

MUESCA: A Multi-Genre Corpus and Annotation Scheme for Subjectivity in Spanish

Alba Pérez-Montero

alba.perezm@ua.es

University of Alicante

Elena Lloret

University of Alicante

Paloma Moreda

University of Alicante

Research Article

Keywords: Subjectivity Detection, Spanish Corpus, Multi-genre, Corpus Linguistics, Annotation methodology, Subjectivity Annotation, Pragmatics

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10521286/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

MUESCA: A Multi-Genre Corpus and Annotation Scheme for Subjectivity in Spanish

Alba Pérez-Montero^{1*}, Elena Lloret Pastor² and
Paloma Moreda Pozo²

^{1*}University Institute for Computing Research (IUII), University of Alicante.

²Dept. of Software and Computing Systems, University of Alicante.

*Corresponding author(s). E-mail(s): alba.perezm@ua.es;
Contributing authors: elloret@dlsi.ua.es; moreda@dlsi.ua.es;

Abstract

This paper presents MUESCA, a multi-genre corpus and annotation scheme for Subjectivity Detection (SD) in Spanish. The corpus operationalizes subjectivity along a continuum across three textual genres, utilizing scientific abstracts and consumer reviews as polar anchors and news articles as an hybrid category. To characterize these texts, the approach employs pragmatic features categorized into high author commitment (encompassing certainty, participation, and specificity) and low author commitment (encompassing doubt, non-specificity, and distancing). A Kruskal-Wallis analysis is conducted to evaluate the distribution and discriminative capacity of these features across the subcorpora. Additionally, the resource is evaluated through subjectivity classification tasks. Experiments using a Multi-Layer Perceptron (MLP) model integrating these pragmatic features result in a peak F1-score of 83.81% in multi-level configurations. Furthermore, combining these features with a BETO transformer architecture yields a global F1-score of 93.8%, particularly affecting the classification of the news genre. Overall, this work provides a methodology to quantify pragmatic characteristics across distinct registers, offering datasets and baselines intended to support further research into modeling complex or hybrid discourses in Spanish.

Keywords: Subjectivity Detection, Spanish Corpus, Multi-genre, Corpus Linguistics, Annotation methodology, Subjectivity Annotation, Pragmatics

1 Introduction

Subjectivity Detection (SD) is a fundamental task in Natural Language Processing (NLP) focused on distinguishing factual information from expressions of personal opinion. By identifying subjective markers, such as feelings, evaluations, and individual perspectives, within a text, SD provides the necessary layer of interpretation for various NLP architectures, such as opinion mining (Kasmuri & Basiron, 2017).

From the NLP perspective, SD has traditionally been approached as a binary classification problem. The traditional distinction between fact and opinion is often rooted in a binary conceptualization of subjectivity as the mere absence of objectivity. However, as noted by Tuchman and Aladro Vico (1998), absolute objectivity is frequently a discursive ideal rather than a linguistic reality, as few texts are entirely separable from the speaker’s perspective. Building upon this concept, this study moves beyond the binary classification (fact vs. opinion) by proposing that subjectivity and objectivity are not discrete, mutually exclusive classes. Instead, we operationalize them as poles of a fluid *continuum*, where the degree of authorial presence determines the position of a text within a discursive gradient. Within this virtual space, linguistic features are often combined and mixed, creating different degrees of intensity rather than a binary division. This landscape makes SD highly dependent on pragmatic context, particularly in morphologically rich languages like Spanish, where subjectivity is encoded through complex inflectional and syntactic structures that go beyond simple lexical choices. For example, knowledge-oriented verbs (e.g., ‘know’ or ‘understand’) encode the speaker’s alignment with the truth of a statement, transforming a simple predicate into a marker of subjective certainty.

An example where this *continuum* is clearly manifest is journalism, particularly within dis- and misinformation phenomena. Research in disinformation has shown that unreliable news strategically navigates this hybrid space, employing a mixture of verifiable data and evaluative framing to enhance its perceived authority (Bonet-Jover, 2023). In these contexts, the primary challenge lies in identifying the “authorial footprint” hidden within discourse that appears superficially neutral. Consequently, modeling subjectivity as a *continuum* is therefore more than a linguistic exercise; it is essential for tasks such as disinformation and misinformation detection, among others.

Despite the critical nature of this task, the advancement of SD is significantly limited by a dual scarcity of resources: a predominant focus on the English language and a persistent restriction to single-genre datasets, that is, collection of texts pertaining to a same textual genre such as news, tweets, etc. To address this gap, our work introduces a novel, multi-genre Spanish corpus that moves beyond a single-domain approach to reflect the fluidity of subjective expression in real-world discourse. Furthermore, this approach emphasizes explainability by grounding the data in hand-crafted linguistic features rather than relying solely on automated patterns, providing a more robust methodology for Hispanic linguistic analysis.

The main contribution of this work is the development of the **MUESCA Corpus**, a novel, multi-genre Spanish corpus for Subjectivity Detection encompassing *scientific abstracts*, *news articles*, and *consumer reviews*. To ensure its quality, robustness, and computational utility, this main contribution is supported by three key methodological advancements:

- **Theoretical and Computational Grounding:** We provide a comprehensive mapping of pragmatic dimensions operationalized through multi-level linguistic markers (lexical, syntactic, and semantic), allowing for a formal characterization of the subjectivity-objectivity gradient that defines the corpus.
- **Pragmatic Annotation Guidelines:** A comprehensive set of guidelines specifically designed to capture the nuances of subjective language in Spanish through two levels of author implication: High Author Commitment (displayed by the dimensions of Certainty, Specificity, and Participation) and Low Author Commitment (entailed by the dimensions of Doubt, Non-specificity, and Distancing).
- **Experimental Validation:** A hybrid expert-in-the-loop annotation methodology and a comparative computational analysis demonstrating that the integration of these pragmatic features significantly enhances the performance and explainability of subjectivity classification models.

The paper is structured as follows: In Section 2 we will revise previous works presenting resources in order to draw gaps that our contributions may cover. In section 3 we detail how the MUESCA corpus is collected (Section 3.2), which annotation guidelines are followed (Section 3.3), and we introduce a newly design hybrid annotation methodology (Section 3.4). After that, we design an evaluation methodology to assess the discriminatory potential of our annotated tags (Section 4.1) and evaluate their efficacy in a three-level subjectivity classification task (comprising objective, hybrid, and subjective levels). (Section 4.2). Finally, we draw some conclusions and open future work lines in Section 5.

2 Related Work

Methodological advancements in Subjectivity Detection (SD) are fundamentally intertwined with the availability and quality of annotated data. Over the past decades, the field has transitioned from small-scale, expert-curated lexicons to massive, multi-speaker corpora that better reflect the pragmatic complexities of modern discourse. This section reviews the current state of these resources and examines how their development has enabled critical applications across diverse NLP domains. A detailed comparison of the reviewed literature, encompassing both the corpora and the specific applications for which they were developed, is presented in Table 2.

2.1 Available Resources for Subjectivity Detection

The development of SD resources has evolved from static, word-level lexicons to complex, context-aware corpora. Early efforts relied heavily on lexical engineering, such as the *Subjectivity Lexicon* (Wilson, Wiebe, & Hoffmann, 2005), which provided a foundational list of evaluative terms. These resources were often applied to domain-specific datasets, such as the BLLIP WSJ corpus (Charniak et al., 2000), which contrasted 74,000 objective news reports with 2,000 subjective letters to the editor (Krüger, Lukowiak, Sonntag, Warzecha, & Stede, 2017). While effective, these early corpora were primarily English-centric and restricted to binary distinctions between news and opinion.

A cornerstone in the field is the MPQA Opinion Corpus (Wilson, Wiebe, & Cardie, 2017), a collection of news articles containing detailed, expression-level annotations of sentiment and subjectivity. Rather than using simple binary tagging, the corpus employs a sophisticated, multi-attribute frame-based annotation scheme that captures nuances such as attitude type, intensity on a four-point scale, and polarity. By providing this granular depth, MPQA shifted the focus from document-level sentiment to the linguistic triggers of subjectivity.

Cross-linguistic resource development has seen significant growth in recent years. In the Italian context, the *SubjectivITA* corpus (Antici et al., 2021) provided 1,841 manually annotated sentences from online newspapers, shifting the focus to sentence-level granularity. More recently, the CLEF CheckThat! Lab (Struß et al., 2024) introduced a massive multilingual resource for SD in news, covering Arabic, Bulgarian, English, German, and Italian. Despite the scale of these large-scale benchmarks, they remain largely focused on the news genre and utilize binary classification schemes.

Concurrently, the integration of automated methodologies has further expanded annotation paradigms. For instance, Shokri, Sharma, Filatova, Jain, and Levitan (2024) utilized Chain-of-Thought (CoT) prompting to evaluate subjectivity in English news, while Song, Hu, and Mahowald (2025) leveraged generative models to analyze metalinguistic dimensions of subjective language across Wikipedia and news articles.

Efforts to include Spanish have often relied on multilingual adaptations of existing tools or localized lexicons. For instance, Volkova, Wilson, and Yarowsky (2013) explored subjectivity in Spanish tweets and news summaries using the *Linguistic Inquiry and Word Count* (LIWC) lexicons (Pennebaker, Francis, & Booth, 2001). In the Luso-Hispanic sphere, researchers have developed Portuguese subjectivity lexicons to analyze fabricated content (Vieira, Jeronimo, Campelo, & Marinho, 2020).

Despite these advancements, a critical research gap remains. Most available resources for Spanish are either domain-specific (restricted to social media or news) or represent mere translations of English methodologies. There is a notable absence of multi-genre Spanish corpora that capture how subjectivity manifests across different communicative registers, from formal scientific discourse to informal consumer reviews. Furthermore, existing datasets often rely on binary labels that fail to provide the pragmatic depth required to model the “linguistic fingerprint” of the subjectivity-objectivity continuum. To address this, our study introduces an annotation scheme designed to capture this nuance by categorizing content based on a spectrum of high author commitment versus low author commitment, shifting the focus from simple sentiment polarity to the structural manifestation of the author’s presence in the text.

2.2 Applications of Subjectivity Detection Resources

The utility of SD resources extends beyond simple text classification, serving as a critical component in diverse NLP tasks where the distinction between fact and opinion is paramount.

The most prominent application lies in the field of Disinformation and Fake News Detection. Research has shown that unreliable information often relies on a strategic blend of factual data and subjective framing to enhance credibility (Bonet-Jover, 2023). Datasets such as those used by Jeronimo, Marinho, Campelo, Veloso, and da

[Costa Melo \(2019\)](#) demonstrate how subjectivity markers can distinguish legitimate news from satirical or fabricated content. By identifying exaggerated language or evaluative bias, SD acts as a primary filter for veracity within automated fact-checking pipelines.

In Sentiment Analysis and Affective Computing, SD serves as an essential pre-processing step to filter non-evaluative content. Datasets like the IMDb/Rotten Tomatoes collection ([Pang & Lee, 2008](#)) have been instrumental in training systems to separate descriptive plot summaries from evaluative critiques. Recent evaluations of Large Language Models (LLMs) across affective computing tasks ([Amin, Mao, Cambria, & Schuller, 2024](#)) suggest that while general-purpose models show promise, specialized subjectivity encoders still maintain a performance advantage in identifying nuanced human perspectives.

Finally, SD resources are increasingly used in Semantic and Metalinguistic Reasoning. Recent studies have utilized established corpora to test how LLMs capture complex authorial perspective; for instance, [Ettinger, Hwang, Pyatkin, Bhagavatula, and Choi \(2023\)](#) utilized Abstract Meaning Representation (AMR) on synthetic data to dissect how modern language models capture semantic modulation, while [Beguš, Dąbkowski, and Rhodes \(2025\)](#) evaluated synthetic inputs through targeted prompting to track metalinguistic variations and parse non-literal language.

These applications highlight that SD is not merely a standalone labeling task but a vital structural layer that enhances model interpretability and reasoning across diverse downstream objectives. Nevertheless, the integration of subjectivity into these broader NLP tasks remains heavily constrained by the scarcity of native, multi-genre benchmarks that move beyond rigid binary distinctions to map the continuous pragmatic nuances of natural text.

Ref.	Lang.	Genre	Size	Task	Classes	Ann. Method
<i>Resources Primarily Developed for Subjectivity Detection (SD)</i>						
Wilson et al. (2017)	EN	News	629 d.	SD/Sent.	Multiclass	Human
Krüger et al. (2017)	EN	News/Letters	1.7k d.	SD	Binary	Lexicon+ML
Wilson et al. (2005)	EN	Lexicon	—	SD	Binary	Lexicon
Charniak et al. (2000)	EN	News/Letters	76k d.	SD	Binary	Human
Antici et al. (2021)	IT	Online News	1.8k s.	SD	Binary	Human
Struß et al. (2024)	Mult.	News articles	10k s.	SD	Binary	Human
Song et al. (2025)	EN	Wiki/News	4k s.	Metaling.	Multiclass	LLM
Volkova et al. (2013)	ES/RU	Tweets/Summ.	12k d.	Disinfo.	Binary	Lexicon (LIWC)
Vieira et al. (2020)	PT	News articles	200k d.	Disinfo.	Binary	Human Lexicon
<i>Resources Integrating SD for Downstream NLP Tasks</i>						
Pang and Lee (2008)	EN	Movie Reviews	2k d.	Sent. Anal.	Binary	Human
Jeronimo et al. (2019)	PT	News articles	3.6k d.	Fake News	Multiclass	Lexical Res.
Amin et al. (2024)	EN	Reviews/Plots	2k d.	Aff. Comp.	Binary	LLM (Persona)
Ettinger et al. (2023)	EN	Synthetic	30 s.	Sem. Mod.	Multiclass	Human (AMR)
Beguš et al. (2025)	EN	Synthetic	90 s.	Metaling.	Multiclass	LLM (Prompt)
Shokri et al. (2024)	EN	News articles	5.7k s.	Metaling.	Binary	LLM (CoT)

Table 2: Comparative analysis of resources related to Subjectivity Detection (SD). The table distinguishes between resources created for SD as a standalone task and those where SD is applied to downstream NLP applications. Units: s. (sentences) and d. (full documents). Tasks: SD (Subjectivity Detection), Metaling. (Metalinguistic Reasoning), Disinfo. (Disinformation), Aff. Comp. (Affective Computing), Sem. Mod. (Semantic Modifiers). Genre: “Synthetic” refers to sentences specifically authored for the study.

3 MUESCA Corpus: Creation and Annotation

This section details the methodology that drives the development and multi-level annotation of the MUESCA corpus. In order to translate our continuous theoretical model of subjectivity into a verifiable dataset, the process is structured into a modular pipeline, as illustrated in Figure 1. This step-by-step modular design is fundamental to our methodology, as it allows for the independent validation of linguistic criteria at each phase while ensuring the computational scalability of the final resource.

The process begins with the Methodological Design (Section 3.1), where the subjectivity-objectivity *continuum* is established through the strategic selection of genre anchors. This design is then populated through systematic Data Collection (Section 3.2), repurposing existing Spanish resources into a unified dataset.

The core of the methodology lies in the Subjectivity Annotation phase (Section 3), where subjectivity dimensions to capture author implication in the discourse (Section 3.3) are operationalized via a Hybrid Annotation Process (Section 3.4). This stage

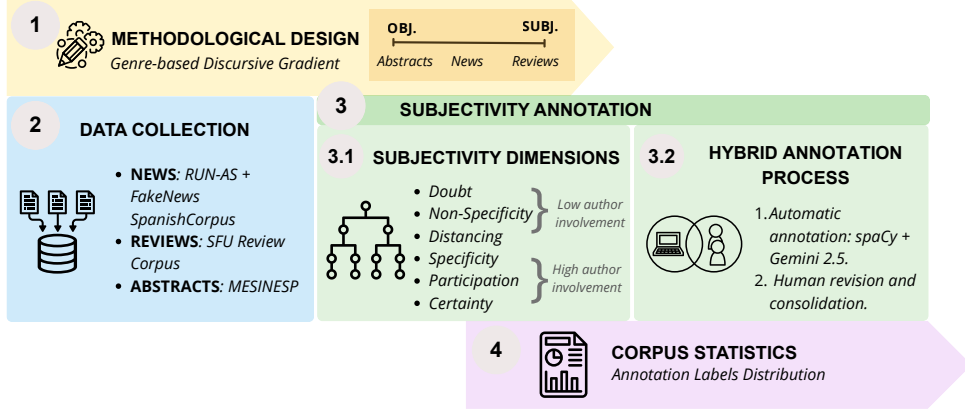


Fig. 1: Methodological workflow for the MUESCA corpus development, structured into four main phases: (1) Methodological Design, establishing the genre-based discursive gradient across abstracts, news, and reviews; (2) Data Collection from specialized Spanish corpora; (3) Subjectivity Annotation, where the low and high author involvement dimensions defined in 3.1 establish the formal labels applied during the hybrid expert-in-the-loop process in 3.2; and (4) Corpus Statistics, reflecting the final annotation labels distribution.

combines automated processing with human consolidation to ensure high-quality tagging. Finally, the resulting Corpus Statistics (Section 3.5) provide a quantitative overview of the distribution, establishing a validated resource for Spanish subjectivity detection.

3.1 Methodological Design

The development of a corpus requires a rigorous planning phase where the features of the dataset are predefined, starting by setting a clear objective to avoid the unproductive collection of random samples (Pérez Hernández, 2002). In this research, the primary objective is to develop a Spanish multi-genre corpus specifically designed for SD. Unlike existing multilingual or translated resources, this dataset is built exclusively from Spanish-language sources to capture the specific idiosyncratic markers of the language.

To ensure a robust corpus construction, our methodology adopts a multi-dimensional design that explicitly prioritizes linguistic representativeness, a foundational requirement established by Biber (1993). Following the methodological guidelines of Pérez Hernández (2002), the systematic design of any specialized corpus must initiate by defining six primary pillars: producer, receiver, medium, purpose, structure, and register. In the development of MUESCA, the textual genre acts as the overriding constraint that dictates the corpus construction. Consequently, rather than applying a rigid, uniform setup, we establish specific criteria tailored to each genre;

these fine-grained parameters and their operational definitions are systematically detailed in Table 4.

Feature	Scientific Abstracts	News Articles	Opinion Reviews
Producer	Researchers	Journalists	General Public
Receiver	Research Community	General Public	General Public
Medium	Scientific Journals	Digital Newspapers	Forums/ Platforms
Purpose	Knowledge dissemination	Reporting events	Evaluation/ Valuation
Structure	Rigid/ Standardized	Inverted Pyramid	Flexible/ Informal
Register	Academic/ Formal	Dynamic/ Standard	Informal/ Colloquial

Table 4: MUESCA methodological design: definition of genre characteristics based on selection criteria.

The architectural design of the MUESCA corpus operationalizes the subjectivity-objectivity *continuum* by integrating three distinct genres that represent varying degrees of authorial presence. This selection is strategically grounded in the diverse levels of individual expression inherent to different discursive genres (Calsamiglia & Valls, 2012). Following the principle that highly standardized genres are less susceptible to reflecting the speaker’s individuality (Bajtín, 1952), the corpus is structured as a discursive gradient defined by varying degrees of subjectivity operationalized by the author commitment fingerprint.

In this methodology, **scientific abstracts** are implemented as the objective pole, characterized by the “erasure of the self”, while **consumer reviews** establish the subjective pole through explicit personal involvement. Occupying the intermediate zone, **news articles** function as a hybrid space where formal objectivity (Tuchman & Aladro Vico, 1998) oscillates with the author’s subjective filtering. By anchoring the dataset at these distinct points, the corpus functions as a diagnostic tool capable of capturing the nuanced transitions between neutrality and evaluation.

The utility of such a multi-genre approach has been demonstrated by Vicente, Maestre, Lloret, and Cueto (2021), who found that analyzing varied textual typologies provides a richer feature set for discourse analysis from a NLP perspective. By utilizing this balanced design, the corpus moves beyond a simple collection of words to provide a representative sample of how subjectivity is encoded in Spanish, ranging from standardized institutional distance to explicit personal involvement.

To account for the distinct linguistic profiles of different genres, our methodology operationalizes subjectivity through six pragmatic dimensions, mapped onto a spectrum of high and low author commitment, rather than simple binary labels. This

genre-independent approach targets underlying functional traits, ensuring the robust identification of speaker perspective across diverse communicative registers.

3.2 Data Collection and Description

Once the corpus requirements are established, the primary texts must be selected for its construction. In this study, previously collected corpora are leveraged as reference material. This approach serves a dual purpose: first, the data quality is reinforced by the validated results of prior research; second, these existing resources provide the necessary raw data, which are subsequently enriched with complementary annotations. The texts were selected to encompass diverse domains while maintaining a minimum threshold for each genre. This approach ensures both topical heterogeneity and quantitative balance across the corpus.

The sources utilized for the collection of raw data are described below:

- **News articles.** This genre was sourced from existing corpora focused on disinformation detection. First, data were extracted from the RUN-AS corpus designed by Bonet-Jover, Sepúlveda-Torres, Saquete, Martínez-Barco, and Nieto-Pérez (2024); this resource is characterized by its high annotation granularity, categorizing texts into binary classes of reliability (reliable vs. unreliable) instead of strictly factual veracity. Concurrently, data were also collected from the FakeNewsSpanishCorpus by Posadas-Durán, Gómez-Adorno, Sidorov, and Escobar (2019), a resource utilized in two prominent shared tasks: MEX-A3T (IberLEF 2020) (Aragón et al., 2020) and FakeDeS (IberLEF 2021) (Gómez-Adorno, Posadas-Durán, Enguix, & Capetillo, 2021). As established by Tuchman and Aladro Vico (1998), news articles constitute a collection of facts structured and presented through the journalist’s interpretive lens.
- **Reviews.** This genre was obtained from the Spanish version of the SFU Review Corpus, designed by Taboada, Anthony, and Voll (2006). The corpus provides a robust foundation for our study, having been successfully employed in previous opinion-related research and classification benchmarks, such as Vicente et al. (2021).
- **Scientific abstracts.** This genre was retrieved from the MESINESP dataset developed for the lab “Evaluation of advanced hierarchical classification techniques for scientific literature, patents and clinical trials” (Gasco et al., 2021), part of the 2021 Conference and Labs of the Evaluation Forum (CLEF)¹. Specifically, texts were extracted from the “Sub-track 1: Scientific Literature Dataset,” which consists of abstracts from peer-reviewed scientific papers.

Following the data collection phase, a rigorous pre-processing pipeline was implemented. In line with previous studies such as Vieira et al. (2020) and Antici et al. (2021), our methodology adopts the sentence as the primary unit of analysis. This approach ensures a high level of granularity, which is essential for capturing the localized nature of subjective markers within larger discursive structures.

Once the texts were segmented into sentences, a manual revision process was conducted to ensure data integrity by removing typos, emojis, and unrecognizable characters. Crucially, we performed a manual quality check on the automatic sentence

¹<https://www.clef-initiative.eu/>

segmentation, which constitutes the foundational unit of our annotation. While formal genres responded well to automatic splitting, informal registers like consumer reviews frequently lack rigid punctuation, causing systematic segmentation mistakes. Consequently, manual correction was necessary, ensuring that every sentence was accurately isolated and well-defined for the subsequent annotation phase.

In Table 5 collected Spanish corpus is presented. As can be seen, the resulting corpus is divided into three balanced sub-corpus, named as *Sub_News*, *Sub_Review* and *Sub_Abstract*, reaching up to 1865 sentences collected in 153 documents.

Set	Texts	Sentences	Tokens	Avg. tokens / sentence
<i>Sub_News</i>	33	634	15,526	24.49
<i>Sub_Review</i>	29	614	15,861	25.83
<i>Sub_Abstract</i>	73	617	15,935	25.83
Total	135	1,865	47,322	25.87

Table 5: Corpus statistics on raw data.

3.3 Theoretical Grounding and Annotation Scheme

Subjectivity is a multifaceted phenomenon constructed through diverse linguistic markers. Our methodology operates at multiple levels to capture the speaker’s perspective, grounded in two primary theoretical pillars: metadiscourse (Hyland, 2005) and modality (Meunier, 1974).

Metadiscourse provides the functional basis for our scheme, focusing on linguistic elements that reflect authorial stance and manage reader interpretation (Hyland, 2005). We specifically adopt Hyland’s distinction between *hedges* and *boosters* to model the tension between low and high author commitment. While hedges signal epistemic uncertainty and caution, boosters reinforce authorial intent to guide the reader’s interpretation (Hyland, 2005). This interplay allows the discourse to reflect the speaker’s unique individuality, aligning with Bajtín’s (2013) view that every text is an inseparable expression of its author’s subjectivity.

Complementing this, we utilize the concept of modality as defined by Meunier (1974), distinguishing between modalities of enunciation (the speaker-listener relationship) and modalities of the utterance (the speaker’s personal view). Within the latter, epistemic modality is crucial, as it indicates the speaker’s degree of responsibility for the statement (Zorraquino, 1999).

Drawing from these foundations, the MUESCA methodology is structured around three primary pragmatic axes, as shown in Figure 2. Following Zorraquino (1999), these axes contrast high author presence against objective detachment through six granular dimensions:

- **Degree of Certainty:** Evaluates epistemic commitment to the proposition (Doubt vs. Certainty).
- **Level of Concreteness:** Distinguishes between vague discourse and text anchored in verifiable data (Non-specificity vs. Specificity).
- **Author Involvement:** Measures the enunciative distance between the producer and the content (Distancing vs. Participation).

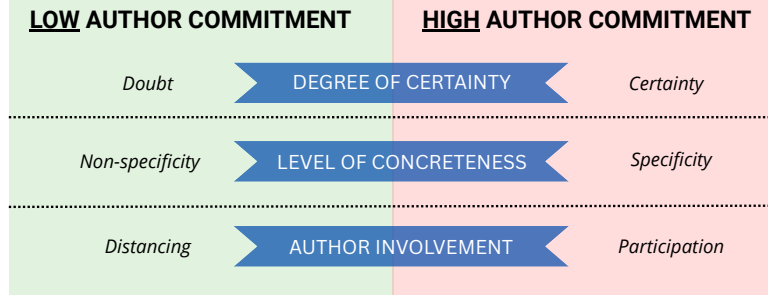


Fig. 2: The MUESCA Subjectivity methodology. The diagram illustrates the two macro-categories of authorial commitment (Low [L] and High [Hi]) intersected by three pragmatic axes. Each axis is defined by a binary tension between opposing dimensions: *Doubt vs. Certainty*, *Non-specificity vs. Specificity*, and *Distancing vs. Participation*.

3.3.1 Annotation Scheme

Building on the established methodology, the annotation process systematically identifies the lexical and grammatical units that operationalize our subjectivity dimensions. While the theoretical axes provide the conceptual foundation, the annotation scheme targets tangible pragmatic markers signaling authorial commitment or detachment.

Table 7 details this taxonomy, mapping the six pragmatic dimensions to specific linguistic features. This granular approach enables the MUESCA corpus to capture the subjectivity-objectivity *continuum* through the identification of pragmatic markers.

To illustrate the interaction between linguistic categories and author commitment in practice, we present representative examples from our corpus across different genres:

Example 1: Abstract (Objectivity Anchor)

Durante la prueba de la sonda, se observó un aumento a 1000 mg/kg. Nuestros resultados sugieren que representa un riesgo.

(During the *probe trial*, an increase at 1000 mg/kg was observed. Our results suggest that it represents a risk.)

In this excerpt, three distinct markers are annotated:

- **se observó /was observed:** Impersonal Reflexive Passive. This marker signals a strategy of *Distancing* within the *Low Author Commitment* category, as the agent avoids direct accountability for the observation.

Dimension	Linguistic Markers and Grammatical Categories
<i>Degree of Certainty</i>	
Doubt (L)	Question marks Conditional tense Subjunctive mood Verbs of belief Stressed interrogative pronoun Conditional adverb “ <i>if</i> ” Adverbs indicating possibility
Certainty (Hi)	Adverb “ <i>si</i> ” Verbs of knowledge Verb “ <i>ser</i> ” (to be) Verb “ <i>estar</i> ” (to be) Verb “ <i>tener</i> ” (to have) Imperative Future tense Periphrases of obligation
<i>Level of Concreteness</i>	
Specificity (Hi)	Numbers and digits Proper names and Named Entities Technical terms
Non-specificity (L)	Generalizing terms Indefinite pronoun Temporal construction
<i>Author Involvement</i>	
Participation (Hi)	First-person verb Second-person verb First-person pronoun Second-person pronoun
Distancing (L)	Impersonal reflexive passive Passive voice Third-person possessive pronoun

Table 7: Detailed taxonomy of linguistic dimensions and associated markers. Each domain categorizes individual grammatical features while maintaining a highly condensed vertical layout. (L) signals Low Author Commitment and (Hi) signals High Author Commitment.

- **1000 mg/kg:** Numerical referent. This particular number denotes *Specificity* within the *High Author Commitment* category, anchoring the text in objective data.
- **Nuestros / Our:** First-person pronoun is indicating explicit author *Participation*, within the *High Commitment* category.

- **representa** /*it represents*: Verb of belief/epistemic evaluation. This verb represents maximum *Certainty* within the *High Author Commitment* category, where the authors state the risk as a factual consequence of their findings.

Example 2: News (Hybrid Zone)

El Congreso ha respaldado la ley para regular la eutanasia que **se convierta** en prestación. Las federaciones piden que **se investigue si** los autores incurrieron en un delito.

(Congress approved the euthanasia law that may become a public service. Federations request an investigation into whether the authors committed a crime.)

This news item illustrates the domain of uncertainty:

- **se convierta/come** and **se investigue/ request**: Triggered by subjunctive mood markers. They map the *Doubt* spectrum, capturing a *Low Author Commitment* level regarding prospective or unrealized events.
- **si/if**: Conditional marker that reinforces the tentative nature of the statement. This marker signals *Doubt*, within the *Low Author Commitment* category.

Example 3: Review (Subjectivity Anchor)

Mi hermano **creo** que **podría** haberlo escrito mejor. Cuando me **enteré** que el escritor era médico...

(I think my brother thinks could have written it better. When I found out that the author was a doctor...)

The informal review genre presents sharp contrasts in author involvement:

- **creo /I think**: The use of the first-person signals *High Author Commitment* through the *Participation* dimension, while the semantic value of this verb of belief is related to the *Doubt* dimension within the *Low Author Commitment* category. In this case find a word that is compelling different pragmatic information.
- **podría /could have written**: The conditional inflection function as tentative devices, signaling *Doubt* within the *Low Author Commitment* category.
- **enteré /I found out**: First-person past tense verb is indicating explicit author *Participation*, within the *High Commitment* category.

3.4 Hybrid Annotation Process

This research employs a expert-in-the-loop strategy to balance computational scalability with the high linguistic precision required for subjectivity detection. The methodology is operationalized in two primary phases: the creation of a gold standard and the subsequent hybrid deployment.

3.4.1 Phase 1: Expert Gold Standard and Pilot Design Assessment

To establish a rigorous baseline, a gold-standard reference was manually annotated using INCEpTION (Klie, 2018). A representative subsample of 340 sentences was selected, exceeding the minimum sample size ($n = 319$) required for a 95% confidence level and a 5% margin of error for the total population of 1,865 sentences

(Cochran, 1977). This process was conducted by an expert annotator to ensure internal consistency and unified interpretative criteria across the complex pragmatic dimensions of the corpus. Crucially, this subsample functioned as a pilot test to assess the discourse analysis capabilities of different computational approaches, balancing traditional Machine Learning (ML) pipelines against Large Language Models (LLMs). This diagnostic phase revealed that while ML models provide reliable baseline morphosyntactic and rule-based tracking, they fall short at the pragmatic level. Conversely, LLMs proved essential for capturing complex linguistic constraints, such as specific terminology and some verbal moods (e.g., the subjunctive) that automatic rule-based systems failed to adequately detect. These preliminary findings provided the empirical justification for developing our hybrid annotation pipeline.

3.4.2 Phase 2: Hybrid Labeling and Expert Validation

Based on the benchmarks from Phase 1, a hybrid annotation pipeline was applied to the remainder of the corpus. This stage utilized a collaborative workflow: initial labels were generated by the automated systems (spaCy and Gemini) to act as a diagnostic proposal, which was then subjected to a 100% manual validation. Throughout this process, the expert annotator functioned as a final judge, reviewing every automated tag for the 1,865 sentences. This expert-in-the-loop refinement allowed for a qualitative discrepancy analysis, where the expert corrected or discarded automated tags based on the established guidelines (Section 3.3), ensuring that the final corpus retains the rigor of a manual dataset with the systematicity of an automated methodology.

To assess the reliability of the hybrid annotation process, the inter-methodological agreement between the manual gold-standard and the final hybrid-expert output was measured using Krippendorff’s Alpha (α). The hybrid annotation methodology achieved an overall score of 0.72. Following the benchmarks established by Krippendorff (2004), where $\alpha > 0.8$ denotes high reliability and values between 0.67 and 0.8 are considered acceptable for drawing tentative yet significant conclusions, this score indicates a substantial level of consistency. This result validates the capacity of the hybrid methodology to maintain linguistic rigor while significantly mitigating the temporal constraints of manual annotation. Consequently, the resulting MUESCA corpus is categorized as a validated Silver Standard Corpus (SSC) (Chowdhury & Lavelli, 2011), offering a robust and scalable resource for subjectivity detection in Spanish. This corpus qualification differs from a Gold Standard Corpus (GSC), as it is not created exclusively by human annotators (Chowdhury & Lavelli, 2011).

3.5 MUESCA Corpus Overview & Statistics

Following the implementation of the hybrid annotation methodology, the resulting MUESCA corpus statistics are detailed in Table 8. To ensure transparency and reproducibility, the annotated dataset and the comprehensive annotation guidelines have been made available for research purposes in an open-access repository (*blind for review purposes*).

Dimension / Marker	Global Stats			Density % (Genre)		
	Freq.	% Tag	% Tok.	Abs	News	Rev
Certainty (Hi)	1,668	15.17	3.52	2.35	3.96	4.28
Adverb “ <i>si</i> ”	33	0.30	0.07	0.01	0.06	0.14
Verbs of knowledge	349	3.17	0.74	0.67	0.79	0.75
Verb “ <i>ser</i> ” (to be)	743	6.76	1.57	1.27	1.34	2.09
Verb “ <i>estar</i> ” (to be)	146	1.33	0.31	0.13	0.46	0.34
Verb “ <i>tener</i> ” (to have)	166	1.51	0.35	0.20	0.28	0.57
Imperative	34	0.31	0.07	0.01	0.10	0.11
Future tense	171	1.55	0.36	0.03	0.81	0.26
Periphrases of obligation	26	0.24	0.05	0.04	0.12	0.01
Participation (Hi)	958	8.71	2.02	0.24	1.00	4.81
First-person verb	432	3.93	0.91	0.20	0.35	2.18
Second-person verb	96	0.87	0.20	0.00	0.13	0.48
First-person pronoun	355	3.23	0.75	0.04	0.46	1.75
Second-person pronoun	75	0.68	0.16	0.00	0.07	0.40
Specificity (Hi)	4,822	43.84	10.19	15.04	9.95	5.55
Numbers and digits	705	6.41	1.49	2.74	1.06	0.66
Proper names and Named Entities	1,545	14.05	3.26	2.20	4.86	2.77
Technical term	2,572	23.39	5.44	10.10	4.04	2.12
High Author Commitment	7,448	67.72	15.73	17.63	14.91	14.64
Doubt (L)	697	6.34	1.47	0.45	2.02	1.97
Question marks	40	0.36	0.08	0.00	0.14	0.11
Conditional tense	75	0.68	0.16	0.08	0.25	0.15
Subjunctive mood	348	3.16	0.74	0.26	1.09	0.86
Verbs of belief	126	1.15	0.27	0.09	0.22	0.49
Stressed interrogative pronoun	3	0.03	0.01	0.00	0.00	0.02
Conditional adverb “ <i>if</i> ”	95	0.86	0.20	0.02	0.29	0.30
Adverbs indicating possibility	10	0.09	0.02	0.00	0.02	0.04
Non-specificity (L)	1,620	14.73	3.42	2.01	3.39	4.88
Generalizing terms	1,200	10.91	2.54	1.69	2.50	3.42
Indefinite pronoun	390	3.55	0.82	0.31	0.75	1.42
Temporal construction	30	0.27	0.06	0.01	0.14	0.04
Distancing (L)	1,233	11.21	2.61	2.33	2.79	2.70
Impersonal reflexive passive	681	6.19	1.44	1.74	1.28	1.29
Passive voice	109	0.99	0.23	0.24	0.26	0.19
Third-person possessive pronoun	443	4.03	0.94	0.35	1.25	1.22
Low Author Commitment	3,550	32.28	7.50	4.79	8.20	9.55
Total/Mean	10,998	100	23.23	22.42	23.11	24.19

Table 8: Granular distribution of subjectivity markers across subcorpora (1,865 sentences). Freq.: Absolute Frequency; % Tag: Percentage relative to total annotations; % Tok.: Percentage relative to total running tokens. Subcorpora density metrics are divided by genre: Abs (Scientific Abstracts), News (News Articles), and Rev (Consumer Reviews). (L) signals Low Author Commitment and (Hi) signals High Author Commitment.

The corpus exhibits a significant annotation density, with nearly one in four words (23.23%) carrying a tagged element related to subjectivity. The data reveals a clear

polarization between genres: Abstracts are defined by high Specificity (15.04%) and minimal Participation (0.24%), whereas Reviews show the opposite trend, with a relevant increase in personal participation (4.81%). These findings confirm that scientific and evaluative texts hold opposite ends of the subjectivity gradient, leaving news articles as a hybrid category that balances technical precision with involvement markers.

4 Validation and Experimental Results

Following the construction and annotation of the MUESCA corpus, an evaluation methodology is established to provide empirical evidence of the resource’s quality and utility. This evaluation is designed to validate: (1) the effectiveness of the cross-genre design in capturing the diversity of subjective expressions, and (2) the discriminatory potential of the pragmatically-oriented subjectivity dimensions.

In the first stage (Section 4.1), a statistical characterization of the corpus is conducted. By employing the Kruskal-Wallis test [Kruskal and Wallis \(1952\)](#), Spearman’s correlation heatmaps ([Spearman, 1961](#)), and Principal Component Analysis (PCA) ([Hotelling, 1933](#)), the capacity of the proposed subjectivity dimensions to differentiate genres is examined.

The second stage (Section 4.2) implements a classification benchmark to evaluate the predictive power of the annotated dimensions across various computational architectures. The performance of traditional machine learning models is assessed by integrating the proposed subjectivity dimensions with semantic vectors extracted from BETO ([Cañete et al., 2023](#)), a Spanish pre-trained transformer. By utilizing BETO as a feature extractor, the complementary nature of manual annotations and automated embeddings is examined. Furthermore, an ablation study is performed to identify the specific combinations of dimensions that yield the highest discriminatory quality.

The methodology concludes with a multigeneric validation of the subjectivity continuum in Section 4.2.4; in this stage, the corpus is utilized as a diagnostic tool to measure the pragmatic alignment of the hybrid genre news relative to the polar extremes of factual abstracts and opinionated reviews.

4.1 Statistical Profiling of Subjectivity Dimensions

This section provides a statistical analysis of the subjectivity dimensions to evaluate their discriminatory power across the objectivity-subjectivity spectrum. While our methodology is theoretically structured around two macro-categories (High and Low Author Commitment) this analysis focuses on the specific dimensions within them to achieve a higher level of descriptive granularity. Throughout the subsequent results and discussion, each dimension is explicitly labeled as belonging to either the High (Hi) or Low (L) commitment pole.

To ensure comparability across the multi-genre corpus and to prepare the data for machine learning tasks, a Z-score normalization was applied to the annotation frequencies. Each feature was transformed such that it possesses a mean of 0 and a standard deviation of 1, as defined by the formula $z = (x - \mu)/\sigma$ ([Han, Pei, & Tong, 2022](#)). This standardization process eliminates potential biases stemming from

different feature scales and varying sentence lengths, allowing the classifiers to evaluate the relative importance of each subjective marker based on its statistical variance rather than its absolute frequency.

4.1.1 Kruskal-Wallis Statistical Analysis

The statistical significance of the proposed linguistic markers is assessed using the Kruskal-Wallis H-test (Kruskal & Wallis, 1952). This non-parametric method was selected because the frequency distributions of our subjectivity dimensions do not follow a normal distribution. The analysis is conducted at two levels of granularity: first, at the macro-dimensions level (e.g., Doubt, Certainty, Participation) to confirm the overall validity of the proposed pragmatic axes; and second, at the individual marker level (e.g., subjunctive mood, generalizing terms, named entities) to identify which specific markers carry the highest diagnostic weight. Following standard conventions in corpus linguistics and computational discourse analysis (Biber, 1991), a confidence threshold of $p < 0.05$ is established as the baseline for statistical significance. This threshold provides an optimal balance between statistical power and the mitigation of Type I errors when verifying that the observed variations among Abstracts, Reviews, and News are linguistically meaningful.

Dimension	H-Statistic	p-value	Significant
Participation (Hi)	480.41	< 0.001	✓
Specificity (Hi)	360.70	< 0.001	✓
Doubt (L)	134.04	< 0.001	✓
Non-specificity (L)	133.91	< 0.001	✓
Certainty (Hi)	79.12	< 0.001	✓
Distancing (L)	0.94	0.624	n.s.

Table 10: Statistical Significance Analysis (Kruskal-Wallis) using Macro-Dimensions. The H-statistic measures the strength of the difference between groups (Abstracts, News, Reviews), while ✓ denotes significant results ($p < 0.05$) and n.s. indicates non-significance.

As summarized in Table 10, the Kruskal-Wallis H-test identifies Participation ($H = 480.41$) and Specificity ($H = 360.70$) as the dominant dimensions across the corpus. The high magnitude of these H -statistics indicates that these two dimensions exhibit the most drastic distributional divergence among the three text genres. In pragmatic terms, this significance proves that the presence of personal involvement (Participation) and information density (Specificity) are not uniformly distributed, but rather serve as the primary linguistic axes that statistically polarize and differentiate Abstracts, News, and Reviews.

To further examine how individual linguistic tags contribute to the macro-dimensions, a granular evaluation at the marker level was performed (see Table 12). This analysis identifies which specific components carry the highest diagnostic weight within each category.

Dimension / Marker	H-Stat.	p-value	Signif.
Doubt (L)			
Question marks	20.63	< 0.001	✓
Conditional tense	11.93	0.003	✓
Subjunctive mood	55.94	< 0.001	✓
Verbs of belief	50.23	< 0.001	✓
Stressed interrogative pronoun	6.12	0.047	✓
Conditional adverb “if”	40.67	< 0.001	✓
Adverbs indicating possibility	7.57	0.023	✓
Certainty (Hi)			
Adverb “ <i>st</i> ”	20.91	< 0.001	✓
Verbs of knowledge	1.36	0.508	n.s.
Verb “ <i>ser</i> ” (to be)	58.16	< 0.001	✓
Verb “ <i>estar</i> ” (to be)	24.65	< 0.001	✓
Verb “ <i>tener</i> ” (to have)	33.01	< 0.001	✓
Imperative	12.38	0.002	✓
Future tense	95.92	< 0.001	✓
Periphrases of obligation	14.04	0.001	✓
Specificity (Hi)			
Numbers and digits	83.04	< 0.001	✓
Proper names and Named Entities	90.84	< 0.001	✓
Technical term	575.16	< 0.001	✓
Non-specificity (L)			
Generalizing terms	80.90	< 0.001	✓
Indefinite pronoun	102.59	< 0.001	✓
Temporal construction	21.30	< 0.001	✓
Participation (Hi)			
First-person verb	296.74	< 0.001	✓
Second-person verb	60.52	< 0.001	✓
First-person pronoun	287.44	< 0.001	✓
Second-person pronoun	83.12	< 0.001	✓
Distancing (L)			
Impersonal reflexive passive	27.67	< 0.001	✓
Passive voice	1.79	0.409	n.s.
Third-person possessive pronoun	59.03	< 0.001	✓

Table 12: Kruskal-Wallis test results for individual linguistic markers across genres. The H-statistic measures the strength of the cross-genre differences, where ✓ denotes significant variations ($p < 0.05$) and n.s. indicates non-significance.

Within Specificity, technical terms emerge as the most powerful discriminator in the corpus ($H = 575.16$), providing a clear linguistic boundary for objective discourse. Conversely, Non-specificity is strongly defined by indefinite pronouns ($H = 102.59$) and generalizing terms ($H = 80.90$), which serve as markers of lower concreteness. The Participation dimension is heavily driven by the presence of first-person verbs (H

= 296.74) and pronouns ($H = 287.44$), signaling explicit personal involvement Doubt is primarily characterized by subjunctive mood verbs ($H = 55.94$) and verbs of belief ($H = 50.28$), while Certainty is mostly identified by the use of the future tense ($H = 95.92$) and the verb ‘to be’ (*ser*). In contrast, Distancing relies on third-person possessive pronouns ($H = 59.03$) and the impersonal reflexive passive ($H = 27.67$) to minimize authorial involvement.

The analysis also highlights certain markers that lack statistical significance, such as verbs of knowledge ($p = 0.508$) and the passive voice ($p = 0.409$). This bottom-up validation confirms that while the macro-dimensions function as holistic categories, their predictive utility is rooted in specific linguistic markers that map onto the subjectivity-objectivity *continuum*.

4.1.2 Dimensions and Markers Correlations using Spearman’s Heatmaps

The internal structure and dependencies of the annotation scheme were evaluated through Spearman’s correlation analysis (Spearman, 1961), a method chosen for its capacity to detect non-linear relationships between linguistic markers and dimensions. The resulting correlation matrices are presented in Figure 3 for macro-dimensions and Figure 4 for individual markers. In these heatmaps, red-toned squares indicate a positive correlation, signaling systematic co-occurrence patterns where specific linguistic markers reinforce each other to constitute unified discursive styles. Conversely, blue-toned squares denote a negative relationship, suggesting structural mutual exclusivity between competing pragmatic strategies. The color intensity represents the association’s magnitude, thereby providing an empirical mapping of how individual markers cluster together. Establishing these correlations allows us to identify the underlying multi-dimensional axes of the corpus, proving that subjectivity is not driven by isolated lexical choices but by combinations of interconnected linguistic features.

The macro-dimensions analysis (Figure 3) highlights the tension between the poles of the subjectivity *continuum*. The most significant negative correlation occurs between Specificity and Participation (-0.28), confirming that higher authorial presence typically leads to a decrease in technical specificity. Conversely, the strongest positive correlation is found between Participation and Doubt (0.25), suggesting that personal involvement in our corpus is frequently expressed through markers of uncertainty and belief. Furthermore, Certainty and Doubt show a notable positive alignment (0.19), indicating that both poles frequently co-occur within the same textual unit. This statistical relationship highlights the multi-layered nature of subjectivity: authors do not maintain a uniform epistemic commitment throughout a text. Instead, they alternate between assertive framing (Certainty) and speculative hedging (Doubt). This alignment confirms that in natural language, subjective positioning operates as a fluid negotiation rather than a binary, mutually exclusive choice.

Beyond macro-dimensions, the detailed matrix in Figure 4 reveals the functional dependencies between individual markers, providing empirical support for the theoretical grouping of our annotation scheme. Specifically, the positive correlations observed among markers assigned to the same macro-dimension (such as verbs in the subjunctive mood, verbs of belief, and conditional constructions within the Doubt axis)

confirm their statistical cohesion. This alignment demonstrates that our categories are not arbitrary taxonomies, but rather reflect structurally interconnected linguistic features that operate in group to fulfill specific pragmatic functions. While most markers within the same dimension exhibit high consistency, the correlations between different categories also confirm the heatmap’s representativeness. A clear example is the positive correlation between Imperative verbs (Certainty) and Second-person pronouns (Participation) (0.18). Since imperative forms are directed at a second-person subject, this statistical alignment captures a grammatical reality. Within the Participation dimension, the strong correlation between First-person verbs and First-person pronouns (0.38) reinforces the subjective density of the texts. Similarly, in the Certainty category, a significant link is observed between the Future tense and Verbs of knowledge (0.15). Within Doubt, the co-occurrence of the Conditional “if” and the Subjunctive mood (0.16) further supports the structural coherence of the methodology. These patterns prove that the annotated markers are not isolated occurrences but rather cohesive linguistic patterns that consistently define the discursive nature of the MUESCA corpus.

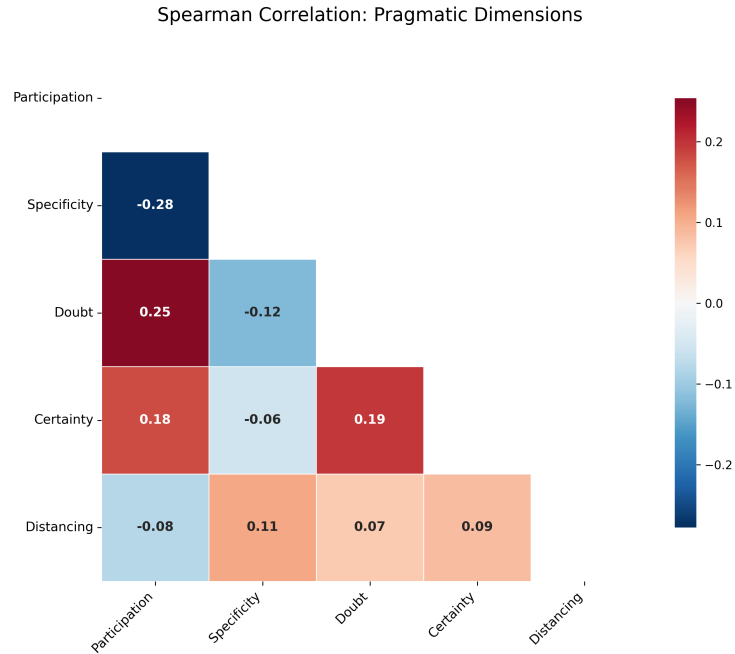


Fig. 3: Spearman correlation heatmap of consolidated pragmatic dimensions. This macro-level analysis shows the relationships between the functional categories defined in the methodology. Note that the *Non-specificity* dimension was excluded from this aggregate analysis due to near-zero variance within the selected sub-corpus, which precludes the calculation of meaningful monotonic relationships.

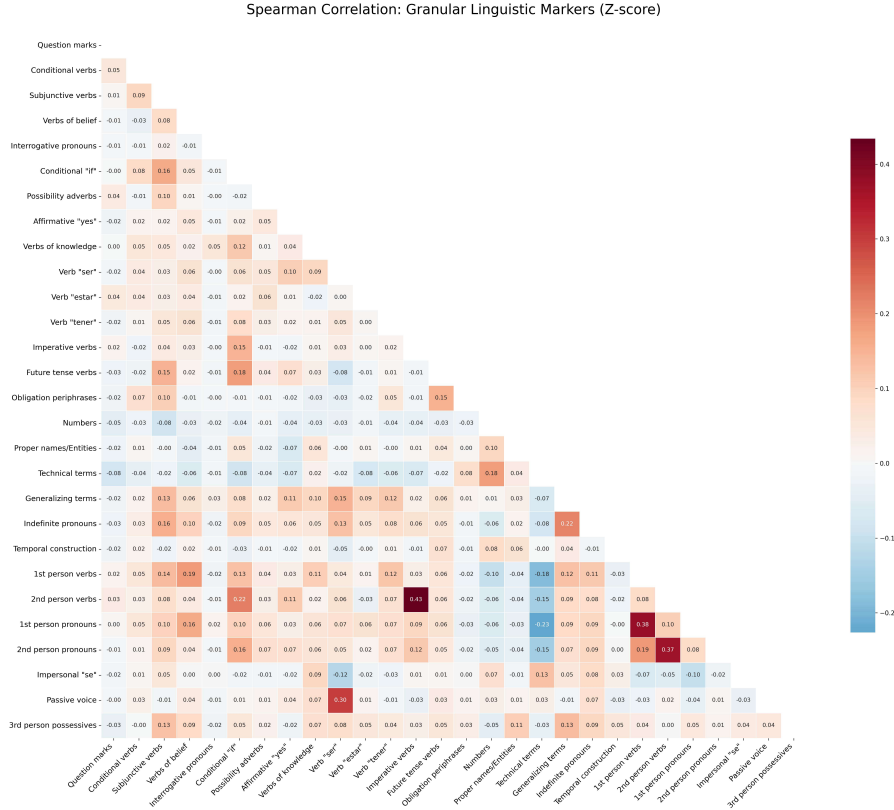


Fig. 4: Spearman correlation heatmap of Z-score normalized individual linguistic markers across the tri-genre corpus. This granular view illustrates the internal dependencies and co-occurrence patterns between specific linguistic tags.

4.1.3 Genre Clustering and PCA Visualization

Finally, a Principal Component Analysis (PCA) (Hotelling, 1933) is performed to visualize the spatial distribution of the MUESCA corpus (Figure 5). By projecting the multi-dimensional annotations into a two-dimensional plane, this analysis reveals how texts naturally cluster.

The PCA visualization (Figure 5) maps the MUESCA corpus onto a spatial gradient that spans from objectivity to subjectivity. Principal Component 1 (8.97% of variance) serves as the primary axis of discrimination, positioning Abstracts at the negative extreme, while Reviews move toward the positive pole as personal involvement increases. Occupying the central area, the News subcorpus exhibits a distribution that overlaps with both poles, confirming its status as a hybrid zone where objective and subjective discourse coexist. Furthermore, Principal Component 2 (5.72% of variance) captures the internal consistency of each genre. The clustering of Abstracts suggests a highly standardized discourse, whereas the broader dispersion of News and

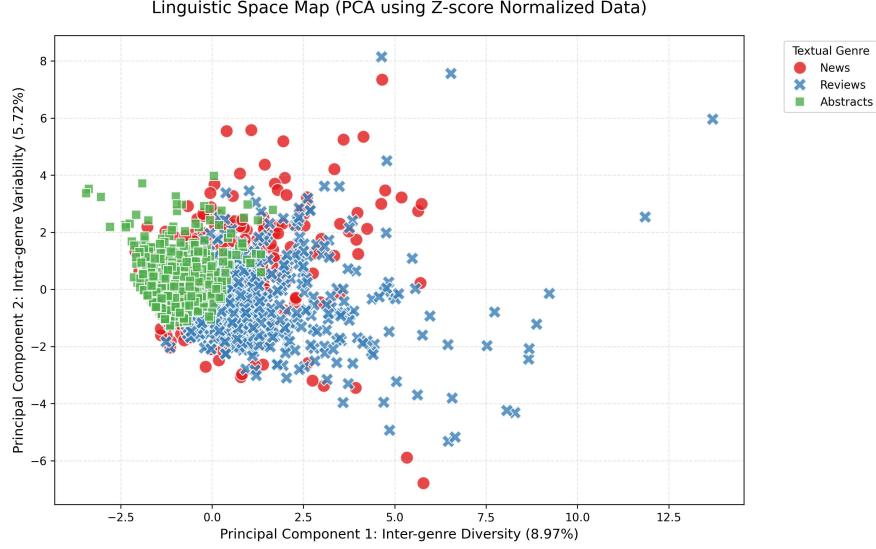


Fig. 5: Principal Component Analysis (PCA) of Subjectivity Dimensions across Genres. Principal Component 1 represents variation within objectivity-subjectivity *continuum* and Principal Component 2 represents variations inside each textual genre.

Reviews reveals greater stylistic flexibility and discursive heterogeneity. This multi-dimensional alignment into three distinct discursive zones confirms that each genre maintains a predictable position within the linguistic space, providing an empirical basis for the classification tasks presented in the following Section 4.2.

4.2 Classification Performance and Feature Evaluation

The second stage of this validation methodology assesses the predictive utility of the annotated markers through a multi-level classification benchmark (comprising objective, hybrid, and subjective levels). In this setup, the three genres serve as representative categories for the subjectivity-objectivity *continuum*: Abstracts represent the objective pole, Reviews denote subjectivity, and News acts as a hybrid intermediate category. Grounded in the statistical findings of Section 4.1, this experiment moves beyond theoretical definitions to evaluate how pragmatically-oriented dimensions enhance the capacity to classify subjectivity across diverse genres.

The experimental design is structured as a sequential process aimed at both technical optimization and linguistic mapping. First, the most effective lexical representation and classifier are established; these optimal settings are subsequently utilized to perform an ablation study of the subjectivity dimensions. Furthermore, the complementarity between our annotated features and deep-learning semantic embeddings (BETO) is explored to determine the unique contribution of pragmatic markers over raw semantic vectors.

Ultimately, this comparison is operationalized through two primary scenarios to evaluate whether these pragmatic macro-dimensions provide a more robust and theoretically-grounded methodology than raw text alone:

- **Baseline (Raw Textual Features):** The models utilize standard lexical representations (TF-IDF/BoW or BERT-based embeddings) to detect subjectivity through general word-level patterns.
- **Pragmatic Dimension Enrichment:** The models are enriched with the six subjectivity dimensions (Doubt, Non-specificity, Distancing (L) and Certainty, Specificity, Participation (Hi)).

This methodology culminates in a cross-genre validation setup designed to test the robustness and generalizability of our pragmatic features across the subjectivity *continuum*. To achieve this, the evaluation is operationalized using a *Leave-One-Genre-Out* strategy, where the classification models are systematically trained on two genres and evaluated on the third, remaining genre as an unseen test set.

By analyzing the fluctuations in a multi-level classification task (comprising objective, hybrid, and subjective levels) performance across these setups, the experiment serves a dual purpose. Methodologically, it verifies whether our annotated pragmatic markers remain predictive when applied to a completely different text type. Theoretically, it quantifies the exact position of the hybrid news genre within the *continuum*; by observing whether a model trained on News generalizes better to the objective pole (Abstracts) or the subjective extreme (Reviews), we can mathematically map the alignment and fluid boundaries of journalistic discourse.

4.2.1 Baseline Configuration and Model Selection

In the initial phase of the evaluation, a feature extraction pipeline is established to determine the baseline performance of standard text representation. A Bag-of-Words (BoW) encoding approach is utilized to evaluate the extent to which subjectivity detection can be achieved through raw lexical frequency alone. To ensure a robust assessment, six classifiers are contrasted: SVM, Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), AdaBoost, and a Multi-layer Perceptron (MLP).

This stage is designed to measure the performance increase (Gain (Δ)) obtained by enriching the raw textual BoW vectors with the Z-score normalized MUESCA dimensions. To ensure statistical robustness and mitigate potential bias, a 5-fold Stratified Cross-Validation is utilized over the entire corpus. This method guarantees that every document serves as both training and testing data across different iterations, providing a more reliable estimation of the model’s performance than a single train-test split.

The results presented in Table 9 reveal a significant performance distribution. The highest absolute performance is achieved by the MLP (0.8367), which effectively integrates the non-linear interactions between lexical frequencies and pragmatic markers. This superiority highlights that subjectivity is driven by combinations of features rather than isolated variables. The MLP architecture successfully captures these dynamics, recognizing that a pragmatic marker (e.g., the subjunctive mood) fundamentally modifies and amplifies the predictive weight of surrounding words depending on the context. While linear models like LR and SVM show marginal negative gains,

Model	F1 Baseline	F1 Enriched	Gain (Δ)
MLP	0.8192	0.8367	+0.0175
LR	0.8356	0.8301	-0.0055
SVM	0.7912	0.7898	-0.0013
RF	0.7620	0.7740	+0.0119
DT	0.6239	0.6598	+0.0358
AdaBoost	0.6131	0.6331	+0.0200

Table 13: Global performance comparison: BoW Baseline vs. Enriched MUESCA methodology (Macro-F1).

suggesting a degree of feature redundancy in high-dimensional spaces- However the tree-based models (DT, RF, and AdaBoost) exhibit the most substantial improvements. Specifically, the Decision Tree (DT) model shows a peak gain of +0.0358, indicating that explicit pragmatic dimensions provide an incipient structural guide for models that otherwise struggle with the sparsity of raw BoW data.

To further validate the subjectivity-objectivity *continuum* proposed in this study, a granular analysis was conducted on the optimal configuration (MLP + Enriched BoW). The performance breakdown by genre is presented in Table 14.

Genre	Baseline F1 (BoW)	Enriched F1 (MUESCA)	Gain (Δ)
Abstracts	0.8712	0.8849	+0.0137
News	0.7610	0.7851	+0.0241
Reviews	0.8255	0.8401	+0.0146
<i>Global Macro-F1</i>	<i>0.8192</i>	<i>0.8367</i>	<i>+0.0175</i>

Table 14: Detailed F1-score comparison per genre for the MLP configuration. Results highlight the pragmatic gain achieved by integrating MUESCA dimensions into the BoW baseline.

The classification of Abstracts yields the highest absolute performance ($F_1 = 0.8849$). This outcome is consistent with the highly standardized and objective structure of scientific discourse, allowing the BoW baseline to perform strongly even before enrichment.

Conversely, the News genre remains the most challenging task for the classifier, yet it exhibits the highest relative improvement (+0.0241). The integration of the MUESCA pragmatic dimensions effectively bridges the classification gap by providing explicit markers of author presence that raw lexical density alone fails to capture in this “heterogeneous space” of the subjectivity continuum.

At the subjective pole, Reviews demonstrate robust performance and a steady improvement ($F_1 = 0.8401$). This result confirms that across the three textual genres, the BoW component identifies the topical or evaluative content (the *what*), whereas

the enriched Z-score dimensions successfully characterize the linguistic construction of the evaluative stance (the *how*).

Consequently, the MLP with BoW and MUESCA enrichment is selected as the optimal fixed architecture for the subsequent transformer-based experiments, as it demonstrates the highest capacity to leverage pragmatic information across all textual genres.

4.2.2 Feature Analysis and Ablation Study

After establishing the core architecture, an ablation study is conducted to isolate the precise contribution of the pragmatic markers. The experimental setup involves testing all possible feature subsets against a raw Bag-of-Words (BoW) baseline. This mapping identifies the “optimal feature subset”: the specific combination of dimensions that maximizes Macro-F1 Gain. The results (Table 16) rank these combinations, revealing the specific weight that subjectivity dimensions carry.

Combination	F1 Avg	F1 Abs	F1 News	F1 Rev	Gain
<i>Top Performing Combinations</i>					
Spec + Cert	0.8381	0.8854	0.7831	0.8459	+1.89%
Spec + Dist	0.8379	0.8878	0.7851	0.8409	+1.87%
Part + Spec + Doubt + Cert + Dist	0.8367	0.8849	0.7851	0.8401	+1.75%
Part + Spec	0.8359	0.8870	0.7807	0.8399	+1.67%
Part + Spec + Cert + Dist	0.8356	0.8792	0.7826	0.8451	+1.64%
BoW Baseline	0.8192	0.8705	0.7588	0.8284	0.00%
<i>Performance Degradation</i>					
Doubt	0.8149	0.8674	0.7533	0.8241	-0.43%

Table 16: Ablation study of subjectivity dimensions (MLP model). Bold values indicate the best performance per column. Abbreviations for text genres and dimensions are defined as follows: Abs (Abstracts), News (News), Rev (Reviews), Part (Participation), Spec (Specificity), Doubt (Doubt), Cert (Certainty), and Dist (Distancing). Gain indicates the percentage change over the BoW baseline.

The findings show that integrating our subjectivity dimensions outperforms the BoW baseline. The most significant finding is the identification of the most discriminative pair: Specificity + Certainty, which achieved the highest global performance with a Macro-F1 of 0.8381 and a substantial gain of +1.89.

Notably, while the combination of all five dimensions remains highly competitive (F1 = 0.8367), the performance of simpler pairs like Spec + Dist and Spec + Cert indicates that Specificity acts as the primary stabilizer for the hybrid News genre, which reached its peak F1-score of 0.7851 in these configurations. Conversely, the Doubt dimension, when used in isolation or in certain three-feature clusters, tended to

introduce noise, even resulting in a negative gain (-0.0043) compared to the baseline. These findings reinforce the hypothesis that pragmatic markers maintain a balanced interaction to effectively guide the classifier.

4.2.3 Impact of Subjectivity Dimensions on BETO

The final stage examines the synergy between state-of-the-art neural embeddings and our subjectivity annotations. While BETO (Cañete et al., 2023) is highly proficient at capturing deep semantic relationships, it may overlook specific discursive nuances. To address this, a feature fusion architecture was implemented via vector concatenation; this allows the classification head to integrate internal Transformer representations with the proposed aggregated linguistic markers.

By comparing a raw BETO baseline against an enriched version via 5-fold stratified cross-validation, the experiment evaluates whether these high-level features provide a specialized “stylistic guide”. This setup determines if explicit subjectivity dimensions can assist models in resolving classification ambiguities, particularly within hybrid genres like News, thereby moving automated subjectivity analysis beyond purely data-driven, statistical approaches.

Metric	Baseline (BETO)	BETO + Best Combo*	Gain (Δ)
<i>Global Performance</i>			
Global Macro F1-Score	0.9353	0.9385	+0.0032
<i>F1-Score by Genre</i>			
F1 Abstracts	0.9609	0.9617	+0.0008
F1 Reviews	0.9388	0.9428	+0.0040
F1 News	0.9063	0.9111	+0.0048

*Best Combo: Specificity and Certainty.

Table 17: Comparative performance between the baseline BETO model and the optimal pragmatic configuration. Results confirm that even high-capacity transformers benefit from explicit markers in hybrid and subjective genres.

As shown in Table 17, the modest F1 gain over the BETO baseline (+0.0032) should be interpreted within the context of the “ceiling effect” (Colombo, Peyrard, Noiry, West, & Piantanida, 2023). Starting from a robust baseline of 0.9353, incremental gains are more difficult to achieve and statistically more significant than larger fluctuations on weaker models. As noted by Ethayarajh and Jurafsky (2020), once models reach such high thresholds, standard metrics often fail to capture qualitative improvements in domain-specific tasks. The ability of just two pragmatic dimensions (Specificity and Certainty) to refine a 110-million-parameter model suggests a potential synergy where MUESCA dimensions capture nuanced variance that may remain latent in large-scale semantic embeddings. This is particularly evident in the News genre, where the hybrid model achieves a peak gain of +0.0048, effectively resolving ambiguities in texts where journalistic neutrality and evaluative cues collide.

Beyond performance metrics, this approach improves model interpretability. While large models like BETO are powerful at identifying patterns in text form, they often rely on statistical shortcuts rather than a true understanding of pragmatically-oriented dimensions (Bender & Koller, 2020). By showing that dimensions like Specificity and Certainty act as key discriminators, this study provides a theoretical foundation for the model’s decisions. Ultimately, the MUESCA dimensions do not replace BETO’s statistical strength; instead, they provide “pragmatic grounding” that allows the model to detect subjective nuances that are otherwise overlooked by statistical analysis.

4.2.4 Cross-genre Validation: Mapping the Subjectivity Continuum

To empirically validate the proposed subjectivity dimensions, a forced-alignment experiment was conducted. This setup uses text genres as linguistic anchors to map the subjectivity-objectivity *continuum*. Rather than treating genres as static labels, the approach utilizes the polar extremes of the corpus: Scientific Abstracts (representing explicit objectivity) and Consumer Reviews (representing explicit subjectivity). Consequently, this setup enables the observation of how News articles (a hybrid discourse) are classified when the model is constrained to binary setups.

Four experimental scenarios were designed (Table 18). In Table 18, the target classification classes are represented by the columns **Obj.** (Objective), **Neu.** (Neutral), and **Subj.** (Subjective). The entry in each column indicates which corpus genre is assigned to that specific class label in a given scenario. A dash (“–”) indicates that the corresponding class role is omitted because that genre was excluded from the binary setup under a Leave-One-Genre-Out protocol. The remaining columns report model performance: **F1 (Enr)** denotes the Macro F1-score achieved by the Multi-Layer Perceptron (MLP) model when enriched with MUESCA pragmatic features, and **Gain** measures the absolute performance increase over the un-enriched baseline.

The four scenarios instantiate the following experimental logic:

- **Scenario 1 (Multiclass Baseline):** Includes all three genres simultaneously. Scientific Abstracts are mapped to Objective, News to Neutral, and Reviews to Subjective.
- **Scenario 2 (Polar Validation):** Excludes News (indicated by “–” in the Neutral column) and directly contrasts Abstracts against Reviews (Abs vs. Rev). This setup establishes an empirical upper bound for the separability of the two continuum endpoints.
- **Scenarios 3 & 4 (Forced Binary Projections):** Exclude one polar genre and assign News to the unoccupied role opposite the remaining pole. Scenario 3 excludes Abstracts (“–” in Obj.) and assigns News to the Objective class against Reviews. Scenario 4 excludes Reviews (“–” in Subj.) and assigns News to the Subjective class against Abstracts.

Comparing performance across these forced re-labelings provides the key analytical insight: a higher F1-score in a given binary pairing indicates that News is more readily separable from that specific pole, implying greater pragmatic proximity to the opposing pole.

Scenario	Obj.	Neu.	Subj.	F1 (Enr)	Gain
<i>Multiclass Classification</i>					
1. Full Corpus (<i>Abs:O, News:N, Rev:S</i>)	Abs	News	Rev	0.8367	+0.0175
<i>Binary Projections & Polar Validation</i>					
2. Polar Validation (<i>Abs vs Rev</i>)	Abs	–	Rev	0.9708	+0.0033
3. News as Obj. (<i>News vs Rev</i>)	News	–	Rev	0.8717	+0.0121
4. News as Subj.	Abs	–	News	0.9176	+0.0105

Table 18: Consolidated classification results for the MLP architecture across scenarios. Target class columns (**Obj.**, **Neu.**, **Subj.**) specify the genre assigned to each role per scenario; a dash (–) denotes role omission due to genre exclusion in binary setups. **F1 (Enr)** reports the Macro F1-score with enriched pragmatic features, and **Gain** denotes the absolute improvement over the raw text baseline.

The results in Table 18 demonstrate a pragmatic distinction between the polar extremes of Abstracts and Reviews (Scenario 2, $F1 = 0.9708$), indicating that the MUESCA dimensions effectively capture the characteristics of both scientific neutrality and evaluative nuances.

A key finding emerges from the performance asymmetry observed in the News genre. When distinguishing News from the objective pole (Scenario 4), the model achieves a higher F1-score (0.9176) than when distinguishing it from the subjective pole (Scenario 3, $F1 = 0.8717$). This indicates that the linguistic markers present in News are more readily separable from scientific objectivity than from evaluative reviews. Consequently, this suggests that journalistic discourse, despite formal conventions of neutrality, exhibits greater pragmatic overlap with subjective stances than with the detachment characteristic of scientific reporting. Overall, this cross-genre evaluation supports the utility of the proposed methodology as a diagnostic approach for analyzing hybrid textual genres.

5 Conclusion and Future Work

This research has demonstrated the utility of the MUESCA methodology in capturing subjectivity as a fluid *continuum* rather than a dichotomous category. By grounding our taxonomy in the functional concepts of metadiscourse (Hyland, 2005) and epistemic modality (Meunier, 1974), we have provided a systematic method to map authorial commitment through three primary axes: degree of certainty, level of concreteness, and author involvement. The integration of Hyland’s distinction between hedges and boosters with Zorraquino’s focus on the speaker’s responsibility has created a robust theoretical bridge, aligning with Bajtín’s (2013) view of text as an

inseparable expression of subjectivity. Rather than enforcing rigid classifications, this explicit linguistic grounding provides the explanatory depth needed to navigate hybrid discourses beyond surface-level labels.

The primary contribution of this research is the MUESCA Corpus, a novel multi-genre resource for Spanish subjectivity detection comprising 1,865 annotated sentences. The empirical value of this corpus was validated through Kruskal-Wallis analysis, which confirmed that its dimensions reflect distinct, statistically significant linguistic behaviors. By encompassing scientific abstracts, news articles, and consumer reviews, the corpus functions as a diagnostic tool that maps the subjectivity-objectivity gradient. The results identify a clear discursive flow where Abstracts and Reviews occupy polar extremes, while News emerges as a hybrid, transitional zone.

The computational utility of the methodology was established through machine learning experiments where an MLP model enriched with MUESCA dimensions reached an average performance of 83.81%. Notably, the integration of these dimensions with BETO semantic vectors yielded a global F1-score of 0.9385. Feature analysis identified traits like Specificity and Certainty as key indicators of a text’s position on the *continuum*. The performance gain observed in the News genre (+0.0048) underscores that while transformers excel at semantic representation, explicit linguistic knowledge is essential for navigating hybrid discursive zones where objective data and subjective stance frequently overlap.

Looking forward, we intend to scale this approach by expanding the size of the corpus to include on the one hand more sentences and, on the other hand, additional genres, such as social media and technical reports, to test the methodology’s resilience in polarized or informal environments. An amplified corpus will be used to fine-tune classification models to enhance the detection of implicit subjectivity. A promising application lies in automated deception detection, where dimensions such as Certainty and Doubt may address the inherent limitations of human judgment (Masip, Sporer, Garrido, & Herrero, 2005). By prioritizing the pragmatic delivery of information over its propositional content (Saquete, Tomás, Moreda, Martínez-Barco, & Palomar, 2020), future iterations of the MUESCA methodology can better navigate the high lexical overlap characteristic of deceptive discourse (Solà-Sales, Alzetta, Moret-Tatay, & Dell’Orletta, 2023).

Left blank for anonymous revision

Acknowledgements

This research is funded by a grant for the recruitment of predoctoral research staff (CIACIF/2023/106) from the Fondo Social Europeo Plus of Generalitat Valenciana - European Social Fund Plus of the Generalitat Valenciana. The research work is part of the R&D projects; “SAFEWORDS”: Language Anonymization with Ethical and Legal Safeguards through NLP (AIA2025-163322-C63); “Mecánica cuántica para la comprensión y generación del lenguaje” (PID2024-160791OB-I00) funded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE; Proyecto Desarrollo de Modelos ALIA within the framework of the Plan Nacional de Tecnologías de Lenguaje -ENIA 2024 del Ministerio para la Transformación Digital y de la Función Pública y PRTR,

NextGeneration EU, Resol. SEDIA 19.08.2024; “Criterios de Evaluación para Corpus de Calidad en Inteligencia Artificial (CRITERIA)”, developed in the II Concurso Nacional para la adjudicación de Ayudas a la Investigación en Humanidades 2025, with the topic “Humanidades Digitales” (Referencia: FRAHUMANIDADES25-01), funded by Fundación Ramón Areces.

Statements and Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Author’s A involvement with the funding organizations listed in the Funding section is limited to the receipt of research grants and predoctoral fellowships for the execution of this study.

References

- Amin, M.M., Mao, R., Cambria, E., Schuller, B.W. (2024). A wide evaluation of ChatGPT on affective computing tasks. *IEEE Transactions on Affective Computing*, 15(4), 2204–2212, <https://doi.org/10.1109/TAFFC.2024.3419593>
- Antici, F., Bolognini, L., Inajetovic, M.A., Ivasiuk, B., Galassi, A., Ruggeri, F. (2021). SubjectivITA: An Italian corpus for subjectivity detection in newspapers. *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12* (pp. 40–52).
- Aragón, M.E., Jarquín-Vásquez, H.J., Montes-y Gómez, M., Escalante, H.J., Pineda, L.V., Gómez-Adorno, H., ... Bel-Enguix, G. (2020). Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish. *IberLEF@SEPLN* (pp. 222–235).
- Bajtín, M. (1952). El problema de los géneros discursivos. *Semiología*, 85.
- Bajtín, M. (2013). *Estética de la creación verbal*. Siglo XXI Editores México.
- Beguš, G., Dąbkowski, M., Rhodes, R. (2025). Large linguistic models: Investigating LLMs’ metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*, 6(12), 3453–3467, <https://doi.org/10.1109/TAI.2025.3575745>
- Bender, E.M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).
- Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257, <https://doi.org/10.1093/lc/8.4.243>
- Bonet-Jover, A. (2023). *Abordando el tratamiento automático de la desinformación: modelado de la confiabilidad en noticias mediante procesamiento del lenguaje natural* (Unpublished doctoral dissertation). Universidad de Alicante.
- Bonet-Jover, A., Sepúlveda-Torres, R., Saquete, E., Martínez-Barco, P., Nieto-Pérez, M. (2024). RUN-AS: a novel approach to annotate news reliability for disinformation detection. *Language Resources and Evaluation*, 58(2), 609–639, <https://doi.org/10.1007/s10579-023-09678-9>
- Calsamiglia, H., & Valls, A.T. (2012). *Las cosas del decir*. Barcelona: Ariel.
- Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., Pérez, J. (2023). *Spanish pre-trained BERT model and evaluation data*. arXiv:2308.02976. Retrieved from <https://arxiv.org/abs/2308.02976>
- Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J., Johnson, M. (2000). *BLLIP 1987–89 WSJ corpus release 1* (Tech. Rep.). Philadelphia, PA: Linguistic Data Consortium.
- Chowdhury, M.F.M., & Lavelli, A. (2011). Assessing the practical usability of an automatically annotated corpus. *Proceedings of the 5th Linguistic Annotation Workshop* (pp. 101–109).
- Cochran, W.G. (1977). *Sampling techniques*. John Wiley & Sons.
- Colombo, P., Peyrard, M., Noiry, N., West, R., Piantanida, P. (2023). The glass ceiling of automatic evaluation in natural language generation. *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)* (pp. 178–183).
- Ethayarajh, K., & Jurafsky, D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4846–4853).
- Ettinger, A., Hwang, J.D., Pyatkin, V., Bhagavatula, C., Choi, Y. (2023). *"You are an expert linguistic annotator": Limits of LLMs as analyzers of Abstract Meaning Representation*. arXiv:2310.17793. Retrieved from <https://arxiv.org/abs/2310.17793>
- Gasco, L., Nentidis, A., Krithara, A., Estrada-Zavala, D., Murasaki, R.T., Primo-Peña, E., ... Krallinger, M. (2021). Overview of BioASQ 2021-MESINESP track. evaluation of advance hierarchical classification techniques for scientific

- literature, patents and clinical trials. *Proceedings of the Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum* (Vol. 2936, pp. 165–187). CEUR-WS.org.
- Gómez-Adorno, H., Posadas-Durán, J.P., Enguix, G.B., Capetillo, C.P. (2021). Overview of FakeDeS at IberLEF 2021: Fake news detection in Spanish shared task. *Procesamiento del Lenguaje Natural*, 67, 223–231.
- Han, J., Pei, J., Tong, H. (2022). *Data mining: Concepts and techniques*. Morgan Kaufmann.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441, <https://doi.org/10.1037/h0071325>
- Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. Continuum.
- Jeronimo, C.L.M., Marinho, L.B., Campelo, C.E., Veloso, A., da Costa Melo, A.S. (2019). Fake news classification based on subjective language. *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services* (pp. 15–24).
- Kasmuri, E., & Basiron, H. (2017). Subjectivity analysis in opinion mining—a systematic literature review. *Int J Adv Soft Comput Appl*, 9(3), 132–159.
- Klie, J.-C. (2018). INCEpTION: Interactive machine-assisted annotation. *DESIREs*, 8, 5–9.
- Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3), 411–433, <https://doi.org/10.1111/j.1468-2958.2004.tb00738.x>
- Krüger, K.R., Lukowiak, A., Sonntag, J., Warzecha, S., Stede, M. (2017). Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5), 687–707, <https://doi.org/10.1017/S1351324917000043>
- Kruskal, W.H., & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621, <https://doi.org/10.1080/01621459.1952.10483441>

- Masip, J., Sporer, S.L., Garrido, E., Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime & Law*, 11(1), 99–122, <https://doi.org/10.1080/10683160410001726356>
- Meunier, A. (1974). Modalités et communication. *Langue française*(21), 8–25.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1–135, <https://doi.org/10.1561/15000000011>
- Pennebaker, J.W., Francis, M.E., Booth, R.J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Posadas-Durán, J.-P., Gómez-Adorno, H., Sidorov, G., Escobar, J.J.M. (2019). Detection of fake news in a new corpus for the Spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5), 4869–4876.
- Pérez Hernández, M.C. (2002). Explotación de los corpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. *Estudios de lingüística del español*, 18, 000–0.
- Saquete, E., Tomás, D., Moreda, P., Martínez-Barco, P., Palomar, M. (2020, March). Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications*, 141, 112943, <https://doi.org/10.1016/j.eswa.2019.112943> Retrieved 2024-04-26, from <https://linkinghub.elsevier.com/retrieve/pii/S095741741930661X>
- Shokri, M., Sharma, V., Filatova, E., Jain, S., Levitan, S. (2024). Subjectivity detection in English news using large language models. *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 215–226).
- Solà-Sales, S., Alzetta, C., Moret-Tatay, C., Dell’Orletta, F. (2023). Analysing deception in witness memory through linguistic styles in spontaneous language. *Brain Sciences*, 13(2), 317.
- Song, S., Hu, J., Mahowald, K. (2025). *Language models fail to introspect about their knowledge of language*. arXiv:2503.07513. Retrieved from <https://arxiv.org/abs/2503.07513>

- Spearman, C. (1961). *The proof and measurement of association between two things*. Appleton-Century-Crofts.
- Struß, J.M., Ruggeri, F., Barrón-Cedeño, A., Alam, F., Dimitrov, D., Galassi, A., ... Siegel, M. (2024). Overview of the CLEF-2024 CheckThat! lab task 2 on subjectivity in news articles. *CEUR Workshop Proceedings* (Vol. 3740, pp. 287–298).
- Taboada, M., Anthony, C., Voll, K.D. (2006). Methods for creating semantic orientation dictionaries. *LREC* (pp. 427–432).
- Tuchman, G., & Aladro Vico, E. (1998). La objetividad como ritual estratégico: un análisis de las nociones de objetividad de los periodistas. *CIC: Cuadernos de información y comunicación*, 4, 199–218.
- Vicente, M., Maestre, M.M., Lloret, E., Cueto, A.S. (2021). Leveraging machine learning to explain the nature of written genres. *IEEE Access*, 9, 24705–24726.
- Vieira, L.L., Jeronimo, C.L.M., Campelo, C.E.C., Marinho, L.B. (2020). Analysis of the subjectivity level in fake news fragments. *Proceedings of the Brazilian Symposium on Multimedia and the Web* (pp. 233–240). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/3428658.3430978>
- Volkova, S., Wilson, T., Yarowsky, D. (2013, October). Exploring demographic language variations to improve multilingual sentiment analysis in social media. D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1815–1827). Seattle, Washington, USA: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/D13-1187/>
- Wilson, T., Wiebe, J., Cardie, C. (2017). MPQA opinion corpus. *Handbook of Linguistic Annotation* (pp. 813–832). Springer.
- Wilson, T., Wiebe, J., Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 347–354).
- Zorraquino, M.A.M. (1999). Aspectos de la gramática y de la pragmática de las partículas de modalidad en español actual. *Español como lengua extranjera, enfoque comunicativo y gramática: actas del IX congreso internacional de ASELE. Santiago de Compostela, 23-26 de septiembre de 1998* (pp. 25–56).