MESironic: A Multimodal Dataset and Benchmark for Irony Detection in Spontaneous Spoken Spanish

MESironic: un corpus multimodal y referencia para la detección de ironía en el español hablado espontáneo

Vicent Ahuir,[1] **Alejandro Joaquín Barceló-Milkova,**[1] **Andreu Casamayor-Segarra,**[1] **Valentina González Verdegais,**[2] **Lluís-F. Hurtado,**[1 3] **Carlos Periñán-Pascual,**[2] **Maria Jose Castro-Bleda**[1 3]

[1]Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València

[2]Departamento de Lingüística Aplicada, Universitat Politècnica de València

[2]Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI)

vahuir@upv.es,ajbarmil@alumni.upv.es,{ancase3,lhurtado,jopepas3,mcastro}@upv.es

Abstract: We introduce MESironic, a multimodal corpus for irony detection in spontaneous Peninsular Spanish. The dataset comprises 6000 short audio clips with manually verified transcriptions, evenly split between ironic and non-ironic instances. Each sample contains the raw waveform and its text transcription. We describe the source material, the annotation protocol, and the main statistical properties of the collection. To provide a first benchmark, we evaluate a text-only baseline (TF-IDF+SVC) and a simple multimodal baseline that concatenates MFCC-derived acoustic features with the same textual representation. The multimodal system outperform the text-only model, confirming that acoustic and linguistic cues are complementary for irony detection. The dataset is publicly available to foster research in multimodal irony understanding in Spanish.

Keywords: irony detection, Spanish language resources, multimodal dataset, multimodal NLP

Resumen: Presentamos MESironic, un corpus multimodal para la detección de ironía en español peninsular espontáneo. El conjunto de datos está compuesto por 6000 fragmentos de audio breves con transcripciones verificadas manualmente, distribuidos de forma equilibrada entre ejemplos irónicos y no irónicos. Cada muestra contiene la señal de audio y su transcripción. Describimos el material de origen, el protocolo de anotación y las principales propiedades estadísticas de la colección. Para ofrecer una primera referencia de rendimiento, evaluamos un sistema base de solo texto (TF-IDF+SVC) y un sistema base multimodal que concatena características acústicas derivadas de MFCC con la misma representación textual. El sistema multimodal supera al modelo solo texto, confirmando que las señales acústicas y lingüísticas son complementarias para la detección de ironía. El conjunto de datos está disponible públicamente para fomentar la investigación en comprensión multimodal de la ironía en español.

Palabras clave: detección de ironía, recursos del idioma español, corpora multimodales, PLN multimodal

1 Introduction

Irony has long been a subject of inquiry across philosophy, linguistics, and communication studies. Its interpretation depends heavily on contextual knowledge, shared assumptions, and pragmatic inference, making it a particularly challenging phenomenon both for humans and for computational systems. If speakers themselves sometimes struggle to identify ironic intent, building systems that can recognize it requires a careful theoretical grounding as well as suitable empirical resources.

Within linguistic theory, one of the most influential accounts of irony is provided by Sperber and Wilson's Relevance Theory (Wilson y Sperber, 2012). Their framework posits that communication is guided by the search for optimal relevance: an utterance is relevant when the cognitive effects it pro-

duces justify the effort required to interpret it. Building on this perspective, Sperber and Wilson describe irony as a form of echoic mention in which speakers implicitly evoke a thought, expectation, or norm and convey a critical or mocking stance toward it (**?**). This view emphasizes that irony is not merely a stylistic device but a pragmatic mechanism grounded in shared knowledge and contextual inference. Such contextual dependence is further elaborated by Kreuz, who notes that ironic meaning often emerges through what he terms *juxtaposition*, namely the temporal or situational relationship between an utterance and the circumstances surrounding it (Kreuz, 2020). For instance, the utterance "What lovely weather!" acquires ironic meaning when produced during a storm but not when referring to a past sunny day. Likewise, irony presupposes a degree of shared common ground between interlocutors, which explains why speakers are more likely to employ it among familiar audiences than among strangers. These pragmatic conditions contribute to the ambiguity of ironic language and to the risk of misunderstanding.

Another well-documented characteristic of irony is its asymmetric polarity. Speakers frequently employ positively valenced expressions to refer to negative situations, a pattern that Sperber and Wilson relate to the ironic echoing of violated social expectations (Wilson y Sperber, 2012). Such mechanisms highlight the strong sociocultural component of irony, suggesting that its interpretation varies across communities and communicative settings.

Beyond contextual factors, irony often relies on extralinguistic cues. Prosody in particular plays a central role: speakers commonly associate irony with a distinctive delivery characterized by flatter intonation, slower tempo, or reduced pitch variability (Wilson y Sperber, 2012). Although no single acoustic pattern universally signals irony, prosodic cues frequently interact with lexical meaning to guide interpretation. This inherently multimodal nature of irony poses a challenge for computational modeling, as most existing resources focus exclusively on textual information.

Indeed, a preliminary survey of available corpora reveals that resources for irony detection in Spanish remain scarce and are overwhelmingly text-based. While datasets derived from social media or online discourse are useful for studying written irony, they do not capture the prosodic and interactional features of spoken communication. Existing multimodal corpora are either limited in size, restricted to scripted audiovisual material, or lack the balanced annotation required for robust machine learning applications.

Motivated by these limitations, we introduce *MESironic*, a multimodal dataset designed specifically for irony detection in spoken Peninsular Spanish. The corpus combines synchronized textual transcriptions with acoustic information extracted from spontaneous speech. In addition to presenting the dataset, we establish baseline results that validate the effectiveness of multimodal modeling for irony detection in Spanish. By making this resource publicly available, we aim to support future research on figurative language, multimodal semantics, and speech-based natural language understanding.

2 Related Work

Research on irony detection has historically focused on written language, particularly short texts from social media platforms such as Twitter (for example, TweetEval (Barbieri et al., 2020)). These resources have been instrumental in advancing computational approaches and have supported the development of both linguistic feature-based models and neural architectures. However, their reliance on textual input alone limits their ability to capture prosodic, acoustic, and visual cues, such as intonation, gesture, or facial expression, that often play a central role in ironic communication (Attardo, 2000).

This limitation is especially evident in Spanish. Existing corpora are predominantly derived from online discourse, including early work on irony and humor detection in written Spanish (Reyes, Rosso, y Buscaldi, 2012). Resources such as the Spanish Online Forums Corpus (SoFoCo) (Justo et al., 2018) have been useful for sentiment analysis in social networks, but they remain restricted to textual data and are primarily oriented toward detecting sarcasm and hostile language. Likewise, the multilingual perspectivist corpus MultiPICo (Casola et al., 2024), while innovative in modeling annotator disagreement and sociocultural variation, is sourced exclusively from written platforms and therefore does not include acoustic or interactional in-

formation.

Efforts to incorporate multimodality into irony detection have been more prominent in English-language research. Datasets such as MUStARD (Castro et al., 2019), which contains audiovisual excerpts from television series, have enabled the exploration of models that jointly process text, speech, and visual signals. Studies using such datasets consistently demonstrate that cross-modal pragmatic incongruities, such as a positive lexical statement delivered with a negative prosodic contour, constitute strong indicators of sarcastic intent (Zhuang et al., 2025; Wei et al., 2024). Despite these advances, comparable multimodal resources for Spanish remain scarce and fragmented (Malik, Tomás, y Rosso, 2023).

Among the few Spanish multimodal initiatives, the corpus compiled by Casa Gómez (de la Casa Gómez, 2021) provides recordings from television programs that include authentic spoken interactions. However, its 432 instances (all ironic) and the lack of non-ironic examples limit its suitability for training machine-learning models. Another relevant effort is the multimodal corpus "¡Qué maravilla!" (Hämäläinen, 2016), which includes both ironic and non-ironic examples aligned with audio and text. Nevertheless, this dataset relies on dubbed animated series and combines Peninsular and Latin-American Spanish varieties, resulting in speech that is partially scripted and produced in controlled recording environments. As Kreuz observes, such "posed" sarcastic speech may reflect speakers' beliefs about irony rather than its spontaneous use in natural interaction (Kreuz, 2020).

Taken together, previous work on Spanish irony reveals a persistent gap: Spanish irony datasets are overwhelmingly text-based, while existing multimodal resources are either limited in scale, derived from scripted media, or not representative of spontaneous spoken language. These limitations restrict the development of models capable of capturing the pragmatic and prosodic dimensions of irony in real communicative contexts.

MESironic directly addresses this gap by providing synchronized audio and text representations of spontaneous spoken Peninsular Spanish. By focusing on spontaneous interaction and balanced ironic/non-ironic annotation, the dataset enables systematic investigation of irony as a multimodal phenomenon and supports the development of models that go beyond text-only approaches, contributing to the growing research area of multimodal NLP for under-resourced languages.

3 The MESironic Dataset

This section introduces MESironic, a multimodal dataset designed for irony detection in spoken Peninsular Spanish. The dataset addresses the lack of publicly available resources that combine acoustic and textual information for the study of irony in natural spoken Spanish.

3.1 Data Collection

The dataset consists of short spoken utterances extracted from publicly available audiovisual content, primarily videos and podcasts published on platforms such as YouTube. In total, 123 hours and 16 minutes of audio from 87 recordings were reviewed to identify relevant segments. Additional material was incorporated from the Val.Es.Co. research group's colloquial speech corpus (**?**). Table 1 presents the distribution of samples by source type.

Sources were selected to prioritize spontaneous, non-formal speech and to capture diverse communicative contexts. Most samples originate from conversational settings and therefore include natural interactional phenomena such as hesitations, interruptions, and overlapping speech. To reduce excessive linguistic and cultural variability relative to dataset size, only Peninsular Spanish was included, since culture-dependent pragmatic factors play a central role in irony interpretation.

Several guidelines guided the extraction of samples. Initially, videos were selected based on their potential to contain ironic utterances, focusing on two types: (1) interactions among speakers who share common ground, which tends to facilitate irony production, and (2) informal debates, where irony may serve to mitigate face-threatening acts. Once videos were selected, the first annotator listened to the audio and extracted segments she identified as potentially ironic. The extracted units correspond to speech acts as defined by the Val.Es.Co. group: the minimal unity of action and intention, that possesses isolability and identifiability properties in a given context (**?**). Ironic samples are speech

Source Type	Samples	Avg. Duration (s)	Avg. Words
Stand-up comedy	36	2.47	8.97
Podcast (one speaker)	324	4.01	13.60
Podcast (multi-speaker)	5387	3.36	11.57
Val.Es.Co. corpus	24	2.05	6.83
Live stream	232	6.15	19.06
Total	6003	3.49	11.93

Table 1: Dataset statistics per source type.

acts produced with ironic intent, constituting minimal units that convey complete meaning. Smaller units (subacts) would lack complete meaning, potentially hindering irony identification, while larger units (interventions) would dilute the ironic effect. To maintain class balance, annotators extracted an equivalent number of non-ironic samples from each source, matched approximately for duration with the ironic samples. This procedure ensured that the final corpus maintains a balanced proportion of ironic and non-ironic instances. The final version of MESironic contains 6003 annotated samples.

3.2 Transcription and Alignment

Audio segments were automatically transcribed using the *Whisper* speech recognition system (Radford et al., 2022). The resulting transcriptions were manually reviewed and corrected by a graduate in Hispanic Studies to reduce recognition errors while preserving natural speech phenomena such as disfluencies and incomplete constructions.

Preprocessing included segmentation based on manually verified timestamps, normalization of transcription conventions, and alignment of acoustic and textual representations at the sample level. As a result, each instance in MESironic contains synchronized audio and text, enabling acoustic, textual, and multimodal modeling.

3.3 Annotation Framework

3.3.1 Definition of Irony

Irony was defined following the rhetorical account proposed by (Kostenko, 2014), according to which irony arises when the speaker's intended meaning contradicts the literal interpretation of the utterance. Such contradiction may be conveyed lexically, prosodically, or pragmatically, and may not be recoverable from text alone.

A binary labeling scheme was adopted:

- Irony: the speaker's intended meaning contradicts the literal content of the utterance. This contradiction may be expressed through wording, prosody, or contextual cues.
- Non-irony: the utterance does not exhibit such contradiction.

3.3.2 Annotator Involvement

Each sample was independently labeled by three annotators. The annotation team consisted of one linguist, one computer scientist, and a rotating pool of ten university-educated native Spanish speakers, who collectively provided the third annotation for each sample.

Annotators followed shared written guidelines and had access to surrounding audiovisual context when making decisions. They were instructed to focus on the speaker's communicative intention, compare it with the literal meaning, and consider prosodic and contextual cues when available. Hyperbolic or exaggerated expressions were not labeled as ironic unless they conveyed a meaning opposite to the literal interpretation.

3.3.3 Annotation Interface and Workflow

To support consistent and context-aware annotation, a custom graphical tool was developed in Python using the Qt framework. The interface allowed annotators to navigate the original audiovisual source, inspect surrounding context, and replay segments before assigning labels.

The overall annotation workflow is summarized in Figure 1. Spoken segments are first extracted from audiovisual sources and automatically transcribed, followed by manual correction. Annotators then label each segment using the custom interface, which provides access to audiovisual context and playback controls. Each instance is independently annotated by three annotators following shared guidelines.

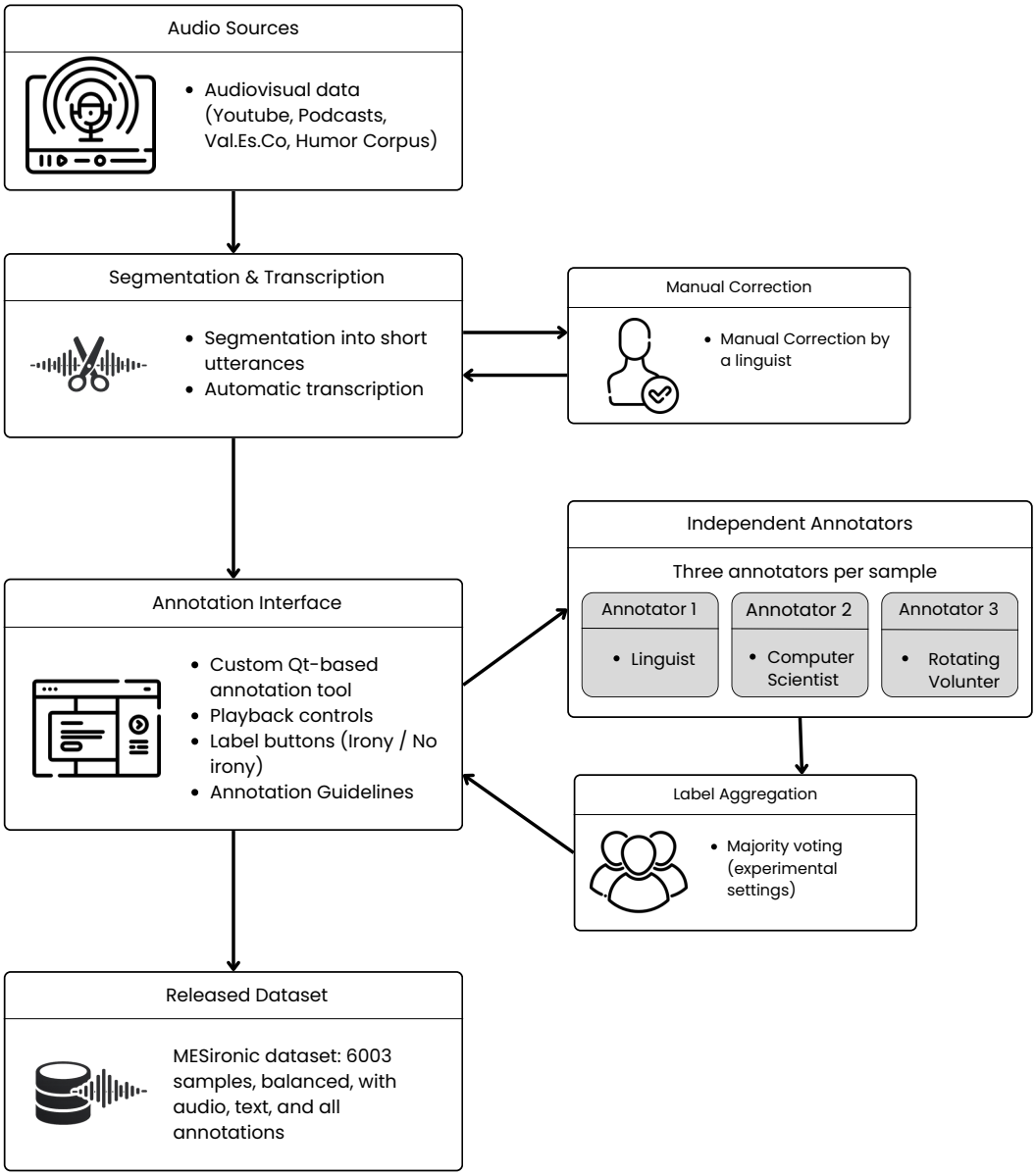

Figure 1: Annotation workflow for the MESironic dataset. Spoken segments are transcribed and manually corrected, then labeled by three annotators using a custom interface with audiovisual context. Final labels are set by majority vote, with all individual annotations retained.

Given the inherently subjective nature of irony perception, all individual annotations are retained in the dataset. For the experiments reported in this paper, labels are derived through majority voting.

3.4 Dataset Statistics and Agreement

The final dataset contains 6003 instances. It is well balanced, with 3045 ironic (50.72%) and 2958 non-ironic (49.28%) samples. The average length is 11.9 words per sample and 3.6 seconds per utterance (see Table 1).

Inter-annotator agreement was measured using Fleiss' Kappa, yielding a value of 0.65, indicating substantial agreement. Despite the inherently subjective nature of irony perception, this value reflects the overall consistency of the annotation process and the clarity of the annotation guidelines.

3.5 Train/Test Split

We randomly split the 6003 samples into a training set of 5100 instances (2587 ironic, 2513 non-ironic) and a test set of 900 instances (458 ironic, 445 non-ironic) using stratified sampling to preserve class balance. The test set is held out and used only for final evaluation.

3.6 Examples

We present representative ironic samples from the MESironic dataset, illustrating the variety of ironic expressions captured:

- Fue Víctor. Haciendo amigos, día uno. (*It was Victor. Making friends, day one*).
  → The speaker attributes a conflict to "Victor" ironically, implying the opposite of making friends.
- Puede ser que los de Pringles ya tuvieran la primera impresora 3D en el 73. (*It's possible that Pringles already had the first 3D printer in '73*).
  → Absurd anachronism mocking the tendency to attribute modern inventions to the past.
- A ver, alguno se pondrá contento porque es una Xbox. (*Well, someone will be happy because it's an Xbox*).
  → Said dismissively about a prize, implying it is not actually desirable.

3.7 Availability and Ethical Considerations

MESironic is derived exclusively from publicly available material and includes only short excerpts necessary for research purposes. No personally identifiable information is included, and speaker identities are not disclosed.

The dataset is released for research use under an open license, together with metadata, annotations, and documentation describing the annotation process and intended use.Links to the original sources are provided to ensure transparency and traceability.

4 Baseline Systems

To establish strong reference points for the MESironic dataset, we implemented a set of baseline models leveraging both textual and acoustic modalities. On the textual side, utterances were represented using TF-IDF features (up to 10,000 dimensions) without stopword removal, followed by MinMax scaling, and classified using a Support Vector Machine (SVC). For the multimodal baseline, we incorporated prosodic information by extracting Mel-Frequency Cepstral Coefficients (MFCCs) from the audio signals, which were subsequently aggregated via mean pooling. These acoustic descriptors were concatenated with the textual features to form a joint representation. A Support Vector Classifier (SVC) with an RBF kernel was then trained, with hyperparameters optimized through grid search and class balancing enabled. These baselines provide a simple yet effective framework to benchmark irony detection on MESironic, demonstrating the contribution of combining lexical and acoustic cues while maintaining a relatively low computational cost.

System	Precision	Recall	F1
ELiRF-UPV	0.60177	0.60153	0.60025

Table 2: Test performance of the ELiRF-UPV baseline on the text-based modality of the MESironic dataset.

System	Precision	Recall	F1
ELiRF-UPV	0.60984	0.60926	0.60865

Table 3: Test performance of the ELiRF-UPV baseline on the multimodal setting of the MESironic dataset.

As shown in Table 2 and 3, the ELiRF-UPV baseline obtains an F1-score of 0.60025 text-based modality, with precision and recall values of 0.60177 and 0.60153, respectively. In multimodal, the system reaches an F1-score of 0.60865, with a precision of 0.60984 and a recall of 0.60926. These results reflect a difference between the text-only and multimodal settings, with higher values observed when acoustic features are incorporated.

5 Conclusion and Future Work

This work introduced MESironic, a new multimodal corpus for irony detection in spoken Peninsular Spanish, and presented a systematic evaluation of unimodal and multimodal modeling strategies on this dataset. The experiments reveal a consistent performance hierarchy across modalities: text-only systems exhibit the lowest stability and highest variance, audio-only systems provide substan-

tial improvements by capturing prosodic and paralinguistic cues, and multimodal systems achieve the highest and most robust performance overall. These findings confirm that irony in spoken discourse arises from the interaction between semantic content and vocal realization, and that reliable detection therefore requires integrating both modalities rather than relying on either in isolation.

Beyond the immediate empirical findings, the results also provide evidence for the linguistic importance of prosody in pragmatic interpretation. Acoustic features alone already offer strong predictive power, underscoring their role in signaling speaker intention. However, combining them with textual representations consistently yields more robust predictions, supporting the view that irony detection (and more broadly, spoken language understanding) requires integrating lexical meaning with paralinguistic expression. In this sense, this study contributes not only a new dataset but also a benchmark for advancing research on multimodal pragmatics in Spanish.

Limitations. Despite these contributions, several limitations remain. First, the dataset is restricted to utterance-level annotations, a design choice that enables controlled experimentation but limits the ability of models to capture broader discourse context or speaker-specific patterns that may be essential for interpreting irony in natural conversation. Second, although the corpus incorporates both text and audio, it does not yet include richer conversational metadata such as interaction structure, speaker relationships, or audience response, all of which may influence ironic interpretation. Third, the models explored in this work rely on relatively simple fusion strategies; while this choice was deliberate to establish clear baselines, more sophisticated multimodal architectures may yield further improvements. These limitations suggest that the results presented here should be interpreted as a strong baseline rather than an upper bound.

Future Work. Building on these findings, future work will explore several directions. First, we plan to investigate more advanced fusion mechanisms, including cross-modal attention and transformer-based architectures that can model interactions between modalities at finer temporal granularities. Second, we aim to incorporate contextual information beyond the utterance level, such as discourse structure or conversational history, to address the context-dependent errors identified in our analysis. Third, we will extend the dataset with additional annotations—for example, speaker demographics, dialogue acts, or audience reactions—to enable richer pragmatic modeling. Finally, we plan to investigate cross-domain and cross-language generalization, building on recent advances in spoken satire detection (Pan et al., 2025; Barceló-Milkova1 et al., 2025; Ahuir et al., 2026). By releasing MESironic as a public resource, we hope to support further research on prosody-aware language understanding and pragmatic modeling in spoken communication, and to invite the community to build upon these baselines with more sophisticated approaches.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033 and ERDF "ERDF A way of making Europe" (project PID2024-155948OB-C55), by MICIU/AEI/10.13039/501100011033 and ERDF, EU (R&D&I project PID2023-147137NB-I00), and by the Spanish Ministerio de Universidades (grant FPU24/01938).

Bibliografía

Ahuir, V., A. J. Barceló-Milkova, A. Casamayor-Segarra, L.-F. Hurtado, y M. J. Castro-Bleda. 2026. Listen, Read, and Detect: Multimodal Satire Classification in Spanish Language Media. *Computer Speech and Language*, In press.

Attardo, S. 2000. Irony as relevant inappropriateness. *Journal of Pragmatics*, 32(6):793–826.

Barbieri, F., J. Camacho-Collados, L. Espinosa Anke, y L. Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. En T. Cohn Y. He, y Y. Liu, editores, *Findings of the Association for Computational Linguistics: EMNLP 2020*, páginas 1644–1650, Online, Noviembre. Association for Computational Linguistics.

Barceló-Milkova1, A. J., A. Casamayor-Segarra, V. Ahuir, y M. J. Castro-Bleda. 2025. ELiRF-UPV at SatiSPeech-IberLEF 2025: Multimodal Speech-text

Satire Recognition in Spanish. *Procesamiento del Lenguaje Natural*, 75(0).

Casola, S., S. Frenda, S. M. Lo, E. Sezerer, A. Uva, V. Basile, C. Bosco, A. Pedrani, C. Rubagotti, V. Patti, y D. Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. En L.-W. Ku A. Martins, y V. Srikumar, editores, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 16008–16021, Bangkok, Thailand, Agosto. Association for Computational Linguistics.

Castro, S., D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, y S. Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). En A. Korhonen D. Traum, y L. Màrquez, editores, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 4619–4629, Florence, Italy, Julio. Association for Computational Linguistics.

de la Casa Gómez, L. 2021. *La ironía verbal: caracterización pragmática a partir de un corpus oral en español*. Ph.D. tesis, Jaén : Universidad de Jaén, Noviembre.

Hämäläinen, M. 2016. *Reconocimiento automático del sarcasmo: ¡Esto va a funcionar bien!* Ph.D. tesis.

Justo, R., J. M. Alcaide, M. I. Torres, y M. Walker. 2018. Detection of Sarcasm and Nastiness: New Resources for Spanish Language. *Cognitive Computation*, 10(6):1135–1151, Diciembre.

Kostenko, A. P. 2014. The concept of irony in the study of literature. *Science and Education a New Dimension. Philology, II (4)*, (24):9–13.

Kreuz, R. 2020. *Irony and sarcasm*. The MIT Press, Massachusetts.

Malik, M., D. Tomás, y P. Rosso. 2023. How challenging is multimodal irony detection? En E. Métais F. Meziane V. Sugumaran W. Manning, y S. Reiff-Marganiec, editores, *Natural Language Processing and Information Systems*, páginas 18–32, Cham. Springer Nature Switzerland.

Pan, R., J. A. García-Díaz, T. Bernal-Beltrán, F. García-Sánchez, y R. Valencia-García. 2025. Overview of SatiSPeech at IberLEF 2025: Multimodal Audio-Text Satire Classification in Spanish. *Procesamiento del Lenguaje Natural*, 75:513–522.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, y I. Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Reyes, A., P. Rosso, y D. Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Wei, Y., M. Duan, H. Zhou, Z. Jia, Z. Gao, y L. Wang. 2024. Towards multimodal sarcasm detection via label-aware graph contrastive learning with back-translation augmentation. *Knowledge-Based Systems*, 300:112109.

Wilson, D. y D. Sperber. 2012. *Meaning and Relevance*. Cambridge University Press, New York, UEA.

Zhuang, X., Z. Li, F. Zhou, J. Gu, C. Zhang, y H. Ma. 2025. Dycr-net: A dynamic context-aware routing network for multimodal sarcasm detection in conversation. *Knowledge-Based Systems*, 310:113029.