Gavaldà, 2014). By scrutinizing the nuances of spoken interactions, forensic linguists can uncover patterns, inconsistencies, or linguistic markers that may be relevant to legal cases. Finally, forensic linguists are also concerned with offering expert witnesses in court proceedings (Olsson, 2008). In this capacity, they provide professional opinions on language-related issues that are pertinent to legal cases. This may involve authorship attribution, where linguists use their expertise to determine the likely author of a disputed text (Turell, 2011). They also contribute to trademark disputes by analyzing the linguistic similarities or differences between competing brand names or slogans. Additionally, forensic linguists assist in meaning

interpretation, helping courts understand the intended or perceived meaning of specific language use in legal contexts.

It is in this last sub-discipline where authorship analysis is situated, which involves studying the linguistic characteristics of a text to draw conclusions about who wrote it (Queralt, 2014: 37) intending to determine who is most likely to be the author of a text when there are several candidates (authorship determination) or decide whether a written text could have been produced by a specific author (authorship attribution). However, authorship analysis can also be practical for establishing the linguistic profile of the possible author of a disputed text (or several disputed texts) (Cicres, 2015; Giménez & Queralt, 2024).

In this regard, forensic linguists employ a diverse array of analytical methods to examine language evidence. One of these methods is stylistic analysis, a fundamental technique used for authorship attribution, which involves studying a text's unique linguistic features and patterns (Queralt, 2020*b*). Moreover, discourse analysis is applied to examine the structure, content, and context of verbal communications, which is crucial in legal settings (Coulthard et al., 2021). Another method that is often used in forensic linguistics is corpus linguistics, which involves the study of large collections of texts, is another valuable tool in forensic linguistics (Queralt, 2020*a*). By analyzing extensive language databases, linguists can identify patterns and norms in language use, which can be crucial in determining whether a particular usage is standard or atypical. Finally, phonetic analysis is employed in speaker identification, where the acoustic properties of speech are examined to match voice samples, identify speakers or even determine the physiological characteristics of health status of the speaker (Cicres & Gavaldà, 2014).

The methods previously mentioned, which are somehow extracted from various linguistic disciplines, are doubtlessly connected with variationist sociolinguistics. The underlying idea in both fields is that any linguistic production contains features that must necessarily be related to the person who produced it (whether oral or written), such as age, gender, profession, educational level, cultural and religious background, political ideology, geographical origin, physiological characteristics, health status or whether the speaker is a native or non-native speaker of a particular language (Cicres, 2015; Queralt, 2020*a*). That is, any text, whether oral or written, contains linguistic data that can be related to the social or individual aspects of the person who produced it (or the people who produced it). Individual data, however, are not always clear-cut. This is why, in the next section, we will deal with the following concepts: *idiolect* /vs/ *idiolectal style*.

1.2. Idiolectal style and sources of linguistic variation

As we have just explained, one of the axioms of forensic linguistics is that any linguistic production provides information about biological aspects (age, gender, physiological characteristics, or health status—which are mainly observed in oral productions), social aspects (religion, social status, educational level, profession, geographical origin), but also aspects related to the speech act, such as social context, topic, textual genre or interlocutor. In this context, however, some features do not depend on external factors but are the result of individual, usually unconscious, factors based on the possibilities offered by the language. These features, along with dialectal and sociolectal characteristics, constitute what is known as an *idiolect*, which is typically defined as an individual's variety of speech. According to Baldwin (1979), an idiolect is the specific form that any dialect takes in a particular individual, and it is composed of their linguistic preferences regarding lexis, morphosyntax, or phonetics. As Coulthard's explains (2004: 431), "Every native speaker has their own distinct and individual version of the language they speak and write, their own idiolect, and the assumption that this idiolect will manifest itself through distinctive and idiosyncratic choices".

However, to determine how an individual creates their own "version" of all the possibilities offered by the language, Turell (2010) considers that it would be necessary to obtain all the linguistic productions of the speaker throughout their life—and, since that it is impossible, she states that the existence of idiolects cannot be empirically demonstrated. To solve this problem, Nolan (1994) proposed the term *idiolectal style*, which, as explained by Turell (2010), entails a change in perspective by recovering the term style from pragmatics. Although the concept of style can be understood from various approaches, idiolectal style relates to the result of the selections an author makes at a specific moment from the different possibilities offered by the linguistic system (without this having to include all the productions of the individual throughout their life). For Nolan (1994), Turell (2010) and ForensicLab (2013), idiolectal style is the set of traits that characterize an individual's "personal dialect," which can change depending on the textual genre or the stage of life they are in. It relates to how each person uniquely, personally, and distinctively uses the linguistic system: "in the context of forensic text comparison, *idiolectal style* could be defined as the set of options that writers take from the linguistic repertoire available to them as users of a specific language" (Turell, 2010: 217). However, the presence of idiosyncratic elements that depend on individual language uses does not prevent variables from appearing due to the various sources of linguistic variation: geographical origin, ethnicity, age, gender, etc.

All these elements have an important implication, which is that it is possible to find characteristics that reveal the geographical origin of a Catalan speaker in both oral and written productions, as variation can be found in phonetics (neutralization or maintenance of unstressed vowels, articulation of affricates, or devoicing of sibilants), morphosyntax (endings of the 1st, 3rd, 4th, and 5th persons of the present indicative, form of the article, form of the subjunctive, negation particles, etc.), and lexis (especially related to basic vocabulary, as home and food). Similarly, non-native speakers of a language may retain features of their first language at the morphosyntactic level (word order, verb regime, use of prepositions, etc.), spelling (accentuation, use of diacritics, representation of certain sounds, digraphs), ortho-typography (capitalization, punctuation), or stylistic issues (degree of formality, greeting and farewell formulas in letters, etc.), which can reveal their ethnicity (Cremades, 2022).

There is also variation according to age, as different generations of speakers tend to use linguistic forms that distinguish them from previous or subsequent generations, related to trends (consider, for example, the influence of media on speech), cultural references (for instance, references to music or certain traditions that have changed due to globalization), technological changes (both in terms of loss of expressions and incorporation of neologisms), etc. Similarly, various studies (especially following Lakoff, 1973) have shown that men and women maintain different linguistic patterns, although there is controversy around this issue (Serrano, 2014; Argenter, 2020).

On the other hand, it seems that some variables, such as the use of specialized terminology, grammatical, orthographic, or stylistic correctness can be related to the profession, the individual's educational level, and others, such as the use of certain metaphors, with religion (e.g. talking about shepherds, brother's keeper or forbidden fruits) or cultural background (Queralt, 2020*a*).

All this allows forensic linguistics to determine, to some extent, specific characteristics about the authorship of anonymous or disputed texts for which there are no candidates (which can be helpful, for example, for police investigations). In this regard, the paper is related to the search for variables that allow distinguishing, in texts that can be found on a platform like X, the social and biological characteristics of the individuals who produced them, since studies in this field are still scarce, in Catalan.

2. Goals of the study

The main goal of this paper is to determine if characteristics related to age and sex can be distinguished in X posts. Regarding gender, according to Cicres (2015), there are studies on English that show that, on X (formerly Twitter), women use more emoticons, ellipses, words with multiplied vowels (*nooo waaay*), repeated exclamation marks, combinations of punctuation symbols (such as *?!*), the abbreviation *omg*, and onomatopoeic words. In contrast, men's only characteristic is a high usage of *yeah* and *yea*. On the other hand, in relation to age, it is often said that referring to the Internet as an invention rather than as a technology would indicate that the speaker or writer is middle-aged or elderly (Cicres, 2015). Therefore, this paper aims to identify whether these differences also occur in the case of X writers in Catalan.

3. Methodology

3.1. X: main characteristics

Social network X is a microblogging platform that allows users to connect, share, and communicate through short messages formerly known as *tweets* and currently known as *posts*. Even though the number of users has recently decreased, it still has around 176 million active daily users (according to 2024 data), which explains its significant impact on disseminating of news or comments on current events.[1] Catalan language was incorporated to the platform in July 2012, both on the web and in the mobile version.

The platform operates as follows: each user creates a profile with a username, bio, and profile picture. Users can follow others to see their posts in their feed, creating a network of connections. Following someone allows users to receive updates from that account, while users can also be followed by others, expanding their reach. Users can choose their messages' privacy level: they can be public (the default option) or restricted to their circle of followers. For many users, the fact that the messages are public influences the topic and type of publication they usually make, so they tend to restrict themselves to comments on current events or topics of interest to their professional sector or social circles (although there are also users whose posts are more personal in nature).

[1] Source of information: https://www.nbcnews.com/tech/tech-news/fewer-people-using-elon-musks-x-struggles-keep-users-rcna144115

Users can write entries of a maximum length of 280 characters (including spaces) on their profile page, which encourages concise communication. Each entry, which can be accompanied by links or images, might automatically be sent (if the algorithm favors it) to the homepage of people who has chosen to follow that user, making the dissemination of each message instantaneous and automatic. It must be considered that the platform employs algorithms to curate users' feeds, prioritizing posts based on relevance, engagement, and the user's interaction history. This means that not all posts from followed accounts will appear in a user's feed, as the algorithm aims to show only some content.

Users can interact with posts through likes (i.e., clicking on the heart icon that appears under each publication), reposts (i.e., copying the post to their own profile) and replies, fostering engagement and discussions around various topics. Additionally, there are specific elements of the platform: hashtags and mentions. A hashtag is a set of characters preceded by a hash symbol (#) that can be clicked on, used to access content indexed by categories or topics in a microblog, and it is a key feature of many social networks like X. By clicking on a hashtag, users can view all posts associated with that topic. The platform also highlights trending topics, which are the most discussed subjects at any given time, helping users discover popular conversations. Conversely, mentions, on the other hand, are posts that include a user's name and are preceded by the *at* symbol (@). These two elements will be interesting from a forensic perspective, as, as we will see, the use of hashtags or mentions can be relevant for characterizing users.

3.2. The corpus

This study was conducted by analyzing 50 posts in Catalan from 10 different X users; thus, a total of 500 posts were analyzed. On one hand, the group was divided by gender (5 of the studied authors were men and 5 were women) and, on the other hand, by age (three authors over 50 years old, four between 30 and 50, and three under 30 years old). To ensure minimal variation, the chosen speakers share certain characteristics: all of them were native speakers of Eastern Catalan and had college degrees in humanities or social sciences.

Since X profiles can be public, and to follow ethical standards regarding privacy, only authors who did not have protected profiles were sought. Also, for practical reasons and in order to protect their anonymity, authors are identified numerically throughout the paper: H1-5 are the men and D1-5 are the women. Of these, H1-2 and D1 are in the 50 to 65 age range; H3 and D2-4 are between 30 and 49 years old, and H4-5 and D5 are users between 16 and 29 years old. Specifically, H1 is a 53-year-old economist; H2 is a 65-year-old Catalan philologist; H3 is a 33-year-old political scientist (and politician); H4 is a 28-year-old librarian, and H5 is a 20-year-

old history student (though in this case, the posts analyzed start from when the student was 16). Regarding D1, she is a 57-year-old writer and journalist; D2 is a 45-year-old cultural manager; D3 is a 45-year-old social sciences professor; D4 is a 33-year-old Catalan philologist, and D5 is a 28-year-old linguist. Table 1 contains a summary of the selected users' information.

Table 1. Characteristics of informants

	Gender	Age	Profession
H1	Male	53	Economist
H2	Male	65	Catalan philologist
H3	Male	33	Political scientist
H4	Male	28	Librarian
H5	Male	20	History student
D1	Female	57	Writer / journalist
D2	Female	45	Cultural manager
D3	Female	45	Social sciences professor
D4	Female	33	Catalan philologist
D5	Female	28	Linguist

A stylistic approach was used to perform the analyses and extract potentially significant features. The following variables were considered, following the research project "Forensic Idiolectometry and Idiolectal Similarity Index" (Forensiclab-Linguistic Variation Unit, 2013):

- Pragmatic-discursive
- Ortho-typographic variables based
- Issues specifically related to the use of the platform X: presence of hashtags (preceded by hash symbols, i.e., #), mentions (preceded by the @ symbol), or the shortening of words due to the character limit (only 280 characters per post).

Morphosyntactic or lexical variables were not considered because these issues depend, at least in most Catalan texts, on the geographical area of origin or the level of linguistic competence (related to the level of studies or profession, which expands vocabulary), which is not the focus of this study. Although, as Cicres (2015) states, in English, men use more determiners and prepositions while women use more personal pronouns, auxiliary verbs, and conjunctions, an empirical demonstration of these characteristics for Catalan would require a corpus of thousands of productions and sufficient resources to conduct a stylometric analysis. This goes beyond the scope of this work and, therefore, was not considered. For all the variables, statistical analyses were performed using the program SPSS Statistics 21 (2012).

4. Results

As specified in the previous section, the analysis was conducted considering pragmatic and stylistic (or ortho-typographic) variables, as well as issues that, while strictly pragmatic, have been distinguished because they are directly related to the medium from which the texts (X) were extracted. Below, the specific variables included in each section are described in more detail.

4.1. Specific variables of X

Regarding the specific variables of X, the following variables were analyzed: (1) average number of words per post, (2) average number of characters per post, (3) number and type of emoticons, (4) number and chaining of hashtags, (5) incorporation of links (number and type), (6) number of mentions, (7) types of abbreviations used, and (8) letter multiplication for expressive purposes.

4.1.1. Number of words and characters per post

Although the number of characters is limited to 280 per post, not all posts use the maximum allowed, so it was considered that the average number of words could help distinguish age or gender. However, the data indicate precisely the opposite: neither the number of characters nor the number of words is distinctive of age or gender. Thus, although it seems that, in general, women write fewer words than men, an ANOVA test performed with IBM's SPSS Statistics 21 (2012) shows that the differences are not significant. Therefore, we can affirm that it is not a variable to consider when determining the author's gender or age. Table 2 shows the average number of words and characters per post for each speaker, showing that even though younger men produce less words than older men, the tendency is not the same among women—indeed, in the case of women the tendency is the opposite. Tables 3 and 4 contain the ANOVA tests showing that the differences insignificant—results show that significance is over 0.05 in all the cases.

Table 2. Average number of characters and words per post

Variable	H1	H2	H3	H4	H5	D1	D2	D3	D4	D5
Words	19.2	16.6	20.9	11.7	11	15.8	8.6	12	13.7	13.4
Characters	118.7	107.8	123	106.1	72.8	121	85.2	110.2	121	103.9

Table 3. ANOVA test for age

Variables		Sum of squares	df	Mean square	F	Sig.
Words	Between groups	41.642	2	20.821	1.618	.264
	Within groups	90.067	7	12.867		
	Total	131.709	9			
Characters	Between groups	776.033	2	388.016	1.616	.265
	Within groups	1681.283	7	240.183		
	Total	2457.316	9			

Table 4. ANOVA test for gender

Variables		Sum of squares	df	Mean square	F	Sig.
Words	Between groups	25.281	1	25.281	1.900	.205
	Within groups	106.428	8	13.304		
	Total	131.709	9			
Characters	Between groups	14.400	1	14.400	.047	.834
	Within groups	2442.916	8	305.365		
	Total	2457.316	9			

4.1.2. Number and types of emoticons

An emoticon (or smiley) is a graphic symbol made with ASCII characters that represents a human face with an expression, often used in email and text messages, or posts on social networks such as Instagram, TikTok or X. Cambridge dictionary defines it as "a combination of symbols used in emails, text messages, and on the Internet to show the writer's feelings. For example :-) represents a smiling face to indicate that you are happy or have made a joke".[2]

According to Cicres (2015), in English, women tend to use more emoticons than men, making the use of emoticons a potentially useful variable for distinguishing the author's gender. However, this is not the case in Catalan. Both men and women use emoticons (although not everyone does) and they all alternate between using ASCII characters (i.e., punctuation) and images. It appears that men under 50 use them more than women, although the differences are not significant. This arbitrary distribution can be observed in Table 5, which includes the number and type of emoticons used by the different authors.

[2] The definition can be found here: <https://dictionary.cambridge.org/dictionary/english-spanish/emoticon> (Date: 15/10/2024)

Table 5. Number and type of emoticons

Variable		H3	H4	H5	D1	D2	D5
Number		5	5	6	4	3	2
Type	ASCII characters	;-) / :-) / xD	:-P/ :3 / xD	:D / :) / :(	-	:))) / ;-) / :))	-
	Images	-	👩🐵	-	🙂		💙💚

4.1.3. Hashtags

Although almost all users of X use hashtags, either as a tool for classifying their posts or to help contextualize their messages, not everyone uses them in the same way or in the same quantity. Thus, even though the number of hashtags itself is not significant by age or gender, as shown by the results of the ANOVA tests (Tables 7 and 8), the use of chained hashtags seems to be distinctive by age, as only those under 50 use them. Table 6 clearly shows that only younger users resort to chaining hashtags—indeed, H1 and H2, who are older than H3, H4 and H5, do not resort to chaining hashtags at all, as shown in the table.

Table 6. Hashtag use

Variable	Number	Chaining
H1	4	-
H2	14	-
H3	19	#independència #sàhara lliure #NO #OXI #CIEalParlament #CIEsNO
H4	67	#trepat #Barberà #enoturisme #revetlla #santjoan #verbena #calanaiara #montgat #maresme #plagi #llibres #PutaRenfe #trolladies #Eurovision #ITA #Eurovision #GER #arxius #museus #bcnesmuseu #nitdelsmuseus #FredericMares #Barcelona #catalunya #seat #133 #radiocassette #barcelona #arxiu #audiovisual #semanasanta #sevilla #resurrección #bibliotecari #bibliotecaria #biblioteques #bibliotecas #aniversari #amics #paella #diumenge #pispetitmoltagent
H5	32	#niparenifill #NovolemRei #CatalàALaFranja #SomPaïsosCatalans #20AnysSenseGuillemAgullo #LaLluitaContinua #NoPassaran #VISCAELBARÇA #FORÇABARÇA #sompaïsoscatalans #prou #euskalherria #London #USA #TotsSomÀlexFàbregas #freedomcatalonia
D1	12	-
D2	28	#cultura #administración #capitalismo #concursigerstgn #mediterràniament #ElBatecdelTemps #patrimoni

D3	67	#Cambrils #Juntsimillor #27sOnTotComenca #cambrils #maredeDéudelCarme #bonareflexió #tenimpressa #tarimaCambrils #bonamúsica #bonamúsica #tarimaCambrils #Cambrils #concerts #creatividad #SantPere #Cambrils #Cambrils #stjoan #celebracions #dinarsdEstiu #familia #qüestiómental #samarretesInde #orgullodemadre #ironia #ComissioFrau #puigcerver #FamíliaEscola #ESO #fentEscola
D4	3	-
D5	29	#MarinaAlta #PaísValencià #Estiu #Summer

Table 7. ANOVA for age and number of hashtags

Variable		Sum of squares	df	Mean square	F	Sig.
Hashtags	Between groups	1312.500	1	1312.500	3.019	.120
	Within groups	3478.000	8	434.750		
	Total	4790.500	9			

Table 8. ANOVA for gender and number of hashtags

Variable		Sum of squares	df	Mean square	F	Sig.
Hashtags	Between groups	.900	1	.900	.002	.970
	Within groups	4789.600	8	598.700		
	Total	4790.500	9			

4.1.4. Links

Posts can include various types of links: some users tend to link to newspaper articles on their profile page to highlight certain aspects; others focus their X comments on images they have posted on another platform (Instagram); some share their writings posted on blogs; and others, when referring to another post, link to it instead of reposting it. This diversity causes the number of links to vary between users, prompting an examination of whether the differences were significant enough to distinguish users by gender or age. While the type of link does not vary much between users (as shown in Table 9, there are five possibilities with no group showing a defined trend), it is significant that women tend to include more links than men. Thus, Table 10 shows that the gender variable, according to ANOVA tests, has a significance level below 0.05 (marked in bold).

Table 9. Links

Links		H1	H2	H3	H4	H5	D1	D2	D3	D4	D5
Number		6	7	1	19	5	14	43	21	50	20
Type	Press articles	x	x	-	x	x	x	x	x	x	x
	Blog entries	x	x	-	-	-	-	x	-	x	x
	Instagram pictures	-	-	-	x	x	-	x	x	-	x
	Other posts	x	x	x	-	-	x	-	-	-	-
	Videos	-	-	-	-	-	-	-	-	x	-

Table 10. ANOVA test of gender and presence of links

Variable		Sum of squares	df	Mean square	F	Sig.
Links	Between groups	1210.000	1	1210.000	8.145	**.021**
	Within groups	1188.400	8	148.550		
	Total	2398.400	9			

Table 11. ANOVA test for age and presence of links

Variable		Sum of squares	df	Mean square	F	Sig.
Links	Between groups	394.971	1	394.971	1.577	.245
	Within groups	2003.429	8	250.429		
	Total	2398.400	9			

4.1.5. Abbreviations

Texts from social networks, instant messaging, or text messages have promoted the shortening of words when writing, either due to the need for immediacy (as in the case of instant messaging) or due to character limitations (as with former SMS or X). The type of abbreviation used can be significant in distinguishing an author (Grant, 2010), but in the messages studied, there do not appear to be significant differences due to age or gender. Thus, abbreviations like *q* (que, 'what'), *pq* (perquè, 'why'), *tb* (també, 'also'), *or d* (de, 'of') seem quite widespread in Catalan, while less frequent ones (such as *pel·li, esp, cat, loc. conj., sv, IB, Gencat*) tend to follow "known" rules for this type of resource, such as vowel suppression, use of numbers, truncation, or standardized abbreviations (*loc. conj*. for *locució conjunctiva* 'conjunctive phrase', *sv* for *sub voce*). [3] Table 12 lists the abbreviations used.[4]

[3] They can be found here: <http://bit.ly/1JkLVXo> (Date: 01/08/2024).
[4] There is one informant (D2) who does not use abbreviations, and therefore is not included in the table.

Table 12. Abbreviations

	q/ pq	d/dl/dls	Numbers	Other
H1	**pq** m'has bloquejat el **q** ha passat La raó per la **q** no... sí **q** han analitzat Veig **q** estàs Suposo **q** Veig **q** m'has... testimonia **q** sou **q** va presentar ahir	-	el Barça pujaria a l'**1** del ranking Forbes és la **2a** més gran	serà reconeguda **x** la UE **pspost** de la Gene
H2	canviar statu **q**	-	sobretot el **2**	A **Cat** Radio Cal requisit de **cat** **Esp** és requisit **loc. conj.** **sv** tema / fer / cobrar
H3	-	El referèndum **d** diumenge	tinguin **2/3** o més tinguin **1/3** o menys calen **2/3** per reforma hi va haver **5** detinguts	**x** gravar reprova Pujol **x** diners que a **Cat** tinguin **Pdr Snchz** al PSOE nous impostos **gencat** bombolla **immo**
H4	-	part **dl** jurat	és el **1** cop que	jurat **mb** @Aina acabant a la **expo**
H5	Article **q** he comentat	article **d** @revistacambrils **d...** l'ús **dls** mitjans	Llegint **1** article	**x** la hipocresia ...he de pujar a **Barna**.
D1	**q** la vida és... el temps **q**... diu **q** no cal anar **q** no se sentiria trencats **q** quan... força **q** té una idea **...q** diu	el trencadís **d** Gaudí	-	Al meu **art**: "La..." Himne **P.** Valencià
D3	els **q** no comparteixen	-	**3** regidors **3** mesos de..	**tb** els q...
D4	Lamentable **q**... atura el seu fill **pq**... pensar **q** això... he vist **q** no... Feia molt **q** no... Denuncia **q**... Segur **q** el **q** diu	l'Inst. Interuniv. **d** Filologia Valenciana	-	**Inst. Interuniv.** d Filologia Valenciana Una **pel·li** espectacular
D5	ja se sap **q**	tracta **d** modes	ells són els **1rs**	A França **tb** ha arribat... **dp** del documental **BCN**-París **OMG** T'animes o **RT**?

4.1.6. Letter multiplication

Finally, as Cicres (2015) states, studies in English have shown that women tend to multiplied vowels more than men when writing on X. However, this is not applicable to Catalan. In the analyzed posts, five contained such multiplications, and two of the Xers who used them were, in fact, men, as shown in Table 13.

Table 13. Letter multiplication on X

H3	H5	D2
una miiiiiica	Twiteraa / twitteroo / ehh	Sorpreeeesaaa!!!

In summary, regarding the specifics of X platform, not all features seem to play a role in distinguishing between authors by gender or age. For instance, the number of words or characters in a post, as well as the use of emoticons, hashtags, abbreviations, or letter multiplication do not show any significant discriminatory power. However, two variables stand out: the chaining of hashtags, which appears to be a tactic used primarily by those under 50, and the prevalence of links, which are shared significantly more by women compared to men.

4.2. Variables related to the pragmatic-discourse module

In this module, the following variables were studied: (1) response to an open question, (2) discourse markers, (3) morphological or lexical intensification of the statement, and (4) intensification of the first-person subject. Although the scarcity of data has prevented verification of all items, it seems that, in general, these issues tend to be idiosyncratic, making them useful in a forensic comparison of texts but not for determining the gender or age of the individual.

4.2.1. Answer to an open question

When asked a question "How do you see Maria?", the response can be formulated in various ways: it can be headed by a discourse marker like *well* (e.g., "Well, I see her well, more cheerful"); it can be introduced by repeating part of the question (e.g., "How do I see her? More cheerful"); it can be formulated as an exclamation (e.g., "Great!"); or it can be answered directly ("Good, more cheerful"). Although it is not common to maintain question-response dialogues on X, it is frequent for users to comment on or respond to other users' posts, thus allowing observation of the various strategies employed. Although we did not obtain data from all users, there does not appear to be a distinction between genders or ages; rather, it is more of an

idiosyncratic issue. Most responses are direct: only H1 uses discourse markers, H1 and H4 use exclamations, and only D5 resorts to questions. This can be verified in Table 14.

Table 14. Answers

Type	H1	H2	H4	D4	H5
Discourse marker	Doncs ara que ho menciones... (‘So now that you mention it…’)	-	-	-	-
Question	-	-	-	-	Really? Vols un entrepà càlid? (‘Really? Do you want a warm sandwich?’)
Exclamation	Bingo!!!	-	Quanta raó! (‘So right!’) Gran! (‘Great!’)	-	-
Direct answer	No, company, no. (‘No, colleague, no.’)	Ben d’acord. (‘Totally agree’)	-	I el caló català, un patrimoni... (‘And Catalan *caló,* a heritage…’)	Ok, jo faré brunch català... (‘OK, I am going to do Catalan brunch’.)

4.2.2. Discourse markers

When distinguishing speakers, the use of discourse markers such as “no?” (‘ain’t it?’), “oi?” (‘right?’), “eh?” can also be significant. However, in the analysed posts, only H1 includes one of these markers (“no Alfons?”, ‘ain’t it, Alfons?’). This indicates that it is an idiosyncratic issue (i.e., related to the idiolect) and therefore does not allow for distinguishing gender or age, but rather discriminates individual speakers.

Related to the previous discourse markers, the presence of interjections like *ep* or *ei* has been observed, and it has been noted that they are more frequent among speakers over 40 years old. For instance, only H1, H2, D1, and D2 contain interjections of this type. However, more data would be needed to confirm this trend.

4.2.3. Utterance intensification

Catalan has various alternatives for intensifying an utterance or a word (for instance, noun or adjective). This can be achieved through morphological resources (such as the addition of derivational affixes—suffixes or prefixes), syntactic resources (introduction of quantifiers or intensifiers), a lexicosyntactic resource such as word repetition, or an ortho-typographic resource like syllabic separation of the word. In the texts analyzed, most authors employ syntactic intensification. As shown in Table 15, only two use morphological intensification; one uses the ortho-typographic method, and none repeat words to intensify the statement. This suggests that it is likely an idiolectal phenomenon or, in any case, related to factors not considered in this analysis (such as socioeconomic status, geographical origin, or professional field). However, women show a more frequent use of the qualifier *gran* (‘big’) than men, which could likely be confirmed with a larger corpus.

Table 15. Utterance intensification resources

Type	Morphological	Syntactical	Ortho-typographical
H1	-	una mica rata la més gran i patètica de totes... molt nerviós	-
H2	-	molt bé	-
H3	Superemocionant. problemón	-	-
H4	redimonis	ben animat	-
D1	-	grans líders grans noms propis gran somni tan elegants tan nou, tan democràticament radical i tan agosarat com...	-
D3	-	gran mestre	-
D4	-	més bonic i més humà	Es-pec-ta-cu-lar
D5	-	gran article	-

4.2.4. Intensification of the first-person subject

Although X is a platform for personal microblogs, profiles are often public. This leads to many posts being dedicated to commenting on current events rather than showcasing individuals’ personal experiences. Nevertheless, some more personal comments allow for an examination of first-person usage. In this regard, according to the analyzed posts, only those under 30 tend to intensify the first-person subject by explicitly using the personal pronoun *jo* (‘I’). This contrasts with the trends observed for Spanish by Carvalho, Orozco and Lapidus Shin (2015), although a

larger data corpus would be necessary to confirm these indications. The obtained data are summarized in Table 16.

Table 16. Intensification of the first-person subject

Variant	H1	H2	H4	H5	D4	D5
Jo	-	-	**Jo** li diria millor...	...colors que **jo** sempre porto al cor **jo** no sé què faré	-	Ok, **jo** faré brunch
Ø	Sé llegir... Celebro que... Hi estic totalment d'acord. Contactaré Veig que...	Diria que... He vist que...	Em sento orgullós	domino una nova llengua he de pujar ja et dire algo	...vaig pensar	...no hauria sabut on aparcar.

As an interim conclusion, a general look at the pragmatic-discursive module shows the findings are quite revealing. Neither the way respondents answer open questions, nor their use of discourse markers or intensification strategies seems to effectively differentiate speakers by gender or age. That said, there is a hint that younger speakers, particularly those under 30, might be more inclined to intensify the first-person subject, though we would need more data to verify this observation.

4.3. Variables related to the stylistic or ortho-typographic module

Regarding the stylistic module, we have observed the following variables: (1) graphical accentuation, (2) use of ellipses, (3) presence of distinctive and intensifying capital letters, (4) use of quotation marks, (5) question marks and exclamation points, (6) combination of punctuation marks, and (7) marking of parenthetical remarks. As with the previous module, not all these variables are distinctive of age or gender, although some can be considered relevant from an idiolectal perspective.

4.3.1. Graphic accentuation

Concerning graphic accentuation, it should be noted that most users of X systematically accentuate words that carry accents—although some make errors. This systematicity may be due to mobile devices that facilitate the accentuation of certain words through autocorrect features. However, the irregularity of accentuation by the only X user under 20 years old considered in this analysis is notable. For instance, H5 accentuated words like *concentració*

(‘concentration’), *més* (‘more’), *perquè* (‘because’), *raó* (‘reason’), *solució* (‘solution’), *ús* (‘usage’), *comunicació* (‘communication’), *públic(s)* (‘public’) , *és* (in some cases) (‘is’), *pròpia* (‘own’), *acció* (‘action’), *revolució* (‘revolution’), *parlarà* (‘s/he will talk’), *però* (‘but’), *espontània* (‘spontaneous’), *són* (third person plural of the present indicative of *ser,* ‘be’), *així* (‘like that’), *imposició* (‘imposition’), and *direcció* (‘direction’). Conversely, the words *despres* (‘after’), *cules* (‘Barça supporters’), *es* (is), *dema* (‘tomorrow’), *classic* (‘classic’), *dire* (first person singular future indicative of *dir,* ‘say’), *mon* (‘world’), *que* (‘what’), *se* (first person singular present indicative of *saber,* know), *fare* (first person singular future indicative of *fer,* ‘do’), *tambe* (‘also’) and *nomes* (‘only’) were not accentuated. This could indicate a difference between authors over 20 years old and those under (i.e., adolescents), but more data would be needed to confirm this trend.

4.3.2. Ellipses

To indicate that a list is incomplete, ellipses or *etc*. (abbreviation for *et cetera*, which literally means ‘and the rest’) can be used. In the analyzed texts, authors quite systematically resort to the former; no latter occurrences were found. However, ellipses can serve a dual purpose: on one hand, they can indicate the incompleteness of a list, as explained; on the other hand, they can be expressive, simply indicating that the sentence’s meaning is suspended. The use of this latter alternative is quite frequent on X, although there are differences in use among Xers, as shown in Table 17. Although it may seem more common among women than men, ANOVA tests (based on the number of sentences with ellipses from each speaker) show that the differences are insignificant enough to constitute reliable evidence (Tables 18 and 19).

Table 17. Ellipses

Variant	Expressive	Incomplete list indicator
H1	[...] enormement irritant... No sé. [...] a unes... plebiscitàries!	-
H3	-	[...]bombolla immo, cohabitació públic-privat...
H4	El fil de l'#humorbibliotecari d'avui està ben animat... La mar està alterada... La #Conca estava trepatejant... @rodalies en silenci... [...] un capítol anomenat "Ales al vent"... [...] és així tòpicament mexicana... Ja anava tocant des del 2009 reformant-se...	-
D1	No ho hauria dit mai... L'enllaç per apuntar-se a la Diada... Estic impressionada amb tanta categoria intel·lectual... Mireu quins carinyos m'envien... Sempre tan elegants... Les sis passes del camí cap a Ítaca... I... Guardiola... Estan dels nervis... Que no és calb...	-
D2	L'estiu ja és aquí... Per... Ara fa uns dies... La Primavera Cultural dels...	-
D3	[...]et feliciti pel Facebook enlloc de trucar-te...és preocupant Doncs, això...	[...] respecte a l'altre, treball en equip, flexibilitat...
D4	Total, per un euro...	-
D5	[...] no hauria sabut on aparcar... Poder tornar a sentir-li la veu... [...], els cap més d'una #llengua a l'etiqueta... [...], things still merge... [...] el cartell de la Policia espanyola... [...] diferents maneres d'entendre el respecte... Per les llengües. [...] la meva escola a prop... [...] més d'1 setmana... OMG!	-

Table 18. ANOVA test for ellipses and gender

Variable		Sum of squares	df	Mean square	F	Sig.
Expressive ellipses	Between groups	22.500	1	22.500	2.055	.190
	Within groups	87.600	8	10.950		
	Total	110.100	9			

Table 19. ANOVA test for ellipses and age

Variable		Sum of squares	df	Mean square	F	Sig.
Expressive ellipses	Between groups	18.683	2	9.342	.715	.522
	Within groups	91.417	7	13.060		
	Total	110.100	9			

4.3.3. Capital letters

Regarding the use of capital letters, it is essential to distinguish between distinctive capitals (those that are needed) and those used to intensify the statement. In the case of distinctive capitals, it should be noted that only H5 omits them in several instances where they would be necessary, such as *Tomić* (Ante Tomić, a basketball player), *twitter*, *xavi* (the former Barça coach), *ari* (for the name *Ariadna*) or *eclipse* (the name of a song). In contrast, D5, who is under 30 years old, only omits it once (*mercadona*, a supermarket chain). These omissions could indicate a difference between younger and older writers, although more data would be needed to confirm this. As for intensifying capitals, it seems that only older and younger speakers use them. They are present in H1, H2, and D1 on one hand, and in H5 and D5 on the other. However, they appear to be much more frequent among the older speakers, and an ANOVA test shows that the difference is significant (Table 20).

Table 20. ANOVA test for age and intensifying capital letters

Variable		Sum of squares	df	Mean square	F	Sig.
Intensifying capital letters	Between groups	140.667	2	70.333	6.902	.022
	Within groups	71.333	7	10.190		
	Total	212.000	9			

4.3.4. Quotation marks

Quotation marks are used on X both for quoting other authors and for usages that would normally be in italics in Catalan, such as titles of books or movies and metalinguistic uses, given that the platform does not support italics. The data show that the most widely used quotation marks are, double high quotation marks. There is one author, H2, who also uses single and angular quotation marks, and others who either use single (D1, D2, and D5) in addition to double, or angular (H4 and D4).

It seems, however, to be an idiosyncratic issue, characteristic of everyone's usage (and their knowledge of ortho-typography), and is not significant regarding sex or age. Certainly, the

youngest X user, H5, does not include any type of quotation marks, but this could also be due to their competence in citing titles or others' words. More data would be needed to ensure whether there is a distinction in this regard.

4.3.5. Exclamation and question marks

Regarding exclamation and question marks, in Catalan, some choose to include them only at the end (the majority option, recommended by the IEC), while others write them both at the beginning and the end of the interrogative or exclamatory sentence (an option recommended by, for example, *Ortotipografia* by Pujol & Solà, 1999).

It should be noted that, likely for practical reasons, all analyzed authors write the mentioned punctuation marks only at the end of sentences, but they vary in the number of symbols used in each case. This was observed to see if it was a "regular" variation or simply idiolectal. According to Cicres (2015), in English, it has been observed that women on social networks tend to use more double or triple punctuation marks and more combinations like *?!*, while men usually restrict themselves to one punctuation mark per sentence.

However, this regularity does not seem to be present in Catalan. Double and triple marks (for instance, *!!!* or *???*) have been found in both men and women of all ages, although they are much less frequent than single marks (i.e., *!* and *?*), and no combinations of marks were observed in any of the texts. This could be partly due to the character limit imposed by X, and therefore, it would be necessary to check if individuals' behavior on other virtual platforms or social networks is similar.

4.3.6. Parentheses

Finally, both the number of parentheses contained in the various texts and the punctuation marks used to mark them were observed. In this case, it was found that there is a difference between men and women: men include more parentheses in their posts than women. Only D4 includes parentheses in their messages, while all the men do. Although the ANOVA test indicates that there are no significant results in this regard, it is likely due to the insufficient data available. As for the punctuation marks used, parentheses are usually employed: only one occurrence of a parenthetical remark limited by dashes was found, in H3, as shown in Table 21.

Table 21. Parentheses

Types	Parenthesis	Dashes
H1	(Twitter?) Xavier (Tejedor) Rius @CDC_Vila_Seca (15) ERC a favor (2) Cs i PP no (2) Als estudiants (i a mi) Tots els polítics catalans imputats (Bustos, Oriol Pujol, etc.)	-
H2	(@pauetvidal) (aquest fatídic "cap" amb no comptables") (tuitejar, en bàrbar) (la gran falta universal) (sv tema) (sv fer) (sv cobrar) (dos anglicismes passats per l'esp, sobretot el 2) (vegeu tweet precedent) (no demanen certificats) (recorrerà contra) (al bar) (ARA) Grup(úscul)	-
H3	La comoditat (retroalimentació) Els pobles (del sud d'Europa en especial) no tenim... ...a tots nivells (democràtic, social i nacional)	...combat -a mort- entre l'Europa dels pobles...
H4	#Bellprat (Anoia) A Rússia ni aigua (encara que la canó mola)	-
H5	el #LAPAO (llengua aragonesa pròpia del #aragóoriental)	-
D4	...de deixar anar segons quines crítiques (= de disparar-se al peu) Sobre la (no) utilitat d'una llengua	-

Regarding stylistic features, therefore, the only strong takeaway is the use of capital letters for emphasis, which appears more frequently among authors over 50. Other patterns emerged as well: women tend to use ellipses more than men, younger users are more likely to skip accents, and men often employ parentheses to indicate parenthetical remarks more than women do. However, it is worth emphasizing that these findings are preliminary. Analyzing a larger dataset could provide a clearer picture and help validate these emerging trends.

5. Conclusion

This study presents an analysis of various variables that can be examined in a forensic comparison of texts, specifically to assess their potential for distinguishing the gender or age of

speakers. The examination of posts from platform X focused on three key aspects: specific platform elements, pragmatic variables, and stylistic features.

The results indicate that not all platform-specific features are effective in differentiating authors by gender or age. Metrics such as word and character count, emoticon usage, and abbreviations did not yield significant distinctions. However, two notable exceptions emerged: the chaining of hashtags, which was predominantly utilized by users under 50, and the higher incidence of links shared by women compared to men.

In terms of pragmatic-discursive features, our findings suggest that conventional markers, such as responses to open questions and discourse markers, do not significantly differentiate speakers by gender or age. Interestingly, younger individuals (under 30) appear to intensify the first-person subject more frequently, though further research is needed to confirm this pattern.

Finally, the stylistic analysis revealed that the use of capital letters for emphasis is more common among authors over 50. Additionally, trends were identified where women tend to use ellipses more, younger users frequently omit accents, and men are more likely to use parentheses for parenthetical remarks.

Overall, while this study provides valuable insights into the relationship between textual features and speaker characteristics, it also highlights the need for further research involving larger datasets to corroborate these trends and deepen our understanding of forensic text analysis.

References

Argenter, Joan A. (2020) "Gènere i substitució lingüística: hipòtesis i etnografies." *Revista d'Etnologia de Catalunya,* 45; 186-200.

Baldwin John R. (1979) "Phonetics and Speaker Identification." *Medicine, Science and the Law,* 19(4); 231-232.

Carvalho, Ana M., Rafael Orozco, Naomi Lapidus Shin (eds) (2015) *Subject Pronoun Expression in Spanish: a Cross-Dialectal Perspective*. Washington, D.C.: Georgetown University Press.

Cicres, Jordi (2015) "Curso de postgrado en determinación/atribución de autoría: presentación introductoria" [didactic material], Barcelona: IDEC-Universitat Pompeu Fabra.

Cicres, Jordi, Núria Gavaldà (2014) "La lingüística forense: la llengua com a evidència." *Revista de Llengua i Dret,* 61; 60-71.

Coulthard, Malcolm (2004). "Author Identification, Idiolect and Linguistic Uniqueness." *Applied Linguistics, 25* (4); 431-447.

Coulthard, Malcolm, Alison May, Rui Sousa-Silva (eds) (2021) *The Routledge Handbook of Forensic Linguistics*. London: Routledge.

Cremades, Elga (2022) "Expressió escrita i català com a llengua addicional: anàlisi d'errors en parlants de llengües eslaves." *Resercle,* 3; 50-69.

Forensiclab - Unitat de Variació Lingüística (2013) "Idiolectometria forense i Índex de Similitud Idiolectal" [research project]. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Gavaldà, Núria (2011) "Sociolingüística de la variació i lingüística forense." *Llengua, Societat i Comunicació,* 9; 49-59.

Giménez García, Roser, Sheila Queralt (2024) "Grounds for Exemption from Criminal Liability? How Forensic Linguistics Can Contribute to Terrorism Trials." *International Journal for the Semiotics of Law,* 37; 623–646.

Grant, Tim (2010) "Txt 4n6: Idiolect Free Authorship Analysis?" [In:] Malcolm Coulthard, Alison Johnson (ed.) *The Routledge Handbook of Forensic Linguistics,* London and New York: Routledge; 508-522.

Lakoff, Robin (1973) "Language and Women's Place." *Language in Society,* 2 (1); 45-80.

Nolan, Francis (1994) "Auditory and Acoustic Analysis in Speaker Recognition." [In:] John Gibbons (ed.) *Language and the Law*. London/New York: Longman; 326-345.

Olsson, John (2008) *Forensic Linguistics (2nd ed.).* London and New York: Continuum.

Queralt, Sheila (2014) "Acerca de la prueba lingüística en atribución de autoría hoy." *Revista de Llengua i Dret,* 62; 35-48.

Queralt, Sheila (2015) "Curso de postgrado en determinación/atribución de autoría" [didactic material], Barcelona: IDEC-Universitat Pompeu Fabra.

Queralt, Sheila (2020*a*) *Atrapados por la lengua*. Madrid: Larousse.

Queralt, S. (2020*b*) El uso de recursos tecnológicos en lingüística forense. *Pragmalingüística* 28;212-237.

Serrano, M. José (2014) "El rol de la variable *sexo* o *género* en sociolingüística." *Boletín de Filología,* 53; 175-192.

Turell, M. Teresa (2010) "The Use of Textual, Grammatical and Sociolinguistic Evidence in Forensic Text Comparison." *The International Journal of Speech, Language and the Law,* 17 (2); 211-250.

Turell, M. Teresa (2011). “La tasca del lingüista detectiu en casos de detecció de plagi i de determinació d’autoria de textos escrits.” *Llengua, Societat i Comunicació*, 9; 69-85.