Corpus Resulting From the DEFALAC Project on Fallacies Detection

Fermín Cruz[1,†], José A. Troyano[1,*,†], Fernando Enríquez[1,†] and F. Javier Ortega[1,†]

[1]*Department of Computer Languages and Systems. School of Computer Engineering. University of Seville. Av. Reina Mercedes s/n. 41012, Seville, Spain.*

Abstract

The *DEFALAC* project focuses on the automatic detection and generation of argumentative fallacies in Spanish using language models based on deep learning. Its main objective is to improve public debate and address the lack of resources in Spanish. This article presents the corpora generated during the execution of the project: *FallacyES*, composed of prototypical fallacies translated and revised from the *Logic* corpus, along with spontaneous fallacies extracted from comments on `meneame.net`; *FallacyES-Political*, containing approximately 2,000 fallacies extracted from 19 Spanish electoral debates over 30 years, classified into 16 types; and *DebatES*, based on the same transcripts as *FallacyES-Political*, enriched with additional information for automated analysis and manual exploration.

Keywords

Detection of fallacies, annotated corpus, large language models

1. Description and objectives of the *DEFALAC* project

The *DEFALAC* project (reference number PID2021-123005), funded by Ministry of Science, Innovation and Universities of the Government of Spain (MCIN), is expected to last four years (2022-2025) and represents the continuation of a consolidated trajectory of ITALICA research team (University of Seville) in the last 20 years. The project proposes the use of large language models and deep learning for the automatic detection and generation of argumentative fallacies, with a focus on Spanish. Using NLP techniques, the project aims to identify, classify, and generate fallacies.

A fallacy is a line of reasoning that appears valid or convincing but is in fact incorrect or misleading. The study of argumentative fallacies dates back to Aristotle, who was the first to compile a catalog of fallacies. Over time, this catalog has expanded and become more structured, with fallacies being classified as either formal (detectable through logical rules) or informal (where this is not possible). Understanding why fallacies are effective is related to the concept of cognitive bias—psychological shortcuts that can distort information interpretation and be exploited by fallacious arguments.

Natural Language Processing (NLP) has advanced in areas such as discourse analysis, argument mining, and *fake news* detection, but research on fallacies is still in its early stages, especially in Spanish. The recent emergence of language models based on deep learning techniques, such as *transformers*, has revolutionized NLP and offers promising tools to address the automatic detection and generation of fallacies.

The motivation behind the *DEFALAC* project stems from the need to improve the quality of public debate, where fallacies distort reasoning and facilitate disinformation in political, health-related, and

SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025.

*Corresponding author.

†These authors contributed equally.

fcruz@us.es (F. Cruz); troyano@us.es (J. A. Troyano); fenros@us.es (F. Enríquez); javierortega@us.es (F. J. Ortega)

https://departamento.us.es/lsi/profesor/cruz-mata-fermin/ (F. Cruz);
https://departamento.us.es/lsi/profesor/troyano-jimenez-jose-antonio/ (J. A. Troyano);
https://departamento.us.es/lsi/profesor/enriquez-de-salamanca-ros-fernando/ (F. Enríquez);
https://departamento.us.es/lsi/profesor/ortega-rodriguez-francisco-javier/ (F. J. Ortega)

0000-0001-9029-9963 (F. Cruz); 0000-0002-9317-3626 (J. A. Troyano); 0000-0002-5427-6331 (F. Enríquez); 0000-0002-6661-2849 (F. J. Ortega)

social contexts. Furthermore, the lack of studies in Spanish represents an opportunity to generate resources and methodologies specific to this language. The main objectives of the project are:

1. Apply the latest advances in language technologies to the automatic detection and generation of fallacies in Spanish.
2. Create and make available to the scientific community resources related to fallacy detection, especially corpora of fallacies in Spanish that can be used to train new models and evaluate systems.

This work focuses on Objective 2, presenting the resources generated during the project's execution, which have been published in various repositories. In the following sections, we present the main features of three resources: the *FallacyES* corpus, the *FallacyES-Political* corpus, and the *DebatES* corpus.

The *FallacyES* corpus is the first annotated fallacy corpus published in Spanish, central to the *DEFALAC* project. It contains two collections: prototypical fallacies, translated and reviewed, and spontaneous fallacies from meneame.net comments. Both include non-fallacious examples, crucial for training fallacy detectors. While prototypical covers 12 types, spontaneous examples focus on 8. Preliminary experiments highlighted the inherent complexity of detecting spontaneous fallacies.

The *FallacyES-Political* corpus comprises 1,965 fallacy instances extracted from 19 Spanish General Election debates spanning 30 years. Transcribed and annotated by experts, each entry provides the fallacy text, type, speaker, and crucial contextual information. This dataset classifies fallacies into 16 distinct types, including new categories like Appeal to Authority. Benchmark evaluations underscored the complexity of the classification task, highlighting benefits of fine-tuning domain-specific models for accuracy.

The *DebatES* corpus, still under development, will facilitate access and analysis of political information. It is based on transcriptions from the same 19 Spanish electoral debates as *FallacyES-Political* corpus, enriched with various data for automatic and manual analyses. Distributed in XML and HTML, annotations include linguistic metrics (e.g., lexical diversity) and political-domain insights. It will also integrate semantic and discourse information, such as thematic blocks and emotions.

2. The *FallacyES* corpus

The first annotated resource produced by the *DEFALAC* project was the FallacyES corpus[1] [1]. To our knowledge, this was the first annotated fallacy corpus published in Spanish. It consists of two collections of fallacies: (1) prototypical fallacies, translated and reviewed from the *Logic* corpus [2], supplemented with manually created non-fallacious examples; and (2) spontaneous fallacies, extracted from online comments on `meneame.net`, also accompanied by non-fallacious examples taken from the same discussions.

The types of fallacies included in the prototypical collection are as follows:

- **Hasty generalization**: drawing a general conclusion based on one or few cases.
- **Ad hominem**: attacking the person or entity making the argument instead of reasoning about the argument itself.
- **Ad populum**: basing the truth (or falsehood) of an argument on its popularity.
- **False causality**: establishing a causal relationship between two phenomena without providing evidence.
- **Circular reasoning**: erroneous reasoning in which the premise and conclusion rely on each other for validation.
- **Appeal to emotions**: attempting to manipulate the audience's emotions to win the debate.
- **Red herring**: introducing a new, unrelated topic to distract from the original debate.

[1] `https://idus.us.es/items/170398a1-81b7-4e08-93cc-ccff0cabd506`

- **Invalid deduction**: presenting a seemingly logical argument with formal errors (false analogy, denying the antecedent, affirming the consequent, etc.).
- **Credibility**: basing the truth (or falsehood) of an argument on the opinion of an authority (argument from *authority*) or on traditionally accepted opinion (*ad antiquitatem* fallacy).
- **False dilemma**: presenting two options as the only possible choices, when in fact more exist.
- **Straw man**: reformulating the opponent's argument in an exaggerated or caricatured way, then attacking that distorted version.
- **Intentional**: any other type of fallacy not included in the above categories, in which the speaker's intent to win the debate without sound reasoning is evident.

For the spontaneous fallacy corpus, examples were collected from comments on the website `meneame.net`, a news aggregator where users discuss current events. To identify fallacies, following the approach used by [3], we searched for comments that mentioned types of fallacies (accusatory comments) and retrieved the messages they referred to (candidate comments). These candidates were then manually reviewed to identify fragments that truly constituted fallacies.

Not all accusations were substantiated, and some comments lacked enough context to clearly identify a fallacy. In total, more than 14,000 candidate comments were analyzed. The aim was to classify fallacies into the same 12 categories as the prototypical collection, but a significant number of examples were found in only 8 categories. The remaining 4 were excluded due to insufficient accusatory comments (fewer than 30) or lack of context needed for accurate classification.

In addition to the type and text, each instance in the resource also includes the headline and summary of the news article from which the comment was extracted. This context could be included as input for fallacy detection and classification models. Table 1 shows the total number of instances for each fallacy type.

Table 1
Number of instances per type included in the spontaneous fallacies section of *FallacyES*.

Type	Instances
Ad hominem	208
Credibility	185
False dilemma	182
Straw man	153
Ad populum	67
Hasty generalization	48
Invalid deduction	42
Appeal to emotions	38
Total fallacies	923
Total non-fallacies	917

A significant contribution of this corpus is the inclusion of non-fallacious examples with similar themes to the fallacies in both corpus sections, which is very useful for training and evaluating fallacy detectors.

In addition to the resource release, preliminary experiments on fallacy detection and classification in Spanish using the *FallacyES* corpus and the language models `Flan-T5` and `RoBERTa-BNE` are detailed in [1]. The results of these experiments highlighted the inherent complexity of the task, especially when dealing with spontaneous fallacies. The performance of transformer-based models was also compared with a logistic regression model using bag-of-words (TF-IDF), showing the importance of the transformers' ability to capture discourse structure in fallacy identification. Finally, experiments were conducted to evaluate the capacity of models trained on prototypical fallacies to classify spontaneous fallacies, revealing limited utility in this scenario.

3. The *FallacyES-Political* corpus

The *FallacyES-Political* dataset[2] [4] consists of examples of fallacies extracted from 19 debates between candidates in Spain's General Elections, broadcast on national radio or television. All debates with a publicly available recording were processed, from the first televised debate in 1993 to those of the 2023 General Elections.

The debates were transcribed using WhisperX [5] and analyzed by three annotators with backgrounds in journalism and philosophy. The resulting dataset comprises 1,965 instances extracted from the speeches of 33 representatives belonging to 11 different political parties. For each instance, the dataset includes the text, type of fallacy, the debate it was extracted from, and the speaker's identity. Additionally, contextual information surrounding the excerpt is provided, as it is essential in some cases to correctly identify the type of fallacy. The fallacy types include 8 that were already present in the *FallacyES* corpus (ad hominem, ad populum, appeal to emotions, red herring, false cause, false dilemma, hasty generalization, and straw man), in addition to the following 8:

- **Appeal to authority**: Refers to a supposed authority (a person, organization, or group) who agrees with or supports a claim, without providing concrete evidence beyond the citation of that authority.
- **Appeal to fear**: A subtype of emotional appeal, in this case to fear. Seeks to persuade the audience that, unless they accept the claim or act in a certain way, negative consequences, disaster, or catastrophe will ensue.
- **Complex question**: A question that contains an implicit assertion or unproven premise, such that answering it implies accepting the hidden assumption.
- **False analogy**: Compares elements or situations that are not truly comparable (or lacks sufficient reasoning to make them comparable), projecting characteristics from one element onto another and drawing conclusions based on this.
- **Flag-waving**: A subtype of emotional appeal that invokes group identity, encouraging audiences who identify with that group to accept the argument uncritically.
- **Poisoning the well**: An intensified form of ad hominem. It consists of a prolonged sequence of negative accusations, or a single extremely harsh accusation, directed at an opponent or group in order to discredit or ridicule them.
- **Slippery slope**: Suggests an improbable or exaggerated outcome as the consequence of a certain action. It typically omits intermediate premises, using an initial assumption as the first step toward an exaggerated conclusion.
- **Slogan**: A short and impactful phrase used to evoke emotions in the audience, often accompanied by another fallacy known as argument by repetition.

The corpus annotation process was divided into several phases. The first step involved creating a catalog of potential fallacy types, including definitions and examples based on existing literature. With this catalog, three annotators analyzed two full debates, identifying as many fallacies as possible. Subsequently, the detected cases were discussed, and 16 fallacy types were selected for the dataset. An initial version of the annotation guide was developed and refined through regular meetings.

The remaining 17 debates were independently annotated by two annotators per debate, aiming to maximize fallacy coverage. Cross-validation was then performed to assess annotation reliability, resulting in an average agreement rate of 47.67%.

Discrepancies were analyzed in meetings between annotators and project researchers, identifying three main causes: (1) overlapping definitions of some fallacies, which led to guide adjustments; (2) subjectivity in identifying certain fallacies, such as false cause; (3) the presence of multiple fallacies within a single text segment, which led to the use of *multilabel* annotation. Finally, the project researchers conducted a final curation, removing doubtful cases and ensuring the text segments were as concise as possible.

Table 2 shows the total number of instances per fallacy type.

[2] https://zenodo.org/records/14836328

Table 2
Number of instances per type included in the *FallacyES-Political* corpus.

Type	Instances
Straw man	351
False cause	304
Ad hominem	285
False dilemma	181
Poisoning the well	152
Appeal to emotions	155
Complex question	74
Ad populum	50
Appeal to fear	115
False analogy	67
Hasty generalization	59
Red herring	28
Appeal to authority	87
Flag-waving	19
Slippery slope	23
Slogan	15
Total fallacies	1,965

To assess the dataset's utility, a benchmark evaluation of state-of-the-art language models was conducted in a *zero-shot* classification setting. The results of the experiments underscored the complexity of the fallacy classification task and the limitations of current LLMs in understanding context-dependent reasoning. Additionally, the experiments demonstrated the benefits of fine-tuning a compact, domain-specific model instead of relying solely on general-purpose LLMs, achieving significant improvements in classification accuracy with a more sustainable approach.

4. The *DebatES* corpus

The analysis of political debates is a task that has attracted significant research interest, as these types of texts exhibit many features that make them particularly suitable for statistical and automated analysis. These include a clear argumentative structure, the repetition of key themes, the use of rhetorical and persuasive techniques, and a wide variety of approaches and styles.

The *DebatES* corpus is the final resource we plan to produce as part of the *DEFALAC* project. It is still under development and therefore not yet publicly available. This corpus aims to serve as a tool to facilitate access to and analysis of political information, particularly the messages that political parties direct at potential voters in the weeks leading up to an election. The resource is based on the textual transcriptions of the same collection of debates used in the *FallacyES-Political* corpus and will include diverse types of information to enable automatic analyses, as well as tools for manual exploration. Specifically, two versions will be distributed: one in XML format for automatic processing, and another in HTML format (with charts, sections, hyperlinks, filters, etc.) to allow convenient navigation throughout the resource. The corpus will include three types of annotations:

1. Descriptive information about the debates, following the style of [6]
2. Linguistic information obtained using natural language processing tools
3. Automatically obtained political-domain annotations, manually reviewed

Regarding linguistic information, the following metrics will be included, automatically computed using the *spaCy* library [7], at both the intervention and participant level for each debate:

- Lexical diversity index
- Proportion of stopwords
- Average sentence length

- Average number of dependencies per verb
- Proportion of punctuation in the text
- Proportion of adjectives
- Proportion of adverbs

To enrich the resource with semantic and discourse-related information, the texts are being processed using the `gemini-2.0-flash-exp` language model [8], with different prompts to automatically obtain the following annotations, which will then be manually validated:

- Thematic blocks
- Summaries of topics addressed at the intervention level
- Relevant mentions of entities
- Concrete proposals made by speakers
- Assertions made by speakers
- Emotions associated with the messages according to Ekman's scale [9]

5. Conclusions

In this work, we have presented the annotated resources developed within the *DEFALAC* project. The project focuses on the automatic detection and generation of argumentative fallacies using large-scale language models based on deep learning, with a particular emphasis on the Spanish language. The main objective of the project is to improve the quality of public debate through the development of tools that can identify and generate fallacies, addressing the current lack of Spanish-language resources for this task. Publishing such resources and making them available to the scientific community contributes to open science and enables other researchers to use them for training new models or evaluating text analysis systems.

We have presented three resources with different characteristics and objectives. The *FallacyES* corpus is composed of two collections: prototypical fallacies translated and reviewed from the *Logic* corpus, with manually created non-fallacious counterparts; and spontaneous fallacies, extracted from online comments on `meneame.net`, also with non-fallacious examples. The second corpus is *FallacyES-Political*, which contains fallacy examples extracted from 19 Spanish electoral debates held over three decades, classified into 16 types of fallacies. The third resource, still under construction, is *DebatES*, based on the textual transcriptions of the same collection of debates used in the *FallacyES-Political* corpus. It will be enriched with various types of information to enable qualitative, quantitative, and automatic analyses.

Acknowledgements

This publication was produced by members of the ITALICA and TIC-134 research groups at the University of Seville. It was funded through the R&D project PID2021-123005, financed by the Ministry of Science, Innovation and Universities of the Government of Spain (MCIN) AEI/10.13039/501100011033/ and "FEDER A way of making Europe".

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check. The authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] F. L. Cruz, J. A. Troyano, F. Enríquez, F. J. Ortega, Detección y clasificación de falacias prototípicas y espontáneas en español, Procesamiento del Lenguaje Natural 71 (2023) 53–62.

[2] Z. Jin, A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, B. Schölkopf, Logical fallacy detection, https://doi.org/10.48550/arXiv.2202.13758, 2022.

[3] S. Y. Sahai, O. Balalau, R. Horincar, Breaking down the invisible wall of informal fallacies in online discussions, in: ACL-IJCNLP 2021-Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.

[4] F. L. Cruz, F. Enríquez, F. J. Ortega, J. A. Troyano, Fallacyes-political: A multiclass dataset of fallacies in spanish political debates, Procesamiento del Lenguaje Natural 74 (2025) 127–138.

[5] M. Bain, J. Huh, T. Han, A. Zisserman, WhisperX: Time-accurate speech transcription of long-form audio, INTERSPEECH 2023 (2023).

[6] J. H. Fiva, O. Nedregård, H. Øien, The norwegian parliamentary debates dataset, Scientific Data 12 (2025) 4.

[7] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength natural language processing in python (2020). doi:10.5281/zenodo.1212303.

[8] G. DeepMind, Gemini 2.0 flash, 2024. URL: https://deepmind.google/technologies/gemini/flash/.

[9] P. Ekman, An argument for basic emotions, Cognition and Emotion 6 (1992) 169–200. doi:10.1080/02699939208411068.