Procesamiento del Lenguaje Natural, Revista nº 77, septiembre de 2026, pp 689–697 recibido 01-07-2026 revisado 17-07-2026 aceptado 28-07-2026

Overview of SpeechMATICS at IberLEF 2026: Speech Multimodal Audio-Text Irony Classification in Spanish

Resumen de SpeechMATICS en IberLEF 2026: Clasificación Multimodal Audio-Texto de Ironía en Español

Vicent Ahuir-Esteve,[1] **Alejandro Joaquín Barceló-Milkova,**[1] **Andreu Casamayor-Segarra,**[1] **Lluís-F. Hurtado-Oliver,**[1,2] **Maria Jose Castro-Bleda**[1,2]

[1]Valencia Research Institute for Artificial Intelligence, Universitat Politècnica de València
[2]Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI), Universitat Politècnica de València
{vahuir, ancase3, mcastro, lhurtado}@upv.es

Abstract: This paper presents an overview of the SpeechMATICS 2026 shared task, organized as part of IberLEF 2026 (Bonet-Jover, González-Barba, and Chiruzzo, 2026). The main goal of the task is to advance research on automatic irony detection in spontaneous spoken Spanish from both textual and multimodal perspectives. The task is divided into two independent subtasks. The first subtask focuses on irony detection using only manual transcriptions of audio fragments. The second subtask introduces a multimodal setting in which systems can combine textual and acoustic information. SpeechMATICS is based on MESironic, a multimodal corpus of short spoken fragments with manually verified transcriptions and irony annotations. A total of 14 teams registered for the task, 4 submitted official runs, and 4 submitted working notes describing their systems. This overview describes the task, dataset, evaluation protocol, participant approaches, official results, and main findings from this first edition.
Keywords: irony detection, multimodal classification, spoken Spanish, speech processing.

Resumen: Este artículo presenta una visión general de la tarea compartida SpeechMATICS 2026, organizada en el marco de IberLEF 2026 (Bonet-Jover, González-Barba, and Chiruzzo, 2026). El objetivo principal de la tarea es avanzar en la investigación sobre la detección automática de la ironía en español hablado espontáneo desde una perspectiva textual y multimodal. La tarea se divide en dos subtareas independientes. La primera subtarea se centra en la detección de ironía utilizando únicamente las transcripciones manuales de los fragmentos de audio. La segunda subtarea introduce un escenario multimodal en el que los sistemas pueden combinar información textual y acústica. SpeechMATICS se basa en MESironic, un corpus multimodal de fragmentos breves de habla con transcripciones verificadas manualmente y anotaciones de ironía. Un total de 14 equipos se registraron en la tarea; 4 enviaron predicciones oficiales y 4 presentaron working notes que describían sus sistemas. Este overview describe la tarea, el conjunto de datos, el protocolo de evaluación, los enfoques de los participantes, los resultados oficiales y las principales conclusiones de esta primera edición.
Palabras clave: detección de ironía, clasificación multimodal, español hablado, procesamiento del habla.

ISSN 1135-5948 DOI 10.26342-2026-77-49 © 2026 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introduction

Irony is a complex pragmatic phenomenon in which the speaker's intended meaning differs from, and often contradicts, the literal content of an utterance. Its interpretation depends on contextual knowledge, shared assumptions, communicative intention, and pragmatic inference. From the perspective of Relevance Theory, irony can be understood as a form of echoic mention in which speakers evoke a thought, expectation, or norm while expressing a critical or mocking attitude towards it (Sperber and Wilson, 1981; Wilson and Sperber, 2012). This contextual dependence makes irony difficult to interpret even for humans, and particularly challenging for automatic language understanding systems.

In spoken interaction, irony is not expressed only through lexical content. Speakers may convey ironic intent through intonation, rhythm, emphasis, pauses, pitch variation, voice quality, or other paralinguistic signals. As noted in previous pragmatic studies, ironic meaning often emerges from the relationship between an utterance and the surrounding situation (Kreuz, 2020), and prosody may guide interpretation when the literal meaning alone is insufficient (Attardo, 2000; Wilson and Sperber, 2012). Therefore, spoken irony is inherently multimodal: the same transcription may be interpreted differently depending on how it is pronounced and on the communicative context in which it appears.

Despite the importance of these acoustic and contextual cues, most computational work on irony detection has focused on written language, especially short texts from social media platforms. In Spanish, early research on irony and humor recognition has also been mainly text-based (Reyes, Rosso, and Buscaldi, 2012). Although these resources have been useful for studying written figurative language, they do not capture prosodic or interactional information. Multimodal irony and sarcasm detection has received more attention in English, with datasets such as MUStARD (Castro et al., 2019), but comparable resources for Spanish remain scarce and fragmented (Malik, Tomás, and Rosso, 2023). Recent initiatives such as SatiSPeech have demonstrated the community's interest in multimodal audio-text classification for figurative language in Spanish (Pan et al., 2025), but irony detection in spontaneous spoken Spanish remains underexplored.

The SpeechMATICS 2026 shared task, organized as part of IberLEF 2026 (Bonet-Jover, González-Barba, and Chiruzzo, 2026), addresses this gap by focusing on the automatic detection of irony in spontaneous spoken Spanish. The task is based on MESironic (Ahuir et al., 2026), a multimodal corpus composed of short spoken fragments, manually verified transcriptions, and irony annotations. To analyze the contribution of each modality, SpeechMATICS is divided into two independent subtasks. Task 1 focuses on text-only irony classification, where systems must rely exclusively on the transcription. Task 2 introduces a multimodal setting in which systems may combine textual and acoustic information to determine whether a fragment is ironic or non-ironic.

A total of 14 teams registered for SpeechMATICS 2026. Among them, 4 teams submitted official runs and 4 teams submitted working notes describing their systems. The participating teams explored a range of approaches, including fine-tuned Transformer-based language models, self-supervised speech encoders, acoustic features, large language models, and early- and late-fusion strategies.

The remainder of this paper is organized as follows. Section 2 describes the two subtasks and the evaluation protocol. Section 3 presents the MESironic dataset used in the competition. Section 4 describes the official baselines. Section 5 summarizes the approaches proposed by the participating teams. Section 6 reports and discusses the official results. Finally, Section 7 presents the main conclusions of this first edition.

2 Task description

Detecting irony in spontaneous spoken language presents several challenges. Ironic meaning may arise from the lexical content of an utterance, but it may also depend on how the utterance is pronounced, the communicative situation, and the relationship between the literal and intended meanings. Combining textual and acoustic information introduces additional challenges, since each modality conveys different and sometimes complementary cues. SpeechMATICS was designed to evaluate these two perspectives through two independent subtasks:

- **Task 1: Text-only Irony Classification.** The goal of this subtask is to determine whether a spoken fragment is ironic or non-ironic using only its textual transcription. Participants were not allowed to use the audio signal in this subtask. This setting evaluates the extent to which irony can be detected from lexical, syntactic, semantic, and discourse-level information alone.
- **Task 2: Multimodal Audio-Text Irony Classification.** This subtask introduces a multimodal setting in which participants could use both the textual transcription and the corresponding audio fragment. The objective is again to classify each instance as ironic or non-ironic, but systems may exploit acoustic, prosodic, and paralinguistic information in addition to the textual content. This setting was intended to assess whether speech information provides complementary evidence for irony detection in spontaneous spoken Spanish.

The competition was hosted on the Codabench platform.[1] It was structured into two main phases: Practice and Evaluation. During the Practice phase, participants were given access to a subset of the training data in order to become familiar with the dataset format, the two subtasks, and the submission requirements. Baseline material was also provided to help participants develop initial systems for both the text-only and multimodal settings. During the Evaluation phase, participants submitted their predictions for the official test set. A total of 100 submissions were allowed during the competition, and the final ranking for each subtask was computed using the official test labels.

Participants could submit predictions for one or both subtasks. Each submission file was required to contain the identifier of each instance and the predicted labels for the corresponding subtasks. Systems were evaluated using macro-averaged precision, macro-averaged recall, and macro-averaged F1-score. The official ranking metric for both subtasks was macro-F1, since it assigns equal importance to both classes and is therefore appropriate for evaluating performance on both ironic and non-ironic fragments.

[1] https://www.codabench.org/competitions/13700/

Following the official evaluation methodology, the reported scores were computed using a bootstrap-based procedure. Specifically, 1000 bootstrap samples were generated from the official test set by sampling instances with replacement. For each resampled test set, macro-precision, macro-recall, and macro-F1 were computed independently, and the final scores correspond to the mean performance across the 1000 bootstrap samples. This procedure provides more robust estimates of system performance and allows confidence intervals to be obtained for each participating system.

3 Dataset

SpeechMATICS is based on MESironic (Ahuir et al., 2026), a multimodal corpus designed for irony detection in spontaneous spoken Peninsular Spanish. Each instance in the dataset consists of a short audio fragment, its manually verified transcription, and irony annotations. The dataset was created to address the lack of resources that combine acoustic and textual information for the study of irony in natural spoken Spanish.

The corpus was built from spontaneous and non-formal spoken material extracted from publicly available audiovisual sources. In total, 123 hours and 16 minutes of audio from 87 recordings were reviewed to identify relevant segments. The selected material mainly comes from podcasts, live streams, stand-up comedy fragments, and the Val.Es.Co. corpus. These sources were chosen to prioritize natural interactional phenomena such as hesitations, interruptions, overlaps, and informal conversational dynamics. To reduce excessive linguistic and cultural variability, the corpus focuses on Peninsular Spanish.

The final version of MESironic contains 6003 samples. The dataset is almost perfectly balanced, with 3045 ironic instances and 2958 non-ironic instances. The average duration of the audio fragments is 3.49 seconds, and the average transcription length is 11.93 words. Table 1 shows the train/test split used in SpeechMATICS.

Audio segments were automatically transcribed using Whisper and then manually reviewed and corrected by a graduate in Hispanic Studies. This manual correction step was intended to reduce automatic speech

Label	Train	Test	Total
Ironic	2587	458	3045
Non-ironic	2513	445	2958
Total	5100	903	6003

Table 1: Distribution of MESironic samples used in SpeechMATICS.

recognition errors while preserving relevant phenomena of spontaneous speech, such as disfluencies and incomplete constructions. Each final instance, therefore, contains synchronized audio and text, enabling text-only, audio-based, and multimodal modeling.

Irony was annotated following a rhetorical definition according to which an utterance is ironic when the speaker's intended meaning contradicts the literal interpretation of the words. This contradiction may be expressed through lexical content, prosody, or contextual cues. Each sample was independently annotated by three annotators: a linguist specialized in colloquial Spanish, a computer scientist, and one annotator from a rotating pool of university-educated native Spanish speakers. Annotators had access to the surrounding audiovisual context and were instructed to focus on the speaker's communicative intention. The final label used in SpeechMATICS was obtained by majority vote. The annotation process achieved a Fleiss' Kappa of 0.65, indicating substantial agreement despite the subjective nature of irony perception.

4 Baselines

To provide participants with reference systems and examples of the expected data format, we released baseline material for both subtasks. The baselines were intentionally simple and computationally lightweight, following the initial benchmark proposed for MESironic. They were designed to offer a reproducible starting point rather than highly optimized systems.

For Task 1, the official baseline represented each transcription using TF-IDF features (Manning, Raghavan, and Schütze, 2008) and trained a Support Vector Classifier (SVC) (Cortes and Vapnik, 1995) for binary irony classification. This baseline used only textual information and therefore adhered to the restrictions of the text-only setting.

For Task 2, the official baseline extended the textual representation with acoustic information extracted from the audio signal. In particular, Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980) were extracted from each audio fragment and concatenated with the TF-IDF textual features. The resulting multimodal representation was then classified using an SVC (Cortes and Vapnik, 1995). This baseline was intended to provide a simple early-fusion approach combining lexical and acoustic cues.

5 Participant approaches

A total of 14 teams registered for SpeechMATICS 2026. Among them, 4 teams submitted official runs and 4 teams submitted working notes describing their systems. This section summarizes the main approaches proposed by the participating teams.

- **carlosdf.** This team participated in both subtasks with a strong Transformer-based system. For Task 1, the team fine-tuned `pysentimiento/robertuito-base-uncased` (Perez, Giudici, and Luque, 2021), using layer-wise learning rates, a freeze-then-unfreeze strategy for the encoder, class-weighted cross-entropy with label smoothing, and decision threshold tuning on the validation set. For Task 2, the team proposed *FusionNet*, a multimodal architecture combining three information streams: text embeddings extracted from the fine-tuned Task 1 encoder, audio embeddings obtained from Wav2Vec2-XLSR Spanish[2], based on Wav2Vec2/XLSR (Baevski et al., 2020; Conneau et al., 2021), and hand-crafted acoustic features including MFCCs (Davis and Mermelstein, 1980), pitch-related features, and spectral descriptors. These streams were projected and fused through a two-block MLP with a residual connection. The final submission combined the outputs of Task 1 and Task 2 using logistic stacking. The system ranked 1st in Task 1 and 2nd in Task 2.
- **migueluc3m.** This team also addressed both subtasks using Transformer-based text and speech representations. For

[2] `jonatasgrosman/wav2vec2-large-xlsr-53-spanish`

Task 1, three Spanish-oriented language models were fine-tuned and compared: BETO (Ca nete et al., 2020), BERTIN (de la Rosa and others, 2022), and MrBERT-es (Pe naranda and others, 2023). BETO was selected as the best-performing text model. For Task 2, the team evaluated three self-supervised audio encoders, Wav2Vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022), and explored three multimodal fusion strategies: a weighted ensemble, a Support Vector Machine (Cortes and Vapnik, 1995), and a Multilayer Perceptron. The best configuration combined BETO text logits and WavLM audio embeddings using an SVM classifier, which was retrained on the full training set for the final submission. The system ranked 2nd in Task 1 and 1st in Task 2.

- **UC-UCO-Plenitas.** This team proposed a configurable pipeline combining several complementary text-based components and an acoustic branch for the multimodal task. The text-only system integrated TF-IDF representations (Manning, Raghavan, and Schütze, 2008), multilingual sentence embeddings based on models such as multilingual E5 (Wang et al., 2024) with a class-balanced linear classifier, and a few-shot large language model used as a training-free classifier. For Task 2, the team incorporated acoustic information through MFCC features (Davis and Mermelstein, 1980). The predictions of the different components were combined using majority voting. The system achieved competitive results in the text-only setting, ranking 3rd in Task 1, but adding MFCC-based acoustic features did not improve the final Task 2 score, where the team ranked 4th.

- **HammerTeam.** This team explored pre-trained Transformer embeddings and fusion strategies for both subtasks. Textual representations were extracted using `xlm-roberta-base` (Conneau et al., 2020) with masked mean pooling, while acoustic representations were obtained from `wav2vec2-large-xlsr-53` (Baevski et al., 2020; Conneau et al., 2021) using temporal mean and standard deviation pooling. For Task 1, the team trained classifiers over the textual embeddings. For Task 2, two main strategies were explored. The first system used early fusion by concatenating textual and acoustic embeddings and training an MLP classifier. The second system used L2-normalized representations, audio truncation, and late fusion, combining the probabilities of independent text and audio classifiers through weighted averaging. HammerTeam ranked 4th in Task 1 and 3rd in Task 2.

6 Results and discussion

The official results for Task 1 and Task 2 are shown in Tables 2 and 3, respectively. Although macro-precision and macro-recall were also computed, the tables report only macro-F1, the official ranking metric, for readability. As described in Section 2, the reported values correspond to the mean macro-F1 across 1000 bootstrap samples of the official test set. Superscript and subscript values indicate the upper and lower bounds of the 95% bootstrap confidence interval.

6.1 Task 1 results

Task 1 evaluated irony detection using only the textual transcription of each audio fragment. As shown in Table 2, all participating systems outperformed the official baseline. The best average result was obtained by *carlosdf*, with a macro-F1 score of 0.659.

Rank	Team	Macro-F1
1	carlosdf	$\mathbf{0{,}659}^{0.690}_{0.629}$
2	migueluc3m	$0{,}658^{0.688}_{0.627}$
3	UC-UCO	$0{,}641^{0.672}_{0.609}$
4	Hammer	$0{,}622^{0.653}_{0.591}$
5	Baseline	$0{,}600^{0.631}_{0.569}$

Table 2: Official results for Task 1. Systems are ranked by macro-F1. Superscript and subscript values indicate the upper and lower bounds of the 95% bootstrap confidence interval.

The results show that the text-only setting was highly competitive. The two best systems obtained almost identical average

macro-F1 scores, and their confidence intervals strongly overlap. Therefore, although *carlosdf* achieved the highest mean score, the difference with *migueluc3m* should not be interpreted as a clear performance gap. This suggests that both systems performed comparably under the bootstrap evaluation.

The remaining systems also outperformed the baseline, although with lower average scores. The overlap between several confidence intervals indicates that differences between adjacent systems should be interpreted with caution. Nevertheless, the overall ranking shows that Transformer-based text representations provided useful improvements over the reference text-only baseline. At the same time, the moderate absolute scores confirm the difficulty of detecting irony from transcriptions alone, since ironic meaning may depend on prosody, speaker attitude, or contextual cues that are not fully represented in text.

6.2 Task 2 results

Task 2 evaluated multimodal irony detection, allowing systems to use both the transcription and the corresponding audio signal. As shown in Table 3, the best average result was obtained by *migueluc3m*, with a macro-F1 score of 0.708.

Rank	Team	Macro-F1
1	migueluc3m	$\mathbf{0.708}^{0.738}_{0.677}$
2	carlosdf	$0.697^{0.727}_{0.666}$
3	Hammer	$0.668^{0.699}_{0.637}$
4	UC-UCO	$0.641^{0.672}_{0.608}$
5	Baseline	$0.609^{0.641}_{0.577}$

Table 3: Official results for Task 2. Systems are ranked by macro-F1. Superscript and subscript values indicate the upper and lower bounds of the 95% bootstrap confidence interval.

Compared with Task 1, the best average macro-F1 increased from 0.659 to 0.708. This improvement suggests that acoustic information provides complementary cues for irony detection in spontaneous spoken Spanish. The best Task 2 system combined BETO-based text logits with WavLM audio embeddings through an SVM classifier, while the second-best system combined Transformer-based text representations, Wav2Vec2 audio embeddings, hand-crafted acoustic descriptors, and logistic stacking. These results indicate that speech representations can improve performance when they are integrated through effective fusion strategies.

However, the confidence intervals of the top systems overlap, especially between *migueluc3m* and *carlosdf*. Therefore, *migueluc3m* can be considered the best system in terms of average macro-F1, but the difference with the second-ranked system should be interpreted cautiously.

The benefit of the audio modality was not uniform across all teams. While *migueluc3m*, *carlosdf*, and *HammerTeam* improved their Task 2 scores with respect to Task 1, *UC-UCO-Plenitas* obtained the same average macro-F1 in both settings. This suggests that simply adding acoustic features is not always sufficient. The choice of audio representation, the fusion mechanism, and the balance between text and audio components appear to be key factors for successfully exploiting multimodal information.

6.3 Main findings

Several conclusions can be drawn from the official results and the participants' working notes. First, textual information is a strong signal for irony detection, as shown by the competitive results obtained in Task 1. Fine-tuned Transformer-based language models and sentence-level representations were commonly used by the best-performing systems.

Second, the multimodal setting achieved the best overall average results. The improvement from 0.659 macro-F1 in Task 1 to 0.708 in Task 2 suggests that acoustic and prosodic cues can complement lexical information. This is consistent with the nature of spoken irony, where the intended meaning may depend on how an utterance is delivered.

Third, not all multimodal approaches were equally effective. Systems based on richer speech representations such as WavLM or Wav2Vec2, combined with supervised fusion strategies, obtained the strongest results. In contrast, simpler acoustic descriptors or basic voting mechanisms did not always improve over text-only systems. This highlights the importance of designing fusion strategies that can exploit the interaction between modalities rather than merely appending acoustic features to textual representa-

tions.

Finally, the bootstrap confidence intervals show that some differences between nearby systems are small and should be interpreted carefully. This is especially clear in Task 1, where the top two systems obtained nearly identical scores. For this reason, the ranking should be read together with the confidence intervals, distinguishing between the best average score and statistically clear separation between systems.

6.4 Post-evaluation phase and future directions

The shared task remains accessible in the post-evaluation phase through the Codabench platform.[3] The competition page includes the resources required to reproduce the baseline systems and to prepare valid submission files, together with the dataset files made available for the task, excluding the gold labels of the official test set. We hope that these resources will be useful for the Spanish NLP and speech processing communities, and that they will encourage further research on multimodal irony detection in spoken Spanish.

The results of this first edition show that there is still considerable room for improvement. Although all participating systems outperformed the official baselines, the absolute scores remain moderate, especially in the text-only setting. This confirms the difficulty of detecting irony from short spontaneous spoken fragments, where the intended meaning may depend on subtle lexical choices, prosody, speaker attitude, and contextual information. The multimodal results suggest that acoustic information can improve performance, but they also show that simply adding audio features is not sufficient. Future work should therefore explore more robust fusion strategies, such as attention-based mechanisms, gated fusion, tensor fusion, or contrastive multimodal learning, in order to better model the interaction between textual and acoustic cues.

Another important challenge is the cultural and pragmatic nature of irony. In Spanish, ironic intent may vary depending on regional varieties, conversational norms, discourse context, and shared sociocultural knowledge. This makes generalization difficult, particularly when systems are evaluated on previously unseen speakers, topics, and communicative situations. For future editions of SpeechMATICS, we plan to expand the dataset with additional examples and a broader range of spoken sources, while maintaining the focus on high-quality transcriptions and carefully validated annotations. Our goal is to continue promoting the development of robust systems for irony detection in spontaneous spoken Spanish.

[3] https://www.codabench.org/competitions/13700/

7 Conclusions

This paper has presented the overview of SpeechMATICS 2026, a shared task organized at IberLEF 2026 on automatic irony detection in spontaneous spoken Spanish. The task was based on MESironic, a multimodal corpus composed of short audio fragments, manually verified transcriptions, and irony annotations. SpeechMATICS was divided into two subtasks: a text-only setting, where systems classified irony using only transcriptions, and a multimodal setting, where participants could combine textual and acoustic information.

A total of 14 teams registered for the task, 4 teams submitted official runs, and 4 teams submitted working notes describing their approaches. The participating systems explored a variety of methods, including fine-tuned Transformer-based language models, self-supervised speech encoders, acoustic descriptors, large language models, and different early and late fusion strategies. The results show that all submitted systems outperformed the official baselines in both subtasks.

The best text-only system achieved a macro-F1 score of 0.659, while the best multimodal system reached 0.708 macro-F1. This improvement suggests that acoustic information can provide useful complementary cues for detecting irony in spoken Spanish. However, the results also show that the benefit of multimodality depends strongly on the quality of the audio representation and the fusion strategy. Systems using richer speech representations and supervised fusion mechanisms obtained the strongest multimodal results, whereas simpler acoustic features did not always improve over text-only approaches.

The bootstrap-based evaluation provided confidence intervals for the official scores, allowing the rankings to be interpreted more

carefully. In particular, several systems obtained close results, especially in the text-only setting, where the confidence intervals of the top systems overlap. Therefore, the results should be read as evidence of the best average performance rather than as clear separation between all adjacent ranks.

Overall, this first edition of SpeechMATICS highlights both the potential and the difficulty of multimodal irony detection in spontaneous spoken Spanish. The task provides a benchmark for future research and shows that there is still considerable room for improvement, especially in the integration of textual, acoustic, and pragmatic information. Future editions will aim to expand the dataset, include additional spoken sources, and encourage the development of more robust multimodal approaches for irony detection.

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).

References

Ahuir, V., A. J. Barceló-Milkova, A. Casamayor-Segarra, V. González Verdegais, L.-F. Hurtado, C. Periñán-Pascual, and M. J. Castro-Bleda. 2026. MESironic: A multimodal dataset and benchmark for irony detection in spontaneous spoken Spanish. *Procesamiento del Lenguaje Natural*, 77(0). To appear.

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

Baevski, A., Y. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Bonet-Jover, A., J. Á. González-Barba, and L. Chiruzzo. 2026. Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org.*

Ca nete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. P'erez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *Proceedings of the Practical Machine Learning for Developing Countries Workshop at ICLR 2020.*

Castro, S., D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria. 2019. Towards multimodal sarcasm detection (an obviously perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Chen, S., C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Zeng, and F. Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Conneau, A., A. Baevski, R. Collobert, A. Mohamed, and M. Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of Interspeech*, pages 2426–2430.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzm'an, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Cortes, C. and V. Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Davis, S. B. and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

de la Rosa, J. et al. 2022. BERTIN: Efficient pre-training of a Spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68:13–23.

Hsu, W.-N., B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

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

Malik, M., D. Tomás, and P. Rosso. 2023. How challenging is multimodal irony detection? In *Natural Language Processing and Information Systems*, pages 18–32, Cham. Springer Nature Switzerland.

Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Pan, R., J. A. García-Díaz, T. Bernal-Beltrán, F. García-Sánchez, and R. Valencia-García. 2025. Overview of SatiSPeech at IberLEF 2025: Multimodal audio-text satire classification in Spanish. *Procesamiento del Lenguaje Natural*, 75:513–522.

Pe naranda, R. et al. 2023. MrBERT: A language model for Spanish social media. *Procesamiento del Lenguaje Natural*.

Perez, J. M., J. C. Giudici, and F. M. Luque. 2021. pysentimiento: A Python toolkit for sentiment analysis and socialnlp tasks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*.

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

Sperber, D. and D. Wilson. 1981. Irony and the use-mention distinction. *Radical Pragmatics*.

Wang, L., N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

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