Int J Semiot Law (2024) 37:595–621
https://doi.org/10.1007/s11196-023-10056-5

Epistemic Modality Constructions as Stable Idiolectal Features: A Cross-genre Study of Spanish

Andrea Mojedano Batel[1] · **Amparo Soler Bonafont**[2] · **Krzysztof Kredens**[1]

Accepted: 20 September 2023 / Published online: 13 November 2023

Abstract

Forensic authorship analysis is based on two assumptions: that every individual has a unique idiolect, and that features characteristic of that idiolect will recur with a relatively stable frequency. Yet, a speaker's language can change with age, affective states, according to audience, or genre. Thus, studies on authorship analysis should adopt the theory that while some linguistic parameters of an idiolect can remain stable, others can change depending on various circumstances. This investigation, which takes a constructional and functional-based approach to discourse-level phenomena in idiolectal stability, analyzes cross-genre data produced by nine Mexican participants throughout a twelve-year time span. This study contributes to a greater understanding of the linguistic elements that survive genre effects and are potentially useful in both investigative and evidential forensic linguistic work. We provide a detailed description of linguistic features, their specific values, and context-dependent interpretation, keeping in mind the context of expert linguistic testimony, with its preference for methods which "employ linguistically motivated analyses in combination with quantitative tools" (Solan & Tiersma, 2004, p.463).

Our findings show that idiolectal style tends to remain stable across genres and communication modes in epistemic modality constructions. Epistemic markers — specifically, markers indicating low commitment by the speaker (e.g., *no sé* 'I don't know') or expressing indirectness when introducing the illocutionary force (e.g., *la verdad [es que]* 'the truth [is that]')— display idiolectal stability, as these markers seem to be the most effective in terms of allowing speakers to strategically manifest the extent of their knowledge regarding what is said.

Keywords Idiolect · Cross-genre analysis · Epistemic modality · Spanish · Authorship analysis · Usage-based linguistics

Extended author information available on the last page of the article

1 Introduction

Central to studies in authorship analysis is the assumption that individuals possess a distinct way of using language, that is, their own *idiolect*. The term, as first used in Bloch [7, p.7], originally refers to "the totality of possible utterances of one speaker at one time in using language to interact with one other speaker." The idea of authorship attribution, in turn, is based on two assumptions: that every language user (henceforth, "user") has a unique linguistic style and that features characteristic of that style will recur with a relatively stable frequency [11].

Most research carried out in authorship analysis seems to stem from the premise that the linguistic parameters of users' idiolects are stable. This notion of stability could have been influenced by Labov's concept of generational change [37], which posits that, after adolescence, the speech patterns of an individual remain mostly unchanged over time. Yet, Sankoff [55, p.1010] notes that "different levels of linguistic structure are differentially susceptible to modification in later life," with research showing that phonology is somewhat more susceptible to change than morphosyntax after the critical period [41] and well into adulthood. We know that a user's language can change with age [3, 35, 52, 53], under stress [51], in response to the audience [5, 13, 26], or with different genres [14, 61]. Moreover, third-wave sociolinguistic studies show that people get involved in a moment-to-moment negotiation of selves as a personal and individual dynamic which is at the same time tied inextricably to larger social orders, because language users are a product of their environment [16, 17]. Thus, the implications for authorship analysis are that while some linguistic parameters of a user's idiolect are possibly stable, others can change depending on various circumstances.

Researchers in authorship analysis have proposed hundreds of style markers and author-matching techniques over the years, with some recent studies reporting attribution rates for closed-set tasks in the region of 95% [e.g., 25, 31, 33, 32, 65]. These studies, which tend to use single-genre data, have contributed to cumulative knowledge in the field of authorship analysis. However, there is still a vital need for forensic authorship analysis research that captures stylistic similarities between texts created in different contexts and for different purposes and audiences, aiding in the identification of patterns of idiolectal stability across genres. Yet, authorship analysis studies using cross-genre and/or cross-domain data are few [e.g., 2, 23, 42, 43]. Ours is the first study to examine idiolectal stability with cross-genre data in Spanish. We show that Spanish epistemic modality constructions (i.e., epistemic expressions and discourse particles) are some of the linguistic features where idiolectal style remains stable across genres and communication modes. Previous studies regarding modality [30, 48, 49] show that a language user's commitment to what is said can be idiosyncratic. Research on modality constructions poses difficulties to analysts, because epistemicity emerges gradually and can indirectly reflect the illocutionary force of the speech acts in which it appears [19].

In this corpus-based study, we take a usage-based constructional approach to discourse-level phenomena in idiolectal stability. We analyze cross-genre data produced by nine Mexican participants throughout a twelve-year time span and provide a detailed description of epistemic modality constructions (henceforth, "EMCs"), their

specific values, and context-dependent interpretations. For this purpose, we have created nine independent subcorpora, each containing varied idiolectal data for each user, comprising different communication channels, genres, and contexts.

Our corpus-based study pursues a threefold goal. First, it offers an overview of EMCs in context, providing insights into idiolectal stability strategies in terms of individual users' constructional networks. Second, by using the Real-Time Construct, the study aims to gain a better understanding of the diachronic process of idiolectal stability to which EMCs are subject. Finally, it provides a linguistic interpretation of stable epistemic features, keeping in mind the context of expert linguistic testimony, with its preference for methods which "employ linguistically motivated analyses in combination with quantitative tools" [58, p.463].

2 Previous Studies of Idiolect

The notion of idiolect has been the subject of a substantial amount of linguistic literature. The idiolectal phenomena analyzed in linguistics so far can be categorized within the fields of, among others, forensic linguistics, corpus and computational linguistics, and variationist sociolinguistics. Generally, studies can be classified into more than one of these fields, as the disciplines tend to overlap. Although a full-scale overview of the linguistic research dedicated to the topic of idiolect is outside the scope of this paper, below we pay special attention to studies in corpus and computational linguistics (Sect. 2.1) and studies in forensic linguistics (Sect. 2.2).[1]

2.1 Cross-genre Idiolectal Studies in Corpus and Computational Linguistics

Within corpus and computational linguistics, there are only a handful of studies whose goal is to assess the level of stability of idiolectal features across genres, modes, topics and/or time of text production. A pioneering investigation was Goldstein-Stewart et al. [23], where the researchers built an English-language corpus consisting of communication samples from 21 participants in six genres on six topics. Their findings reveal that individuals can be identified with samples of their communication (a) across genres (accuracy of 71%); (b) in specific genres other than the one being tested (accuracy of 81%); and (c) in specific topics other than the one being tested (accuracy of 94%). However, identifying an author in one spoken genre after training with data from another spoken genre showed less than 48% accuracy in their study.

Litvinova, Litvinova, and Seredin [42] studied idiolectal variation in Russian written and spoken modes. Their findings point to low intra-individual variability and high inter-individual variability regarding the use of periods, conjunctions, and discourse particles across different types of texts. For spoken discourse, the same holds true for various markers at the level of elementary discourse units. There is another Russian-data study which tackles idiolectal cross-topic variation, namely, Litvinova, Seredin, et al. [43]. Their findings reveal that some linguistic parameters are relatively stable while others are largely variable. The most stable parameters in

[1] For a comprehensive overview of the concept of idiolect see Wright [66].

their study include the proportion of words that are more than six characters long, function words overall, prepositions, words describing cognitive processes, words in the "space" category, and punctuation marks.

Baayen et al. [2] investigate idiolectal cross-genre variation with Dutch data. A series of linear discriminant analyses show a "considerable authorial structure" based on a research design involving eight writers, each having produced three texts in the genres of fiction, argument, and description.

2.2 Idiolectal Studies in Forensic Linguistics

In authorship analysis studies, a common method used to identify authors and idiolects are n-gram analyses. Stamatatos [60], for example, examine the use of character n-grams in the identification of idiolects, and note that while most authorship attribution studies examine cases where topic and genre are controlled, such control is rarely the case in practical applications and can influence the effectiveness of the models when applied to real cases. The advantages of character n-gram features, according to Stamatatos, are simplicity of measurement, language independence, tolerance to noise, and the fact that character n-gram features are less likely to succumb to deception attempts due to their high-dimensional representation, which is based on information too difficult for humans to understand. However, and as Stamatatos himself notes, a limitation of using character n-gram attribution models in authorship research is the difficulty for the linguist to be able to explain the automatically derived decisions in court.

Kredens [34] carries out a computer-assisted study on idiolectal variation by comparing two English corpora. The two language users he studies share similar social and biological traits: Kredens assumes that if individual variation is found between such similar corpora, this individual variation will be even greater in dissimilar corpora. This is a notion we follow in the methodology for this study. To gauge idiolectal differences, Kredens compares a set of linguistic features with a differentiating potential in terms of frequencies of use. The three categories that show the highest potential to discriminate between the two idiolects ($p < .001$) are the most frequent words, the frequency of adverbs, and the use of discourse particles.

Wright [65] takes a corpus approach to test the accuracy of word n-grams in identifying authors of disputed email samples. Through the analysis of specific word strings that aided the identification of one author in particular, Wright shows how speech act realization is individual, and even if more than one author uses the same words, they do not do so in the same collocations. Wright's findings point to theories of entrenchment being helpful in authorship analysis tasks.

3 Epistemic Modality

Modality is the domain of expressions of possibility (epistemic modality) and necessity (deontic modality) [63]. It entails the activation of a gradual, non-discrete meaning that bears on its propositional value, and above all, on the relationship between a proposition and the participants in a specific communicative act. It is for this rea-

son that modality is sometimes analyzed from semantico-pragmatic, and even purely pragmatic, points of view [54].

Our study focuses on *epistemic modality*, which involves an estimation, generally made by the language user, of the probability that a state of affairs expressed in a clause applies in the world [50]. The frequent use of EMCs in Spanish [e.g., 15, 39, 59] can be ascribed to epistemic modality entailing strategies for the user to make themselves present in their utterance, thus regulating the commitment they desire to show in what they say. *Epistemic commitment*, the self-evaluation of the degree of truth with which a user expresses different propositions in their discourse, has to do with an idiosyncratic subjectivity in the way individuals position themselves against aspects of reality.

Our initial findings (see Sect. 6.1.1) show that users generally remain stable in terms of expressing epistemic modality. If it is possible to distinguish some EMCs in a user's linguistic output —with these constructions showing idiolectal regularity and distinctiveness—, then carrying out a description of said EMCs would allow us to establish specific idiolectal recognition patterns. All the EMCs that we analyzed in this study are used by participants in an intersubjective manner. Nuyts [48, 49] links the concept of *intersubjectivity* to the idea of a common ground between speaker and addressee, noting that speakers make use of certain linguistic devices when they assume that knowledge is shared between them and the addressees. It is in this common ground between interlocutors where modality constructions operate [15].

Regarding the operability of EMCs in idiolectal analysis, this research focuses on any epistemic forms capable of expressing a commitment to the user's reality or extra-linguistic truth. These forms can be verbs, adverbs, nouns, adjectives, or constructions showing conditional value [64].

4 Theoretical Framework: A Usage-Based Constructional Approach to Idiolectal Stability

This corpus-based study is situated within the field of forensic authorship analysis, addressing idiolectal stability in Spanish. Wright [65] argues that there are sequences of particular words functionally linked to recurring contexts and specific communicative purposes; while some of these links are part of the linguistic production of a specific community of practice, others are generated by only one of its members [46, 65]. Our initial findings point to EMCs being idiolectal sequences for Spanish.

Our theoretical framework is guided by usage-based construction grammar, with its concept of entrenchment [40]. Grammar is seen here as emergent from experience [28], with *entrenchment* referring to the level at which the formation and activation of a cognitive unit is automated through routinization, where the more a structure is used, the more entrenched it becomes, and vice versa. This framework can be useful in authorship analysis by illustrating the tendency for language users to utilize more than one variant within the same constructional network (albeit with different frequencies) instead of producing only one constructional variant within a network repeatedly. The notion of entrenchment lends itself well to forensic cases, where one must determine how consistent the linguistic output of a writer is, and what its

level of distinctiveness is [24]. We must therefore examine whether stable linguistic constructions —EMCs in this study— produced by one user are idiosyncratic enough to be distinguishable among various users who belong to the same community of practice.

Our analysis is based on the premise that the level of cross-genre consistency of EMCs (defined by the number of communicative genres in which an EMC appears in an idiolect) can indicate entrenchment of said constructions in the language user's mind. Specifically, the more genres where a specific EMC appears, the more entrenched this construction will be in a user's mind, since, as we mentioned before, users tend to modify their speech according to the situational context in which they find themselves and to their audience [5, 21, 52]. Our main aim is to find constructions that appear in most or all genres of a user's idiolect, and not on the amount of times each construction appears within one genre. In other words, our focus is not on frequency-based differences between idiolects but rather on identifying EMCs that survive genre effects for any given idiolect, in such a way that we could potentially consider them individualizing traits, which is a key issue in forensic linguistics.

In a usage-based approach to constructions, a string of words or morphemes used with a certain frequency can be considered a *construction* even when its form, function and/or meaning is entirely predictable [8].[2] This approach allows us to understand how constructions with unpredictable features, such as special pragmatic functions, arise. According to Enghels [18], relationships between patterns within a constructional paradigm are part of a language user's knowledge; therefore, the use of one constructional variant —in our specific case, the use of one EMC— must be understood in connection with the functioning of other variants related to it. Within these constructional networks of related items, new patterns can arise by becoming more frequent and invading the domains of another pattern, while older ones can disappear [18].

The network approach to constructions considers different types of associations among constructions. In the classical inheritance model [22], linguistic generalizations are represented in schematic higher-level constructions from which lower-level constructions inherit shared features. In the complete mode of inheritance, lower-level constructions are fully consistent with their higher-level counterparts; but in the default mode of inheritance, there can be some differences in value higher- and lower-level representations. We, like Goldberg [22], adhere to the default mode of inheritance. In this model, inheritance links assign semantically-driven inheritance relations between related constructions within the same constructional network. This helps us to understand how some constructions within a network can show both their basic meaning as well as other related meanings, which Goldberg [22] designated extensions of their basic meaning.

Finally, we examine the idiolectal use of EMCs as found in various points in time, spanning more than a decade (2008–2019). Studies have documented the use

[2] Most construction-based theories understand any linguistic patterning of words and morphemes as constructions if there is some aspect of form, function or meaning that is not predictable either from the component parts of the pattern or from other constructions in the language [e.g., 20, 22], contrary to what Bybee and Eddington [8] propose.

of EMCs in several dialects of Spanish at one point in time, but there is no research, to the best of our knowledge, examining their *idiolectal* use in what sociolinguists term "real time." Following Labov [37, p.73], *real-time* studies "observe a speech community at two discrete points in time." The default assumption when examining a community of practice is that, according to Sankoff [55], at any given time period, most aspects of language will be stable. Taking into consideration the context of forensic cases, more research is needed in diachronic idiolectal stability; this study is a first step toward analyzing if and how certain idiolectal linguistic features change over time.

The fact that this is a real-time panel study is important for two reasons. First, Sankoff [55, 56] already points out that longitudinal studies of real-time language change, where researchers return to the site of a previous sociolinguistic study to see what has happened in the interim, are both uncommon and relatively recent. The majority of longitudinal studies are *trend studies* (e.g. 6, 9, 62), examining a later sample of the community, where participants from the earlier sample are usually not included. By contrast, *panel studies* [e.g. 4, 12, 27, 57, 52, 53] examine data from the same participants at later points in time. Second, a real-time panel study is a fundamental undertaking within forensic linguistics. As expert witnesses, forensic linguists sometimes compare disputed documents from one time period with another. Investigating, then, if individuals show a proclivity to maintain the same epistemic modality constructions across time proves invaluable in terms of advancing research in idiolectal variation and change.

5 Participants and Data

The participants in our study are six Mexican women and three Mexican men who were between 30 and 60 years old at the time of the sociolinguistic interview we administered in 2020; all were born and raised in central Mexico, and currently work as lecturers and/or researchers at a public university in Mexico City, where they also live; their native language is Spanish. They have a similar social background and share a community of practice. All data have been collected and handled in accordance with Aston University's Policy on Research Ethics (Aston University Ethics Committee, 2020) and the study has Ethics Review Committee (ERC) approval. The data collection process involved obtaining the following linguistic data from each of the participants:

1. 30 text messages (specifically WhatsApp messages), amounting to a total of 6,122 tokens. All were written in 2020.
2. 30 emails, amounting to a total of 33,203 tokens. All were written between 2019 and 2020.
3. One semi-directed sociolinguistic interview with each participant, conducted by one of the researchers between 2019 and 2020 and amounting to a total of nine interviews, 18,842 tokens, and 174 min of audio. The main conversational topic was food, although the interviewer also let the participants talk freely about any

Table 1 Number of tokens for each user and communicative genre

	Text messages Token count	Emails Token count	Sociolinguistic interviews Token count	Work meetings Token count	Total
User 1 - female	326	1179	2919	65,858	**70,279**
User 2 - female	503	6549	2148	81,560	**90,760**
User 3 - male	270	3185	4259	34,297	**42,011**
User 4 - female	2114	5306	1562	19,980	**28,962**
User 5 - female	856	4916	956	1266	**7994**
User 6 - male	397	3114	1104	429,681	**434,296**
User 7 - male	332	1562	1387	146,647	**149,928**
User 8 - female	636	6388	3110	47,056	**57,190**
User 9 - female	688	1004	1397	17,759	**20,848**
Total	**6122**	**33,203**	**18,842**	**844,104**	**902,271**

Table 2 Communicative characteristics per genre

Genre	Production mode	Context	Audience
WhatsApp messages (30 per participant, 2020)	Written (phone keyboard)	Informal	Socially close
Emails (30 per participant, 2019–2020)	Written (full-sized keyboard)	Formal	Socially semi-distant
Sociolinguistic interviews (1 per participant, 2020)	Spoken (one-to-one interaction)	Informal	Socially semi-distant
Work meetings (monthly, 2008–2019)	Spoken (group interaction)	Formal	Socially distant

topic they chose; she also asked them questions in line with the principle of tangential shifting [36].

4. Transcripts of monthly work meetings in which the participants had been present and had spoken, amounting to a total of 844,104 tokens. These transcripts are in the public domain and fully accessible online. The meetings took place between 2008 and 2019, with people discussing university administrative matters.

Table 1 below displays the total number of tokens for each language user and genre in our corpus.

Table 2 below offers a summary of the data with their respective communicative characteristics and diaphasic variation related to each genre.

Data from dissimilar genres may lead to the presence of different patterns in the production of EMCs by the same user. However, there is a unity criterion that serves as the basis for the choice of these four genres: their permanent character of interlocutionary intersubjectivity [10, 15, 44]. All four genres are dialogic, in the sense that they either presuppose an answer to previous initiative interventions or are situated within a superior interaction frame. As Maldonado [44] notes, dialogic genres not only respond to the emissary's argumentative needs, but also operate consistently regarding the hearer/reader.

6 Methodology

In line with the methodological demands regarding language evidence in legal contexts, this corpus-based study follows a corpus linguistics methodology [45], offering two lines of exploration: frequency data and concordances, which respectively exemplify quantitative and qualitative analyses. Both types of analyses are equally important, since quantitative methods provide objectivity, and qualitative methods offer thorough linguistic explanations of the results. In the field of forensic linguistics, this mixed methodology has been supported by Solan and Tiersma [58] and applied by Johnson and Wright [29].

Quantitative methods (Sect. 6.1), that is, the analysis of frequency data, include the computational extraction of adequate tokens and determining the consistency and distinctiveness of EMCs through a basic measure of genre counts. Qualitative methods (Sect. 6.2), namely, looking at constructions in context, involve the coding of all clauses that contain EMCs and a thorough analysis of EMCs by examining their specific meanings and communicative functions in context.

6.1 Quantitative Analysis Method

6.1.1 Computational Extraction of Adequate Tokens

We analyzed token n-grams to investigate stable idiolectal features. The term *token* in our study refers to words and non-words in our corpora (e.g., word forms, punctuation signs, digits, abbreviations, etc.) and serves as the basis of our analysis. An *n-gram* is a contiguous sequence of *N* words or tokens. We conducted the analysis through lemmatized forms to obtain an ampler view of the possible idiolectal patterns; however, when it was deemed relevant, we analyzed declined and conjugated forms.

Through computational tools, we obtained a list of all n-grams that occurred in more than one genre per participant, capturing all token strings between one and four tokens in length. Subsequently, we interpreted each of the lists to determine which constructions to examine qualitatively and which constructions to discard due to their low discriminatory potential. The results show that five out of the nine study's par-

Table 3 Examples of raw (token) and genre count (GC) frequencies of constructions

User	Construction	Raw frequency	Genre count (GC)
User 1	*creo que*	62	2
User 2	*creo que*	191	4

ticipants remained stable in terms of expressing epistemic modality.[3] We examine the contextual use of stable expressions of epistemic modality, having obtained the constructions and their contexts of use through the AntConc concordancer [1].

Our main motivation for extracting token n-grams as linguistic features for analysis, and excluding others, such as character n-grams, is that token n-grams provide the explanatory power to support both linguistic theory and forensic linguistic research [29]. This type of n-gram analysis has been previously used in forensic authorship analysis studies, achieving high accuracy rates [e.g., 65].

Our data are not normalized, and there is some corpus imbalance in terms of the subcorpora sizes —something to be expected due to the nature of our data, but nonetheless a limitation in our study. However, our focus is not on frequency-based differences between idiolects but rather on identifying epistemic modality constructions that survive genre effects for any given idiolect, in such a way that we could potentially consider them individualizing traits.

6.1.2 Determining the Consistency and Distinctiveness of Epistemic Modality Constructions

Grant [24] argues that in forensic linguistic casework we need a methodology to determine how consistent and distinctive the linguistic output of a language user is. We must examine whether stable linguistic constructions produced by a language user are regular and distinctive enough to be distinguishable among different users.

To operationalize said methodology, we used a genre count (henceforth, "GC") as the basic measure to show the number of genres in which a participant produced a linguistic construction. For example, User 1 produced the bigram *creo que* ('I believe that') 62 times, in two genres; therefore, the total frequency of *creo que* for User 1 is 62, and the GC value for this bigram is 2. As a comparison, User 2 produced the bigram *creo que* 191 times, in four genres. The total frequency of *creo que* for User 2 is 191, and its GC value is 4. A genre count allows us to assess the stability of a user's constructions across genres. Table 3 offers these frequencies in a streamlined manner.

In the present study, we place our focus on the analysis of linguistic constructions relating to epistemic modality with a GC of 3 or 4, that is, on EMCs that appear in at least three genres of a user's linguistic output, i.e., EMCs that appear in written and spoken output and in formal and informal contexts. In other words, we analyze EMCs that have survived genre effects.

[3] We also identified other areas of idiolectal stability such as intensifying constructions (e.g., *qué* 'how' + adjective), quantifiers (e.g., *un poco* 'a little bit'), and deontic modality constructions (e.g., *tener que* 'to have to' + clause). We leave the analysis of these stable areas to future studies.

It is vital to carry out genre counts even though we are only focusing on GCs of 3 and 4. The rationale for this recount is that if a construction with a GC of 3 or 4 appears in the linguistic production of more than one user in our corpus, we must check if this construction is either (a) potentially distinctive of a user's idiolect, or (b) a common construction, widely adopted by various members of the community of practice under analysis. To account for this, we compare a user's GC with the mean genre count (MGC) calculated for the remaining eight users in the corpus. For example, the GC for *creo que* for User 2 is 4, while its MGC is 3.25. In other words, the bigram *creo que* was produced by User 2 in all four genres, while, for the remaining eight participants, *creo que* was produced, on average, in more than three genres, suggesting that this bigram is quite common in most of the participants' idiolects, and therefore, it is not potentially differentiating among idiolects.

Finally, Grant (24) argues that it is very difficult to distinguish individual linguistic productions from the linguistic productions of the social group to whom the individual belongs. As such, if we manage to find idiolectal differences between members of the same speech community, it follows that idiolectal differences from speakers outside the community will be even greater. Further, we do not present here a comparison against reference corpora. The problem with reference corpora is that these do not include cross-genre data produced by the same individual, making it impossible to draw conclusions about cross-genre idiolectal stability from them. In this investigation, we are not examining total frequencies: our focus lies in cross-genre frequencies, so that a reference corpus would not aid us in our research.

6.1.3 Type and Token Frequencies

As mentioned in Sect. 6.1, through computational tools, we obtained a list of all uni- to four-grams that occurred in more than one genre per user, and here we focus on the n-grams that occurred in three or four genres per user. When analyzed in context, these n-grams were usually part of larger constructions. For example, User 2 produced the n-gram *verdad* 'truth' in three genres, but in two different constructions: *Es verdad que* 'it is true that' and *la verdad es que* 'the truth is that.' Both constructions appeared in two (partly different) genres: *Es verdad que* appeared in emails and work meetings and *la verdad es que* appeared in the participant's sociolinguistic interview and work meetings. We can therefore say that the constructional network *verdad* contains two variants for User 2, namely, *es verdad que* and *la verdad es que*.

We analyzed all EMCs in the corpus that possessed a GC of 3 or 4 (for any individual participant) and an MGC of 2.5 or less; that is, all EMCs that showed potential in differentiating among idiolects. Overall, we analyzed a total of 315 tokens, which came from 26 different constructions (that is, type frequency is 26), as observed in Table 4 below.

For most unique constructions under examination (N = 24) we analyzed all examples produced by the users who had shown idiolectal stability for those items; the only two unique constructions for which we did not analyze all tokens were *o sea* 'I mean,' which appeared 111 times in User 4's subcorpus, and *por ejemplo* 'for example,' which appeared 44 times in User 8's subcorpus; we examined 30 tokens of each of these two constructions, due to time and space constraints.

Table 4 Type and token frequencies of epistemic modality constructions

Users who produced the same n-gram in 3 or 4 genres	Constructions containing said n-grams, with n-grams in bold (i.e., type frequency)	Token frequency
User 4	**¿no?** (Eng. *no?*)	18
User 4	**a la mejor** (Eng. *maybe*)	5
User 4	**ahora** (Eng. *now*)	22
User 2	algo **así** (Eng. *something like this*)	8
User 2	**así** como (Eng. *and/likewise*)	15
User 2	**así** pues (Eng. *therefore*)	2
User 2	**así** que (Eng. *hence/so*)	6
User 2	**cierto** (Eng. *certain*)	15
User 4	**como que** (Eng. *like*)	5
User 2	**de pronto** (Eng. *suddenly*)	27
User 8	de **verdad** (Eng. *really*)	1
User 2	es **verdad** que (Eng. *it is true that*)	3
User 8	esa es la **verdad** (Eng. *that is the truth*)	2
User 4	**estar viendo** (Eng. *to be seeing*)	13
User 8	la **verdad** (Eng. *the truth*)	3
User 2, User 8	la **verdad** es que (Eng. *the truth is that*)	9
User 4	**más bien** (Eng. *rather*)	8
User 8	**más o menos** (Eng. *more or less/ sort of*)	5
User 2	**nada más** (Eng. *nothing else/only*)	25
User 4	**no sé si** (Eng. *I don't know if*)	26
User 4	**o sea** (Eng. *I mean*)	30
User 8	**por ejemplo** (Eng. *for example*)	30
User 2	**sobre todo** (Eng. *above all/ specially*)	18
User 4	**todo lo que** (Eng. *everything that*)	8
User 8	**veo** por qué (Eng. *I see why*)	1
User 8	**veo** que (Eng. *I see that*)	10
Total	26 unique constructions	315 unique tokens

6.2 Qualitative Analysis Method: Parameters of Analysis

We analyzed the formal, semantic, and functional-discursive parameters shaping participants' statements. Below, we describe the variants and subvariants of said parameters.

(a) Formal parameters.

From a formal perspective, and as the basis for examining epistemic modality in the present study, we analyzed participants' whole interventions in oral genres and whole clauses in written genres. Interventions are recursive, as they can contain various clauses; in such cases, we isolated every clause, extracting all the features under analysis. Following a cognitive-functional framework, we understand the clause to be the smallest unit of analysis.

Within interventions and clauses, we examined two possible phenomena: the emergence of certain morphological categories with epistemic value and the co-occurrence of EMCs and other relevant syntactic, semantic or pragmatic segments in the immediate co-text.

(a.1) Morphosyntactic phenomena.

- Nominal constructions: focalizing nouns (*algo así*, *todo lo que*) or certainty adjectives (*cierto*).
- Verbal constructions: knowledge verbs (*(no) sé si*) or perception verbs (*veo que*, *estar viendo*).
- Full sentential constructions: qualifying certainty (*esa es la verdad*).
- Discourse connectors: consecutive and explanatory (*así pues*, *así que*, *o sea*) or adversative (*de pronto*, *ahora*), restrictive or delimiting in terms of truth attribution (*nada más*, *más bien*, *más o menos*, *sobre todo*, *por ejemplo*), expressing doubt or a low degree of commitment to what is said (*como que*, *así como*, *a lo/la mejor*, *¿no?*), and conveying certainty or a high degree of commitment to what is said (*verdad*, *la verdad es que*).

As can be observed in the above classification, while the units of analysis are syntactic, we subclassified them according to their semantic value. In this sense, the *Así* constructional network is the sole network in our study in which some constructions, apart from having a basic shared meaning, show a (related) extension of this meaning. More concretely, all constructions found in the *Así* network share a basic meaning of approximation, which possesses a mitigating value. Two of the network's constructions —*algo así* and *así como*— only show this original approximative value, whereas the remaining two constructions in the schema —*así pues* and *así que*— display an extension of the original meaning by introducing a possible, attenuating, conclusion to the communicative exchange.

We also examined if the units of analysis appeared within negative constructions (e.g., *no sé si* 'I don't know if') and their syntactic position within the clause or utterance (absolute initial position, relative initial position after a discourse particle or an absolute participle construction, intermediate position with positional mobility, or absolute final position).

(a.2) Co-occurrence phenomena.

Previous work [19, 59] has noted that EMCs tend to co-occur with other epistemic features in their immediate co-text. With this in mind, we examined the co-occurrences listed below.

- Co-occurrence with relevant syntactic constructions: conditional protasis, counter-arguments, direct questions, other epistemic constructions, overt personal pronouns, etc.
- Co-occurrence with relevant semantic constructions: probing constructions, suggestions, or conclusions.
- Co-occurrence with relevant pragmatic features: focus strategies, blurring strategies, or the introduction of explicit personal views.

(b) Semantic parameters.
(b.1) Basic semantic values.

- Meaning: EMCs were categorized as having one of the following basic informative values: focalizing, degree of certainty, knowledge, perception, consecution, explanation, or counter-position value.
- (Inter)subjective scale: We categorized each token depending on whether the user's commitment to certainty was subjective, that is, individual, or intersubjective, that is, presented as a shared truth. All examples were categorically classified as intersubjective.

(b.2) Argumentative values.
We distinguished between clauses in terms of their strong and weak counter-argumentative connectors. Strong- intensity clauses introduce a specific point of view or defend a proposition, whereas weak-intensity clauses introduce a doubt value.
(c) Functional-discursive parameters.

- Type of speech act in which the EMC is found: In our data, all EMCs were found in assertive speech acts, where the language user intended to increase the flow of (shared) knowledge. We distinguished between descriptive (i.e., objective) and evaluative (i.e., subjective) assertives, following Soler [59].
- Pragmatic intensity: Through the analysis of face negotiation, we differentiated between mitigating and intensifying discourse functions.

7 Results and Discussion

7.1 An Overview of Idiolectal Stability of Epistemic Modality Constructions Across the Community of Practice

Our initial computational analysis showed that four participants in the study (Users 1, 3, 5, and 9) did not produce the same EMC in at least three genres, a finding in line with authorship analysis studies, which indicate that the level of stability of idiolectal features is much lower across genres than in just one genre [e.g., 23].

Within the EMCs produced by participants who displayed cross-genre idiolectal stability in their EMC use (N=5, Users 2, 4, 6, 7, and 8), the EMCs *creer que* 'to think that' and *bueno* 'well' were discarded because they did not show potential as idiolectally distinguishing features, due to their high MGCs (3.25 and 2.6, respectively). Two male participants (Users 6 and 7) only presented cross-genre stability for *creer que* and *bueno*, so their idiolects were not further analyzed qualitatively. The frequent use of some EMCs in our corpus by all participants indicates that language users draw some of their epistemic constructions from the community of practice's shared feature pool.

Finally, three female participants in the study (Users 2, 4, and 8), who are all in their thirties or forties, evidenced stable EMC patterns that are potentially dif-

Table 5 Absolute frequencies of EMCs for User 2

EMCs	texts	email	interview	meeting	total
de pronto			6	21	**27**
nada más	1	1		23	**25**
sobre todo		1	1	16	**18**
cierto	1	1	1	12	**15**
así como		3	1	11	**15**
algo así		1	1	6	**8**
así que	1	4		1	**6**
la verdad es que			1	4	**5**
es verdad que		1		2	**3**
así pues		1	1	1	**3**
Total	**3**	**13**	**12**	**97**	**125**

Table 6 Absolute frequencies of EMCs for User 4

EMCs	texts	email	interview	meeting	total
o sea	1	3	21	10	35
no sé si	1	7	3	15	26
ahora	1	5		16	22
¿no?	1	1	15	1	18
estar viendo	1	5	1	6	13
todo lo que	2		2	4	8
más bien		4	1	3	8
a la mejor	2	1		2	5
como que	1	1	2	1	5
más o menos	2	1			3
Total	**11**	**28**	**45**	**47**	**143**

ferentiating in terms of idiolect. These findings are in line with the fact that young women tend to lead linguistic innovation and conform less to non-overtly prescribed norms: in linguistic changes from below, women show higher usage rates of innovative forms than men [38, 47].

We qualitatively analyzed all EMCs that either on their own or as part of a constructional network showed a GC of 3 or 4 and an MGC of 2.5 or less. These constructions come from Users 2, 4, and 8, and are discussed in Sect. 7.2, 7.3 and 7.4.

7.2 Stable Epistemic Modality Constructions that are Potentially Differentiating Idiolectally

7.2.1 User 2

This participant makes use of EMCs less frequently than User 4 and more frequently than User 8.

Table 5 below offers the reader the absolute frequencies of use of EMCs for User 2 for all genres.

In User 2's linguistic production, *cierto* displays an indefinite quantifying value. This use of *cierto*, along with the use of other epistemic constructions, such as the variants in the *así* 'like this' constructional network, allows us to discern semantic

values of qualification to what is said by User 2, who also presents the information as a personal qualification that can be easily shared by all, since it is an intersubjective truth.

The *verdad* constructional network (with variants *es verdad que* 'it is true that' and *la verdad es que* 'the truth is that') appears in clauses that entail strong arguments, where User 2, with great conviction and a high degree of commitment, presents her own point of view. See (1), obtained from one of her emails:

(1) ***La verdad es que desde mi punto de vista*** *no hay una buena traducción al español*

> **The truth is that from my point of view** there isn't a good translation into Spanish.

In (1), the EMC in absolute initial position anticipates a qualification of what follows. Nevertheless, User 2 resorts to another subjective expression ('from my point of view'), which explicitly reinforces her assertion, delimiting what is said. With this mitigating approach, User 2 tries to convince her interlocutor of her point of view without being imposing. She achieves this through an intersubjective technique, with the ultimate intention of intensifying her position, even if in a veiled way.

The *así* constructional network variants (i.e., *algo así*, *así como*, *así pues* and *así que*) introduce clauses in absolute initial or intermediate position that entail strong arguments, which User 2 upholds, treating them as conclusions to what is said. The constructional variants *algo así* and *así como* are produced by User 2 in an attempt to mitigate her strong assertions, in line with the constructional network's basic meaning. User 2 employs the network's remaining variants, namely, *así que* and *así pues*, to hedge her concluding remarks: here, the network's basic meaning still stands, but is somewhat blanched, so that the user's strong assertions are only minimally mitigated. In sum, this participant categorically displays with the *así* network a strong commitment to what is said, effectively changing objective [assertive] speech acts into subjective [assertive] speech acts, where a directly stated mitigating value prevails. Observe (2), obtained from one of her work meetings:

(2)

> ***Yo sé*** *que se oye muy feo decir que: […] pero lo más propio es que* ***justamente*** *es el acuerdo UACM/CU, entonces* ***yo propondría*** *que pudiéramos encontrar un sinónimo en el primer acuerdo para que diga* ***algo así*** *como "conviene"* ***o*** *"determina",* ***estoy de acuerdo.***
> **I know** that it sounds very bad to say that: […] but the most appropriate thing is that it is **precisely** the UACM/CU agreement, so **I would propose** that we could find a synonym in the first agreement so that it says **something like** "it is convenient" or "determines," **I agree**.

In (2), the *algo así* variant hedges User 2's proposal; she commits as little as possible with what is said, providing a suggestion together with an alternative. She is saving face while trying to reduce the impact on her interlocutor. Additionally, other epistemic constructions occur in the near co-text to *algo así* (e.g., the delimiting adverb

precisely, the verb *to know*, an agreement statement, and an overt first-person subject pronoun), all with a mitigating aim, which strategically strengthens her arguments by creating an intersubjective context and the pursuit for a shared understanding to what is said.

To summarize, User 2's idiolect shows stability in the use of two epistemic constructional networks, which, together with the frequent use of other EMCs, point to her epistemic idiolectal style as one where she either mitigates what is said to covertly intensify her position (e.g., with *verdad*) or minimizes the scope of what is said (e.g., with *así*).

7.2.2 User 4

This participant makes use of EMCs more frequently than Users 2 and 8. While User 4 does not produce any constructional networks, she displays a varied use of EMCs.

User 4's epistemic idiolectal style shows a focalizing semantic value achieved through mitigation, which in this case does not display an intensifying purpose (vs. User 2), but instead directly minimizes what is said. Furthermore, the co-occurrence of various EMCs within a clause produces an effect of low commitment to what is said. See (3), obtained from User 4's Whatsapp texts:

(3)

> ***No sé si*** *vaya un 01 antes*, ***a la mejor*** *por eso nunca me pude comunicar*, ***aunque*** *sí sonaba el teléfono y luego* ***de poquito*** *se oía* ***como que*** *no válido*.
> **I don't know if** a 01 should be placed before [the phone number], **maybe** that's why I could never get through, **although** one could hear the tone, and then **little by little** you could hear **something like** [the phone line] wasn't valid.

Here, User 4 produces 'I don't know if' in absolute clause-initial position to introduce a notion of doubt in her assertion. The participant wishes to commit only minimally to what is said, either because she does not know if she dialed the correct number or because she does not have enough proof to assert her claim more strongly. The co-occurrence of other EMCs in the co-text of 'I don't know if' provides a hedging value to the clause, helping weaken its argument through mitigation by offering an alternative ('maybe […]'), a weak counter-argumentative ('although'), a diminutive morpheme (*-ito*) within a diminutive adverb (*de poquito* 'little by little'), and the co-occurrence of another hedge ('something like') in the final segment of the user's intervention.

7.2.3 User 8

User 8 makes use of EMCs less frequently than Users 2 and 4, yet displays a tendency to make ample use of two EMCs (*veo que* and *la verdad es que*) which serve as the base [49] of two constructional networks.

Table 7 below offers the reader the absolute frequencies of use of EMCs for User 8 for all genres.

Table 7 Absolute frequencies of EMCs for User 8

EMCs	texts	email	interview	meeting	total
por ejemplo	1	1	8	10	**20**
veo que		3		7	**10**
la verdad es que	1	1	2	2	**6**
más o menos		1	1	3	**5**
esa es la verdad				2	**2**
de verdad		1			**1**
la verdad			1		**1**
veo por qué	1				**1**
Total	**3**	**7**	**12**	**24**	**46**

In line with a usage-based theory of language, both the overall higher token frequency of *la verdad es que* over other variants in its constructional network, as well as its use across all genres, points to this EMC being more entrenched than others in this network for User 8, indicating that epistemicity is a site of idiolectal stability, although at different levels according to each participant's style. All constructional variants in the *verdad* network generally provide a personal point of view with a high degree of commitment, by either pragmatically intensifying what is said or, through the intersubjective generalization of the certainty of what is said, strategically intensifying the utterance. See (4), obtained from User 8's sociolinguistic interview:

(4)

*Porque-**es que** ¿sabes qué?, **la verdad es que** en las vacaciones sí- estoy súper cansada, súper súper súper cansada y, este, y entonces, pues, **prácticamente** no hicimos nada.*
Because- **it's that**, you know what?, **the truth is that** when I'm on holiday yes- I'm super tired, super super super tired and, uh, and then, well, we **practically** didn't do anything.

The construction *la verdad es que* appears in non-absolute initial position in order to describe a personal state which can be face-threatening to the user herself. This is the reason why the construction appears alongside an excuse (*es que* 'it's that') and an adverb (*prácticamente* 'practically'), among other units, which together lead the evaluation of what is said toward a low commitment scale and to a direct mitigation of the assertion from a pragmatic perspective. The intersubjectivity that *verdad* introduces also has a mitigating aim, as what is said is only a product of a personal reflection on the user's face.

The perception verb *ver* 'to see,' conjugated in the present indicative, first-person singular form *veo* 'I see,' is the second most frequent EMC for User 8 when found in the construction *veo que* 'I see that.' This constructional variant, together with *(no) veo por qué* 'I don't see why,' establish a constructional network. Perception verbs have traditionally been understood as concrete representations by language users through which they present their worldview, and therefore, their commitment to the truth of their assertions [39, 48]. In (5), obtained from User 8's emails, we observe how this construction, which generally appears along other EMCs, displays a high commitment to what is said through intensification devices.

(5)

Igual ***ya*** *le pedí a Blanca si me puede entregar las que ella debía hacer y no hizo.* ***Pero veo*** *complicado que eso vaya a pasar*.
I also **already** asked Blanca if she can give me the ones that she had to do and didn't do. **But I see** [it is] unlikely that it will happen.

When producing *veo*, User 8 shows more self-assurance in what is said than in other utterances with different EMCs. *Veo* appears in a strong argument that must be interpreted as a conclusion or preferred option, introduced by the strong counter-argumentative *pero* 'but.' The absolute clause-initial position of *veo*, as well as the initial premise that repeats information, maximize the user's point of view, committing her to what is said, regardless of any possible damage to her or her interlocutor's face. This intensifying use shows that epistemicity is a gradual phenomenon, appearing in discourse usually through a joint use of various EMCs, which together co-direct the argumentative force.

7.3 An Overview of the Idiolectal Stability of Epistemic Modality Constructions Across time

Due to the nature of our data, the principal manner in which we obtained an overview of idiolectal stability traits over time was to compare linguistic output from work meetings (the genre which included the vast majority of longitudinal data) with present-day (i.e., 2020) productions in all genres.[4] In this section, we analyze data from the three participants (Users 2, 4, and 8) who have shown stable idiolectal use of EMCs throughout the last decade.

As discussed in Sect. 4, more productive constructional patterns can invade the domain of less productive patterns, or, inversely, some patterns can lose their vitality and leave a functional gap. Thus, constructional networks can either remain stable or change by either reorganizing themselves, increasing, or decreasing [18]. In the following subsections, we provide a closer look at epistemic modality constructions and constructional networks, how these change over time, and what this can tell us about idiolects.

7.3.1 User 2

The adverb *así* 'like this' creates a constructional network made up of the constructions *algo así* 'something like that,' *así como* 'and/likewise', *así pues* 'therefore,' and *así que* 'hence/so.' User 2 has been using *algo así* and *así como* since at least 2013 (the initial year we have data for this participant), and *así pues* and *así que* since at least 2014; this participant continues using these constructions today.

Table 8 below offers the reader the absolute frequencies of use of all EMCs found in the *Así* network for User 2, for all genres.

Out of these four constructions, the most frequent one, and possibly the most entrenched, is *así como* (N = 15), which introduces an expression of doubt and of low

[4] The only exceptions to these data are five of User 2's emails that were written in 2019, while all other emails for this participant as well as for Users 4 and 8 were written in 2020.

Table 8 Epistemic modality constructions in the Así network for User 2

EMC	Genre	Frequency	Year
Algo así	Email	1	2019
	Interview	1	2020
	Work meetings	6	2013 & 2014
Así como	Email	3	2019 & 2020
	Interview	1	2020
	Work meetings	11	2013–2015, 2018
Así pues	Email	1	2020
	Interview	1	2020
	Work meetings	1	2014
Así que	Email	4	2019 & 2020
	Work meetings	1	2014
	Chat	1	2020

Table 9 Epistemic modality constructions in the Verdad network for User 2

EMC	Genre	Frequency	Year
Es verdad que	Email	1	2020
	Work meetings	2	2017
La verdad es que	Interview	1	2020
	Work meetings	4	2013 & 2014

commitment to what is said. It is therefore unsurprising that two of the three remaining constructions show *así* at the beginning of the bigram, just like *así como* (i.e., *así pues* [N=3] and *así que* [N=6]). *Así pues*, which is syntactically separated from its host, shows the same hedging semantic value as *así como*, which acts as a connective conjunction. In sum, the extension of this constructional network is due to a mitigating pragmatic function.

Another n-gram that serves as the base of a constructional network is the lexeme *verdad* 'truth,' appearing in the variants *es verdad que* 'it is true that' (N=3) and *la verdad es que* 'the truth is that' (N=5). This constructional network remains diachronically stable across three genres, with *es verdad que* being solely used in formal contexts and *la verdad es que* exclusively produced in spoken contexts. User 2 has been producing the more frequent construction *la verdad es que* since at least 2013 (the initial year we have data for this participant), and *es verdad que* since at least 2017, and is still using both constructions today. The noun *verdad* is, together with the noun *realidad* 'reality,' one of the most productive forms in Spanish to introduce a user's presence into the text [19]. In the *verdad* constructional network, the semantic values of focalization and intersubjectivity on the one hand, and pragmatic intensity phenomena (i.e., mitigation and intensification) on the other, are employed with a similar frequency.

Table 9 below offers the reader the absolute frequencies of use of all EMCs found in the *Verdad* network for User 2, for all genres available.

The remaining EMCs that User 2 produces across genres (*cierto* 'certain,' *de pronto* 'suddenly,' *nada más* 'nothing else/only,' and *sobre todo* 'above all/specially') have also been employed in her speech since 2013 and are still used today, underscoring the idiolectal diachronic stability of epistemic constructions.

7.3.2 User 4

User 4 displays diachronic idiolectal stability throughout all EMCs that have been analyzed for her (see Table 6 for a list). Most EMCs appear in the data as far back as 2009 (N=4) or 2010 (N=4), highlighting the usefulness of EMCs in forensic linguistics. Yet, one construction seems to have emerged very recently (namely, 2019) in User 4's linguistic production: *como que* 'like,' a possibly newer quotative and discourse particle in the Spanish language which has not been examined in the literature so far, to the best of our knowledge. This finding also aligns with the fact that women tend to lead linguistic innovation [47] and conform less closely than men to sociolinguistic norms that are not overtly prescribed [38]. Lastly, this participant did not produce any epistemic constructional networks.

7.3.3 User 8

The construction *verdad* 'truth' establishes a constructional network for User 8, as it does for User 2, pointing to its productivity when it comes to generating EMCs, although as would be expected for different language users, most constructional variants in both networks are idiosyncratic of the user's linguistic productions; just one variant, *la verdad es que* 'the truth is that' overlaps in both networks. It does not seem coincidental that this constructional variant is the most frequent one in both networks, serving as its base construction (N=5 for User 2 and N=6 for User 8), and is the only variant that appears in more than one genre in the linguistic output of User 8 (GC=4).

The remaining variants for User 8's constructional network of *verdad* are *de verdad* 'really,' *la verdad* 'the truth,' and *esa es la verdad* 'that is the truth,' all showing a frequency of one. Of these three variants, only *esa es la verdad* appears in data from work meetings (2011–2012) but does not emerge in any other genre; the only variant that shows usage across time is the more entrenched *la verdad es que*, appearing in data from work meetings (2010–2011) and in data from 2020 from all other genres.

Table 10 below offers the reader the absolute frequencies of use of all EMCs found in the *Verdad* network for User 8, for all genres.

The n-gram *veo* 'I see' forms a second constructional network for User 8, showing two variants: *no veo por qué* 'I don't see why' (N=1) and *veo que* 'I see that' (N=10). While *no veo por qué* is constrained to the Whatsapp chat genre, *veo que* – the more frequent of the two – appears in three genres and has been produced by User 8 since at least 2011. It seems that this network has developed over time, where newer vari-

Table 10 Epistemic modality constructions in the Verdad network for User 8

EMC	Genre	Frequency	Year
De verdad	Email	1	2020
Esa es la verdad	Work meetings	2	2011 & 2012
La verdad	Interview	1	2020
La verdad es que	Email	1	2020
	Interview	2	2020
	Work meetings	2	2010 & 2011
	Chat	1	2020

ants (i.e., *no veo por qué*) could stem from older, more entrenched forms (i.e., *veo que*). Such constructional extensions are probably developed out of a need to broaden the types of personal viewpoints that User 8 introduces into her speech.

Additionally, both constructional networks (i.e., *verdad* and *veo*) generally introduce a personal point of view displaying a high degree of commitment. Thus, they either pragmatically intensify what is said or serve a strategically intensifying purpose, which stems from the generalization of the certainty of what is said, therefore becoming an intersubjective truth, since what is said is presented as an obvious fact.

Table 11 below offers the reader the absolute frequencies of use of all EMCs found in the *Veo* network for User 8, for all genres available.

All remaining EMCs that User 8 produces across genres (namely, *más o menos* 'more or less/sort of' and *por ejemplo* 'for example') have also been part of her linguistic production for years: the first instance in the data for *por ejemplo* is in 2010 and 2011 for *más o menos*, and she still uses both constructions today. Most of these remaining constructions show a focalizing semantic value. User 8 prefers this type of epistemic constructions over others such as comparatives (chosen by User 4) or qualifying EMCs (selected by User 2).

7.4 Forensic Clues for Idiolectal Recognition Based on Epistemic Resources

With these results, we see how the participants' use of EMCs can provide access to their cognitive styles. Through our analysis of language users who showed a cross-genre stable use of EMCs, we learned that each individual user maintains the same epistemic stance patterns across genres, achieving this through similar linguistic means in each genre; yet, the proportion of use of EMCs by user varies significantly. These findings are potentially useful in forensic cases, as we will explain below in more detail.

First, our analysis shows that in the close co-text of the EMCs we examined, other relevant constructions tend to appear. These constructions oftentimes serve as a guide to understand the specific semantic value of the EMC under analysis (see example 5). In other cases, they come together to induce and reinforce the same argumentative sense as the EMC under analysis in the utterance (see example 2).

Second, there are a series of individualized and recurrent epistemic patterns for each of the participants whose data were qualitatively examined. User 2's epistemic patterns are based on the semantic values of hedging, delimitation, and comparison, all focusing on a specific aspect of what is said. User 4's patterns are based on the production of various discourse particles, most of them oriented towards a similar semantico-pragmatic value, namely, mitigation. Finally, User 8's epistemic patterns are based on perceptual values expressed through overt first-person pronouns, exemplifying introductions, etc.

Table 11 Epistemic modality constructions in the Veo network for User 8

EMC	Genre	Frequency	Year
No veo por qué	Chat	1	2020
Veo que	Email	3	2020
	Work meetings	6	2011 & 2012
	Chat	1	2020

Third, the pragmatics of the EMCs produced by Users 2, 4, and 8 are decisive in terms of the semantic values that these users choose, consciously or not. The mitigating pragmatic category is ever present, to protect (mainly in clause-initial position) or repair (chiefly in clause-medial or final position) the user's face. Furthermore, EMCs produced by these users appear in indirect speech acts, instead of the more usual direct speech acts, where mitigation is strategically employed, responding to a veiled intensifying purpose.

Fourth, the diachronic stability of epistemic modality constructions and of the constructional networks they instantiate underscores these features' potential in aiding forensic linguists when acting as expert witnesses in legal contexts.

These four phenomena, namely, (1) the use of various epistemic modality constructions, co-directed either towards the same semantic value or towards the same argumentative force; (2) the use of individualized and recurrent epistemic patterns; (3) the employment of direct or indirect mitigation devices to present what is said; and (4) the diachronic stability in EMCs, could help the forensic linguist identify idiolectal patterns. Additionally, the layout of mitigating EMCs (i.e., clause initial, medial, or final), their purpose (i.e., to protect or repair the user's or their interlocutor's face), and their configurational strategy all serve to differentiate the communicative style of a language user.

8 Conclusions

Our initial list of all token n-grams that occurred across various genres identified four areas of linguistic stability, with epistemic modality constructions being the most promising area in terms of authorship analysis tasks. This finding is in line with results from studies in English idiolectal variation [34] and cross-genre idiolectal studies for Russian, where there is low intra-individual variability and high inter-individual variability in the use of discourse particles [42] and of words that describe cognitive processes [43]. These similar results for Spanish, English, and Russian point to epistemic modality constructions as promising sites of cross-linguistic idiolectal stability which could be used in investigations to help deanonymize offenders posting multilingual content in online criminal environments.

Regarding overall cross-genre idiolectal stability in the community of practice we studied, the results were quite promising, as five out of the nine language users we studied produced the same EMC in at least three genres; more importantly, three of these five users displayed a consistent and distinct production of epistemic modality constructions, potentially idiosyncratic enough to be distinguishable among a large pool of authors. The known difficulty in authorship analysis studies to distinguish individual linguistic patterns from communal linguistic patterns in a community of practice [24] could be thus partially tackled by analyzing epistemic features.

Interestingly, the three users who show a regular and idiosyncratic use of epistemic modality constructions are all women of a similar age (between 30 and 40 years old). From a sociolinguistic perspective, these findings relating to gender are in line with what Labov calls the gender paradox: "women conform more closely than

men to sociolinguistic norms that are overtly prescribed, but conform less than men when they are not" [38, p.293].

In terms of idiolectal stability of EMCs through time, we find that the three participants for whom we analyzed data qualitatively have maintained their epistemic patterns for at least seven years. These results are in line with the well-known fact that people retain their original patterns throughout their lives, where most aspects of the language of a speech community will not be involved in ongoing change [55].

As regards epistemic constructional networks, language users tend to produce more than one variant within a paradigm, where the use of different, yet related, items answers to different pragmatic needs. Due to the diachronic nature of our data, we could observe in real time how a network develops over time, with newer variants originating from older, more entrenched forms, in line with Enghels's [18] theories of variation and change in constructional networks. This is again promising as the current successful forensic authorship analysis methods are mostly based on character or word n-grams, which are likely to capture such derivational change.

Finally, this study adds to the body of literature on cross-genre authorship analysis studies and on idiolectal variation in Spanish by being the first to study cross-genre idiolectal variation in Spanish. It shows that there are idiolectally stable linguistic features across genres in Spanish, even within a relatively homogenous group of language users. As expert witnesses, forensic linguists often compare disputed documents in one genre with known-authorship documents in another; they might, for example, be asked to compare a set of WhatsApp messages with a handwritten diary. This study thus contributes to a greater understanding of the linguistic elements that survive genre effects and are therefore potentially useful in both investigative and evidential forensic linguistic work.

References

1. Anthony, Laurence. 2017. *AntConc (Version 3.5.0) [Computer Software]*. Tokyo, Japan: Waseda University.
2. Baayen, Harald, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. 2002. An experiment in authorship attribution. In *6th JADT: 6es Journées internationales d'Analyse statistique des Donnés Textuelles* 1: 69–75.
3. Bailey, Guy, Tom Wilke, Jan Tillery, and Lori Sand. 1991. The apparent time construct. *Language Variation and Change* 3(3): 241–264. https://doi.org/10.1017/S0954394500000569.
4. Baugh, John. 1996. Dimensions of a theory of econolinguistics. In *A Social Science of Language: Papers in Honor of William Labov*, ed. Gregory R. Guy, Crawford Feagin, Deborah Schiffrin and John Baugh, 397–419. Philadelphia, Pennsylvania: John Benjamins.

5. Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2): 145–204. https://doi.org/10.1017/S004740450001037X.
6. Blake, and Renee and Meredity Josey. 2003. The /ay/ diphthong in a Martha's Vineyard community: what can we say 40 years after Labov? *Language in Society* 32: 451–485.
7. Bloch, Bernard. 1948. A set of postulates for phonemic analysis. *Language* 24(1): 3–46. https://doi.org/10.2307/410284.
8. Bybee, Joan, and David Eddington. 2006. A usage-based approach to spanish verbs of becoming. *Language* 82(2): 323–355. http://www.jstor.org/stable/4490159.
9. Cedergren, Henrietta. 1988. The spread of language change: Verifying inferences of linguistic diffusion. In *Language Spread and Public Policy: Issues, Implications, and Case Studies (Georgetown University Round Table on Languages and Linguistics 1987)*, ed. Peter H. Lowenberg, 45–60. Washington, D.C.: Georgetown University Press.
10. Cornillie, Bert. 2010. On conceptual semantics and discourse functions: the case of spanish modal adverbs in informal conversation. *Review of Cognitive Linguistics* 8(2): 300–320. https://doi.org/10.1075/rcl.8.2.03cor.
11. Coulthard, Malcolm, Tim Grant, and Krzysztof Kredens. 2011. Forensic linguistics. In *The SAGE handbook of Sociolinguistics*, eds. Ruth Wodak, Paul Kerswill, and Barbara Johnstone. 531–544. London: Sage.
12. Cukor-Avila, Patricia. 2002. She say, she go, she be like: verbs of quotation over time in african american Vernacular English. *American Speech* 77: 3–31.
13. Cutillas-Espinosa, Juan, Antonio, and Juan Manuel Hernández-Campoy. 2007. Script design in the media: radio talk norms behind a professional voice. *Language & Communication* 27(2): 127–152. https://doi.org/10.1016/j.langcom.2006.04.001.
14. D'Arcy, Alexandra, Bill Haddican, Hazel Richards, Sali Tagliamonte, and Ann Taylor. 2013. Asymmetrical trajectories: the past and present of–body/–one. *Language Variation and Change* 25(3): 287–310. https://doi.org/10.1017/S0954394513000148.
15. De Cock, Barbara. 2014. *Profiling discourse participants. Forms and functions in spanish conversation and debates*. Amsterdam and Philadelphia: John Benjamins.
16. Eckert, Penelope. 2012. Three waves of variation study: the emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41: 87–100. https://doi.org/10.1146/annurev-anthro-092611-145828.
17. Eckert, and Penelope and Sally McConnell-Ginet. 1999. New generalizations and explanations in language and gender research. *Language in Society* 28(2): 185–201. https://doi.org/10.1017/S0047404599002031.
18. Enghels, Renata. 2018. Towards a constructional approach to discourse-level phenomena: the case of the spanish interpersonal epistemic stance construction. *Folia Linguistica* 52(1): 107–138. https://doi.org/10.1515/flin-2018-0002.
19. Fernández, Sara, and M. Amparo Soler. 2020. La combinación de marcadores discursivos epistémicos y contraargumentativos: una estrategia atenuadora fundamentada en el contraste. In *Aportaciones desde el español y el portugués a los marcadores discursivos. Treinta años después de Martín Zorraquino y Portolés, ed. Antonio Messias Nogueira da Silva, Catalina Fuentes Rodríguez, and Manuel Martí Sánchez*, 209–226. Sevilla: Editorial Universidad de Sevilla.
20. Fillmore, Charles, Paul Kay, O. Mary Catherine, and 'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language* 64(3): 501–538. https://doi.org/10.2307/414531.
21. Giles, Howard Peter Powesland. 1975. *Speech style and social evaluation*. London and New York: Academic Press. https://doi.org/10.1017/S0047404500005820.
22. Goldberg, Adele. 1995. *Constructions: a construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.
23. Goldstein-Stewart, Jade, Ransom Winder, and Roberta Sabin. 2009. Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*: 336–344.
24. Grant, Tim. 2010. Text messaging forensics Txt 4n6: Idiolect free authorship analysis? In *The Routledge handbook of forensic linguistics*, eds. Malcolm Coulthard, and Alison Johnson. 536–550. Abingdon: Routledge. https://doi.org/10.4324/9780203855607.cU33.
25. Grieve, Jack. 2007. Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing* 22(3): 251–270. https://doi.org/10.1093/llc/fqm020.

26. Hay, Jennifer, Stefanie Jannedy, and Norma Mendoza-Denton. 1999. Oprah and /ay/: Lexical frequency, referee design and style. In *Proceedings of the 14th international congress of phonetic sciences*, 1389–1392. Berkeley, CA: University of California.
27. Hernández-Campoy, Juan Manuel. 2003. Complementary approaches to the diffusion of standard features in a local community. In *Social Dialectology: in Honour of Peter Trudgill*, eds. David Britain, and Jenny Cheshire. 23–37. Amsterdam, The Netherlands: John Benjamins.
28. Hopper, Paul. 1987. Emergent Grammar. *Berkeley Linguistics Society* 13: 139–157.
29. Johnson, and Alison and David Wright. 2014. Identifying idiolect in forensic authorship attribution: an n-gram textbite approach. *Language and Law/Linguagem e Direito* 1(1): 37–69.
30. Kärkkäinen, Elise. 2003. *Epistemic stance in English conversation. A description of its interactional functions, with a focus on I think*. Amsterdam and Philadelphia: John Benjamins.
31. Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* 60(1): 9–26.
32. Koppel, Moshe, Jonathan Schler, and Schlomo Argamon. 2013. Authorship attribution: what's easy and what's hard? *Journal of Law and Policy* 21(2): 317–331. https://doi.org/10.2139/ssrn.2274891.
33. Koppel, Moshe, and Jonathan Schler. 2007. and Elisheva Bonchek-Dokow. Measuring differentiability: unmasking pseudonymous authors. *Journal of Machine Learning Research* 8(6).
34. Kredens, Krzysztof. 2002. Towards a corpus-based methodology of forensic authorship attribution: a comparative study of two idiolects. In *PALC'01: practical applications in Language Corpora*, ed. Barbara Lewandowska-Tomaszczyk. 405–437. Peter Lang: Frankfurt am Mein.
35. Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania.
36. Labov, William. 1984. Field methods on the project on linguistic change and variation. In *Language in Use*, eds. J. Baugh, and J. Sherzer. Englewood Cliffs: Prentice Hall.
37. Labov, William. 1994. *Principles of linguistic change: internal factors*. Blackwell.
38. Labov, William. 2001. *Principles of linguistic change. Volume 2: social factors*. Blackwell.
39. Landone, Elena. 2012. El alcance interpersonal de los marcadores del discurso en la dinámica conversacional. El ejemplo de la cortesía verbal. *Verba* 39: 301–313.
40. Langacker, Ronald. 1987. *Foundations of cognitive grammar, vol. I: theoretical prerequisites*. Stanford, CA: Stanford University Press.
41. Lenneberg, Eric. 1967. *The Biological foundations of language*. New York, NY: Wiley.
42. Litvinova, Tatiana, Olga Litvinova, and Pavel Seredin. 2018. Assessing the level of stability of idiolectal features across modes, topics and time of text production. In *2018 23rd Conference of Open Innovations Association (FRUCT)*, 223–230. IEEE.
43. Litvinova, Tatiana, Pavel Seredin, Olga Litvinova, Tatiana Dankova, and Olga Zagorovskaya. 2018. On the stability of some idiolectal features. In *International Conference on Speech and Computer*, 331–336. Springer.
44. Maldonado, Ricardo. 2018. Certezas atenuadas. *Rilce Revista de Filología Hispánica* 34(3): 1129–1153. https://doi.org/10.15581/008.34.3.1129-53.
45. McEnery, Tony, Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
46. Mufwene, Salikoko. 2010. SLA and the emergence of creoles. *Studies in Second Language Acquisition* 32(3): 359–400. https://doi.org/10.1017/S027226311000001X.
47. Nevalainen, and Terttu and Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language Change in Tudor and Stuart England*. Pearson Education Ltd.
48. Nuyts, Jan. 2001a. Subjectivity as an evidential dimension in epistemic modal expressions. *Journal of Pragmatics [special Issue]* 33(3): 383–400. https://doi.org/10.1016/S0378-2166(00)00009-6.
49. Nuyts, Jan. 2001b. *Epistemic modality, language, and conceptualization: a cognitive-pragmatic perspective*. John Benjamins.
50. Nuyts, Jan. 2016. Analyses of the Modal Meanings. In *The Oxford Handbook of Modality and Mood*, ed. Jan Nuyts & Johan van der Auwera, 31–49. Oxford: Oxford University Press.
51. Pennebaker, James. 2002. and Thomas Lay. Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality* 36(3): 271–282. https://doi.org/10.1006/jrpe.2002.2349.
52. Rickford, and John and Faye McNair-Knox. 1994. Addressee-and topic-influenced style shift: a quantitative sociolinguistic study. In *Sociolinguistic perspectives on register*, eds. Douglas Biber, and Edward Finegan. 235–276. Oxford: Oxford University Press.
53. Rickford, John Mackenzie Price. 2013. Girlz II women: Age-grading, language change and stylistic variation. *Journal of Sociolinguistics* 17(2): 143–179. https://doi.org/10.1111/josl.12017.

54. Rodríguez Bravo, Ana Eva. 2017. *Modalidad y verbos modales*. Madrid: Arco Libros.
55. Sankoff, Gillian. 2005a. Cross-sectional and longitudinal studies. In *Volume 2: an International Handbook of the Science of Language and Society*, eds. Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier, and Peter Trudgill. 1003–1013. Berlin and New York: De Gruyter Mouton. https://doi.org/10.1515/9783110171488.2.7.1003.
56. Sankoff, Gillian. 2005b. Age: Apparent time and real time. In *Elsevier Encyclopedia of Language and Linguistics (2nd edition, vol. 9)*, ed. Keith Brown. Oxford, U.K.: Elsevier.
57. Sankoff, and Gillian and Helene Blondeau. 2007. Longitudinal change across the lifespan: /r/ in Montreal French. *Language* 83: 560–588.
58. Solan, and Lawrence and Peter Tiersma. 2004. Author identification in American Courts. *Applied Linguistics* 25(4): 448–465. https://doi.org/10.1093/applin/25.4.448.
59. Soler, M., and Amparo. 2017. La verdad (es que): significado nuclear y atenuante. *Signos Estudios de Lingüística* 50(95): 430–452. https://doi.org/10.4067/S0718-09342017000300430.
60. Stamatatos, Efstathios. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy* 21(2): 421–439.
61. Taylor, Ann. 1994. Variation in past tense formation in the history of English. *University of Pennsylvania working papers in linguistics* 1(1): 143–158.
62. Trudgill, Peter. 1988. Norwich revisited: recent linguistic changes in an english urban dialect. *English World Wide* 9: 33–49.
63. Van der Auwera, Johan. 1996. Modality: the three-layered scalar square. *Journal of Semantics* 13(3): 181–195. https://doi.org/10.1093/jos/13.3.181.
64. Von Fintel, Kai. 2006. Modality and Language. In *Encyclopedia of philosophy*, ed. Donald M. Borchert. 2nd ed. Detroit: MacMillan. http://mit.edu/fintel/www/modality.pdf.
65. Wright, David. 2017. Using word n-grams to identify authors and idiolects: a corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics* 22(2): 212–241. https://doi.org/10.1075/ijcl.22.2.03.
66. Wright, David. 2018. Idiolect. *Oxford Bibliographies in Linguistics*. https://doi.org/10.1093/OBO/9780199772810-0225.

Authors and Affiliations

Andrea Mojedano Batel[1] · Amparo Soler Bonafont[2] · Krzysztof Kredens[1]

✉ Andrea Mojedano Batel
a.mojedanobatel@aston.ac.uk

[1] Aston University (Aston Institute for Forensic Linguistics, English department), Birmingham, UK

[2] Universidad Complutense de Madrid (Spanish and literary theory department), Madrid, Spain