Natural Language Generation research at GPLSI group

Elena Lloret Pastor

elloret@dlsi.ua.es

Wednesday 28th May 2025

Natural Language Processing group @ Utrecht University

Universitat d'Alacant
Universidad de Alicante

GPLSI

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES

Outline

1. About me
2. About GPLSI
3. On-going projects
    1. CORTEX
    2. ILENIA
4. Future projects
    1. ALIA
    2. QUMLAUDE

About me

- Full Professor at University of Alicante (Computer Science background)

Universitat d'Alacant
Universidad de Alicante

TEACHING

Department of Software and Computing Systems

lsi Departamento de Lenguajes y Sistemas Informáticos

RESEARCH

University Institute for Computer Science Research

CENID CENTRO DE INTELIGENCIA DIGITAL PROVINCIA DE ALICANTE

- **Databases Design** (Degree in Computer Science)
- **NLP Introduction** (Master of English and Spanish for Specific Purposes)
- **NLP Applications** (Master in Artificial Intelligence)

GPLSI Language Processing and Information Systems research group(GPLSI)

Elena's research profile:
https://observatorio-cientifico.ua.es/investigadores/362015/detalle

Pictures from: www.google.com, www.ua.es

About me

- Main research topics

GPLSI

2014-2025

NATURAL LANGUAGE PROCESSING

2007-2014

TEXT SUMMARIZATION

My PhD:

Text Summarisation based on Human Language Technologies and its Applications (2011)

NATURAL LANGUAGE GENERATION

2014-2025

PhDs supervised:

- *Proposal of a Hybrid Approach for Natural Language Generation and its Application to Human Language Technologies (2019)*
- *A Discourse-Aware Macroplanning Approach for Text Generation and Beyond (2021)*

About GPLSI

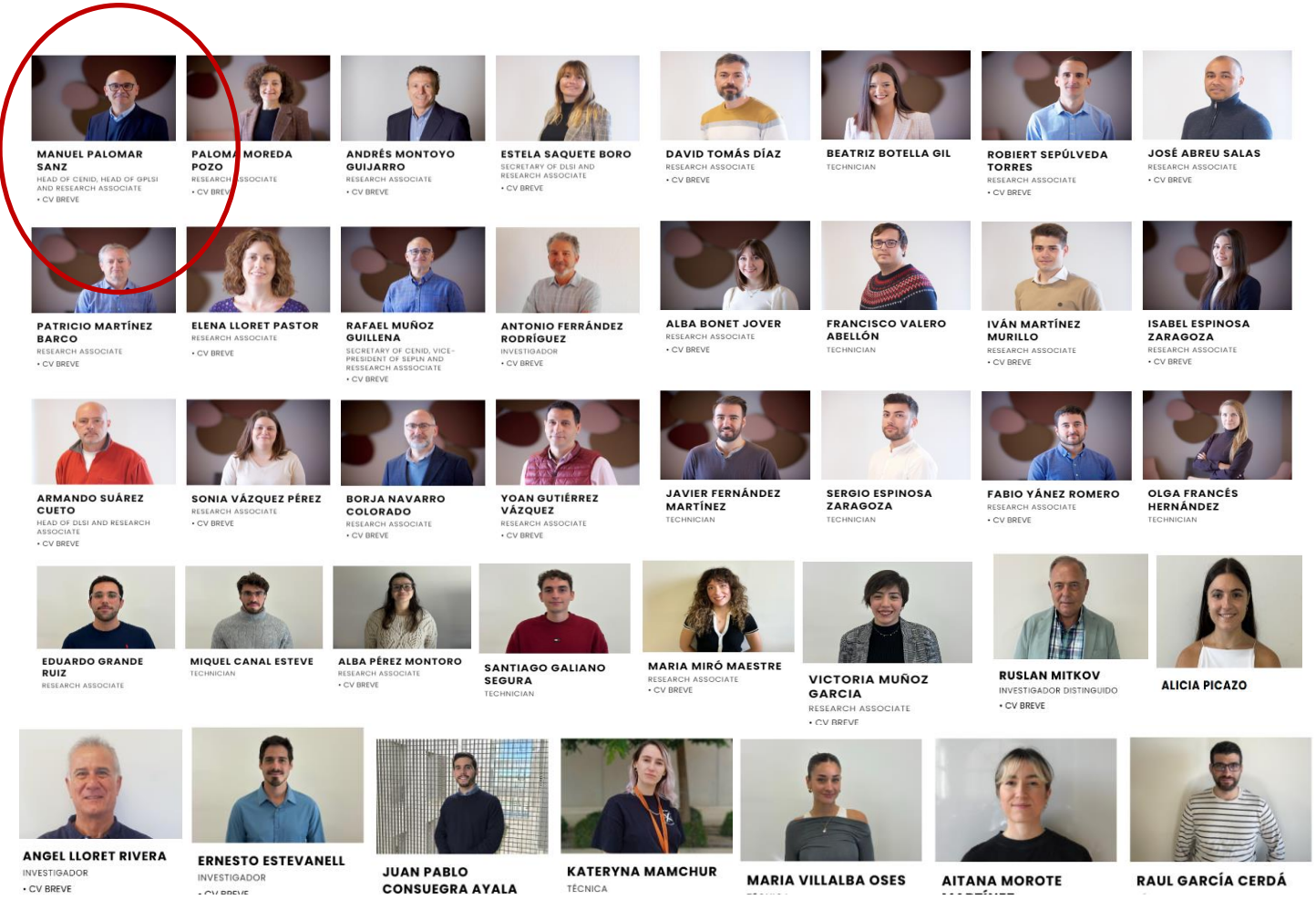
https://gplsi.dlsi.ua.es/

Social networks: @gplsi

About GPLSI

- **Origin**: since 1993
- **University**: University of Alicante (UA)
- **Projects**: working on 9 projects currently
- **Members**: aprox. 40 people
  - Male researchers (5 Full Professors): ~24 (60%)
  - Female researchers (2 Full Professors): ~16 (40%)

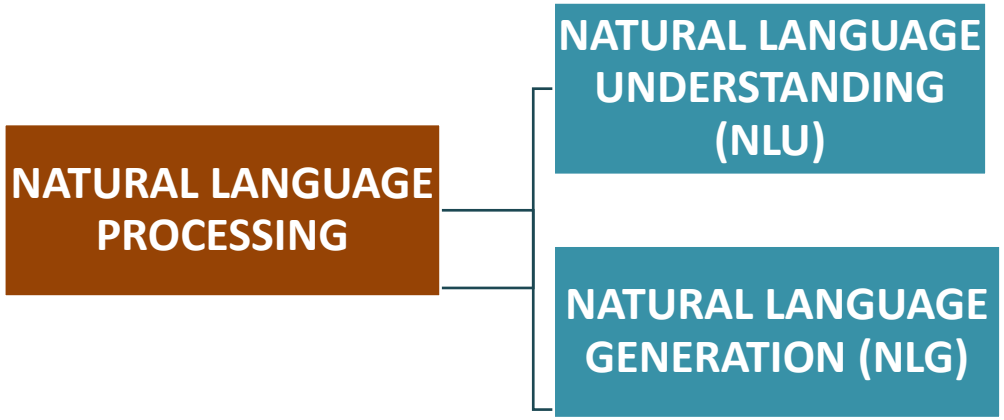
About GPLSI



Team: aprox 40 members

- 7 Full Professors
- 1 Distinguished Fellowship
- 6 Associate Professors
- 1 Assistant Professor
- 3 Post-docs
- 7 PhD students
- 11 technicians associate to projects
- 3 administrative staff

About GPLSI



GPLSI

https://gplsi.dlsi.ua.es/

NATURAL LANGUAGE GENERATION

- Text generation (NLG)
- Text simplification
- Text summarisation
- Conversational robotics (Chatbots)

About GPLSI

CLEARTEXT: Enhancing the modernization of public sector agencies through the implementation of Natural Language Processing to make their digital content CLEARER for individuals with cognitive disabilities.

T2KNOW: Advanced analysis platform for scientific and technical texts for trend extraction and knowledge using NLP techniques

ILENIA: Impulso de las lenguas en la Inteligencia Artifi
[VIVES: Language Technology Plan for Valencian]

CORTEX: Conscious text generation

COOLANG: Language technologies for digital cont

SocialFairness: Analysis of honesty in digital m

NL4dismis: Natural Language Technologies for dealing with disinformation [CIPROM/2021/021]

EATITALL: Artificial intelligence platform for the design and development of new healthy products in the agricultural sector – INNEST/2023/10

GEO.IA: Artificial geointelligence platform to solve problems for citizens and facilitate strategic decision-making in public administrations – INNEST/2023/11

About GPLSI

NLP AREAS

CLEARTEXT: Enhancing the modernization of public sector agencies through the implementa Processing to make their digital content CLEARER for individuals with cognitive disabilities.	Text simplification
T2KNOW: Advanced analysis platform for scientific and technical texts for techniques	Biomedical (NER + classification+ knowledge graphs)
ILENIA: Impulso de las lenguas en la Inteligencia Arti [VIVES: Language Technology Plan for Valencian]	NLU& NLG
CORTEX: Conscious text generatic	NLG
COOLANG: Language t	Fake News + Hate Speech
SocialFairness: Anal	Fake News + Hate Speech
NL4dismis: Natural Language Technologies for dealing with disinformation [CIPR	Fake News + Hate Speech
EATITALL: Artificial intelligence platform for the design and development c sector – INNEST/2023/10	NER and classification
GEO.IA: Artificial geointelligence platform to solve problems for citizens an administrations – INNEST/2023/11	NER and Knowlege-graphs

About GPLSI

NLP AREAS

CLEARTEXT: Enhancing the modernization of public sector agencies through the implementa Processing to make their digital content CLEARER for individuals with cognitive disabilities.	Text simplification
T2KNOW: Advanced analysis platform for scientific and technical texts for techniques	Biomedical (NER + classification+ knowledge graphs)
ILENIA: Impulso de las lenguas en la Inteligencia Arti [VIVES: Language Technology Plan for Valencian]	NLU& NLG
CORTEX: Conscious text generatic	NLG
COOLANG: Language t	Fake News + Hate Speech
SocialFairness: Anal	Fake News + Hate Speech
NL4dismis: Natural Language Technologies for dealing with disinformation [CIPR	Fake News + Hate Speech
EATITALL: Artificial intelligence platform for the design and development c sector – INNEST/2023/10	NER and classification
GEO.IA: Artificial geointelligence platform to solve problems for citizens an administrations – INNEST/2023/11	NER and Knowlege-graphs

On-going projects

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)

https://cortex.gplsi.es/en/home/

Start: 01/09/2022 → End: 31/08/2025

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

On-going projects

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)

https://cortex.gplsi.es/en/home/

Start: 01/09/2022 → End: 31/08/2025

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

How to generate language in reliable manner, without hallucinations?

On-going projects

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)

https://cortex.gplsi.es/en/home/

Start: 01/09/2022 → End: 31/08/2025

How to generate language in reliable manner, without hallucinations?

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

How can we preserve languages in danger of digital extinction and develop LLMs that understand and speak such languages?

On-going projects

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)

https://cortex.gplsi.es/en/home/

Start: 01/09/2022 → End: 31/08/2025

How to generate language in reliable manner, without hallucinations?

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

How can we preserve languages in danger of digital extinction and develop LLMs that understand and speak such languages?

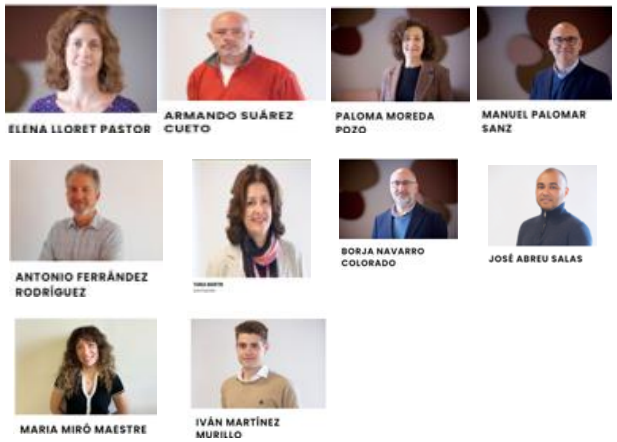
CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)

https://cortex.gplsi.es/en/home/

Start: 01/09/2022 → End: 31/08/2025

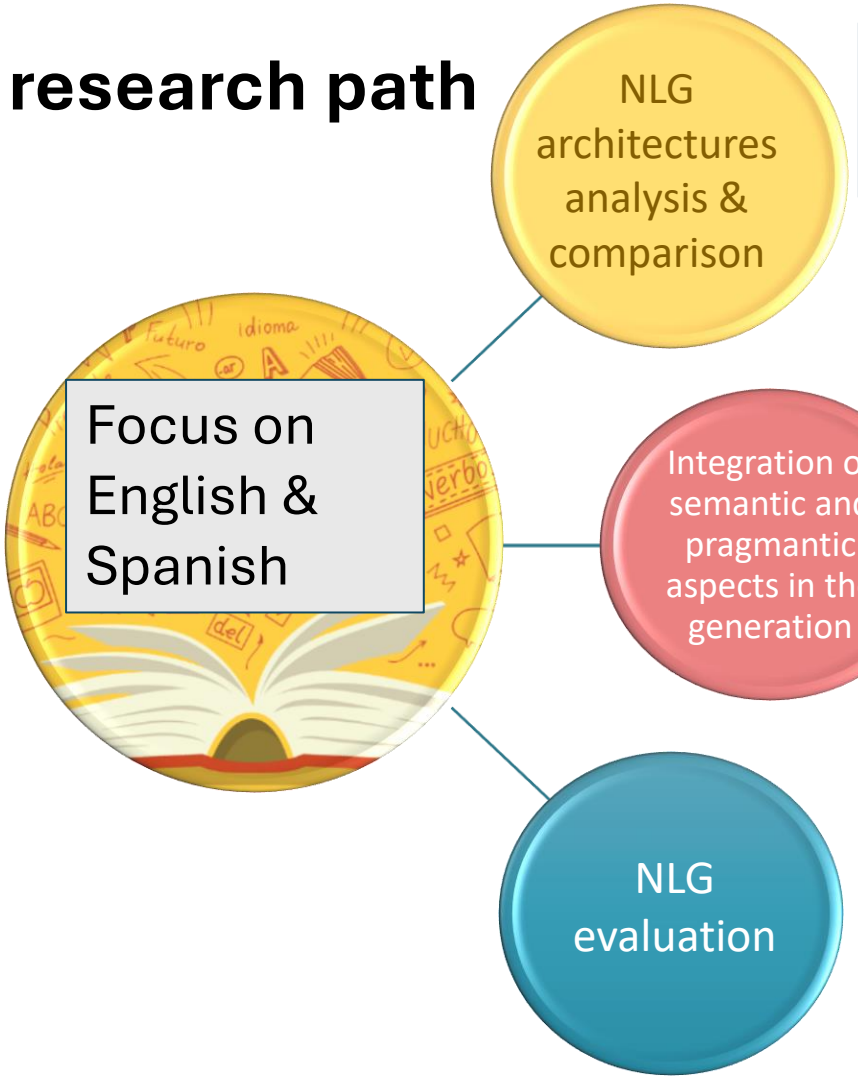
- **Challenge/objective** → to investigate, propose and improve knowledge-enhanced NLG architectures to automatically produce reliable, truthful, and quality texts, avoiding hallucinations.



CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Main research path**



Define a roadmap for NLG and detect LLMs gaps

Knowledge-guide generation (commonsense generation)

Controlled generation

On-going PhD:

Controllability and external knowledge integration as key factors for developing efficient NLG models

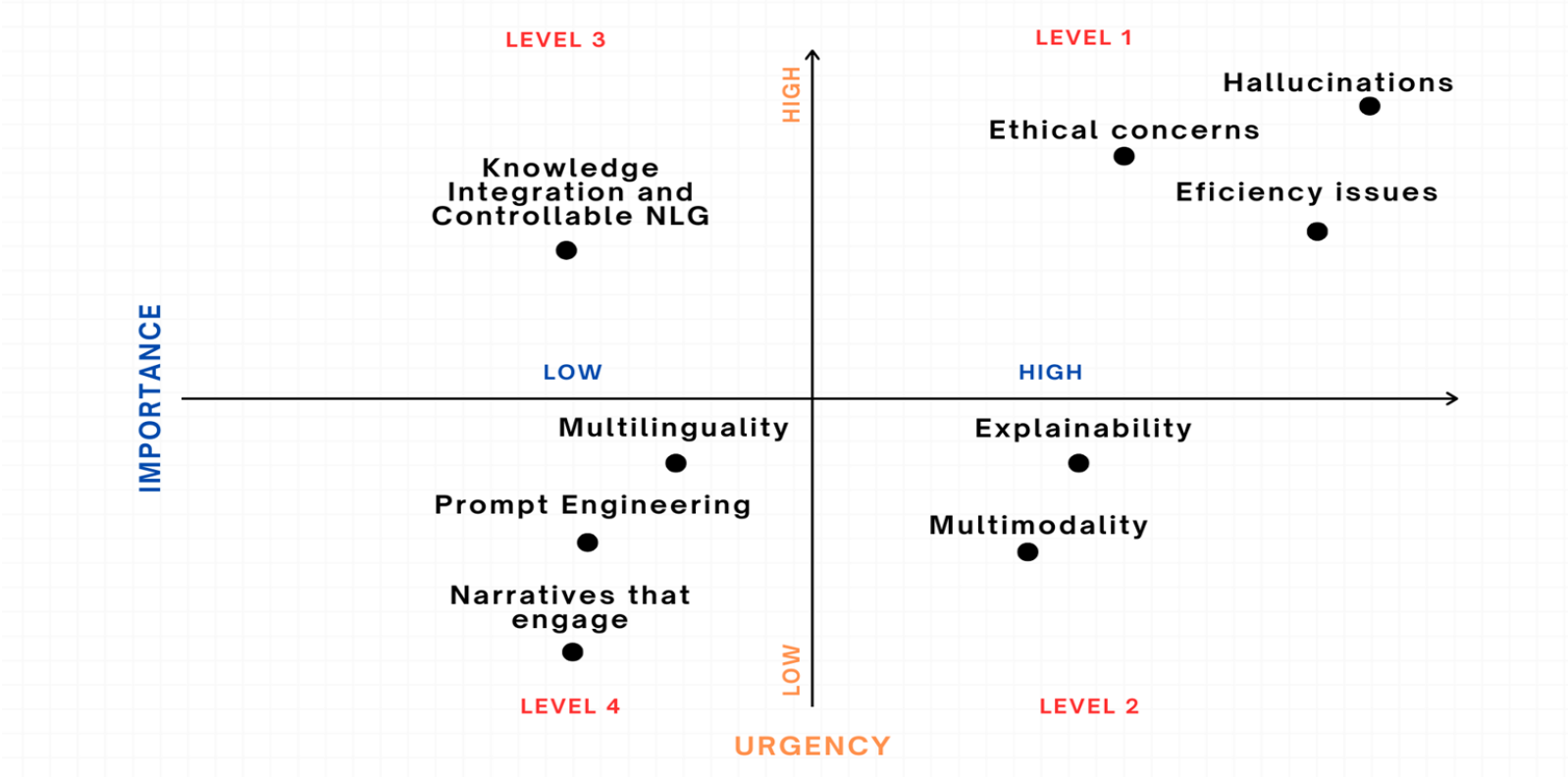
Student: Iván Martínez
Supervisors: Elena Lloret and Paloma Moreda

Analysing the problem of current evaluation metrics

CORTEX project

CORTEX: Conscious Text Generation
(PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Main and current NLG problems**

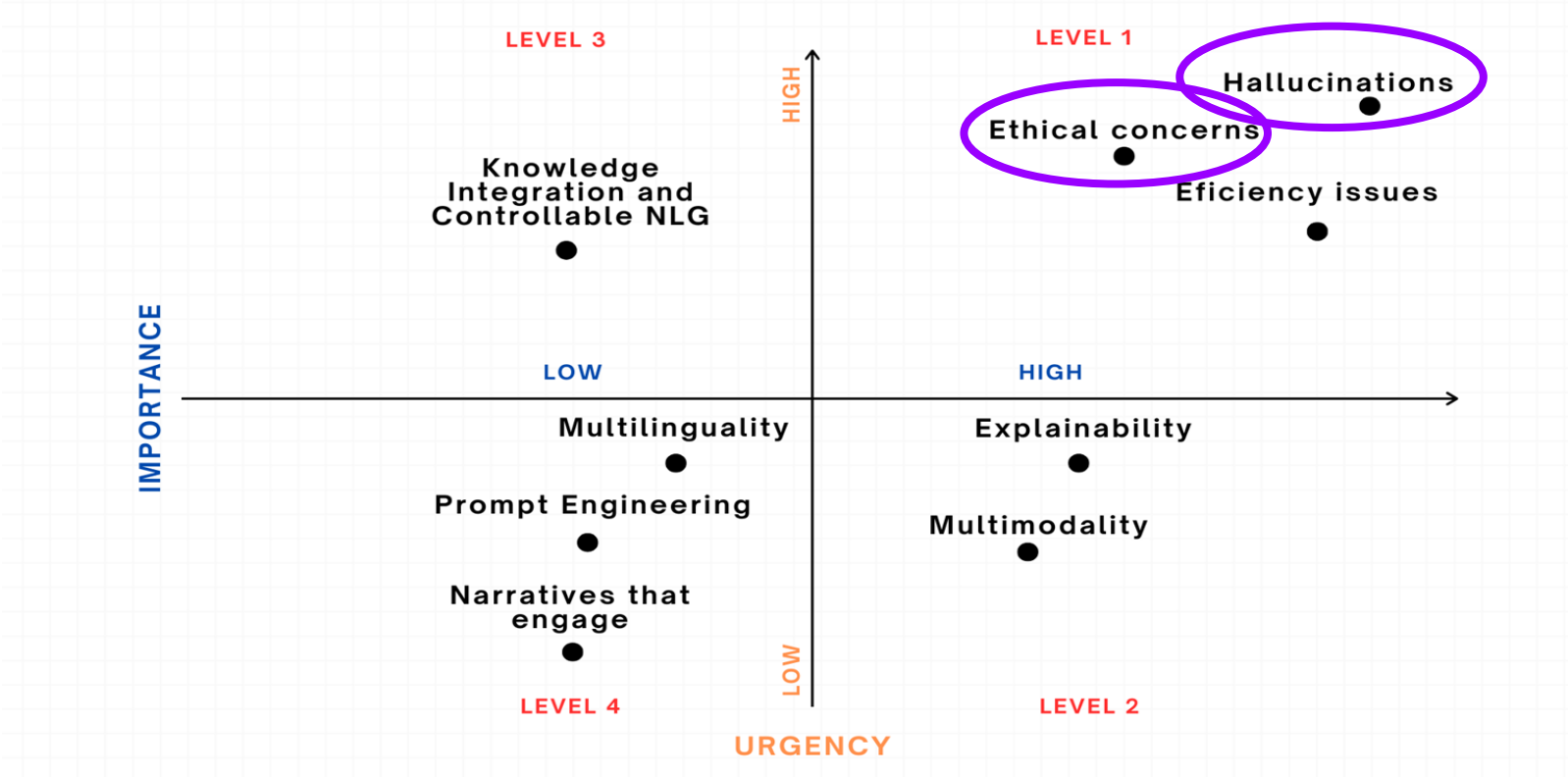


Maestre, M. M., Martínez-Murillo, I., Martin, T. J., Navarro-Colorado, B., Ferrández, A., Cueto, A. S., and Lloret, E. (2025). ***Roadmap for Natural Language Generation: Challenges and Insights***. Procesamiento del Lenguaje Natural, 74.

CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Main and current NLG problems**

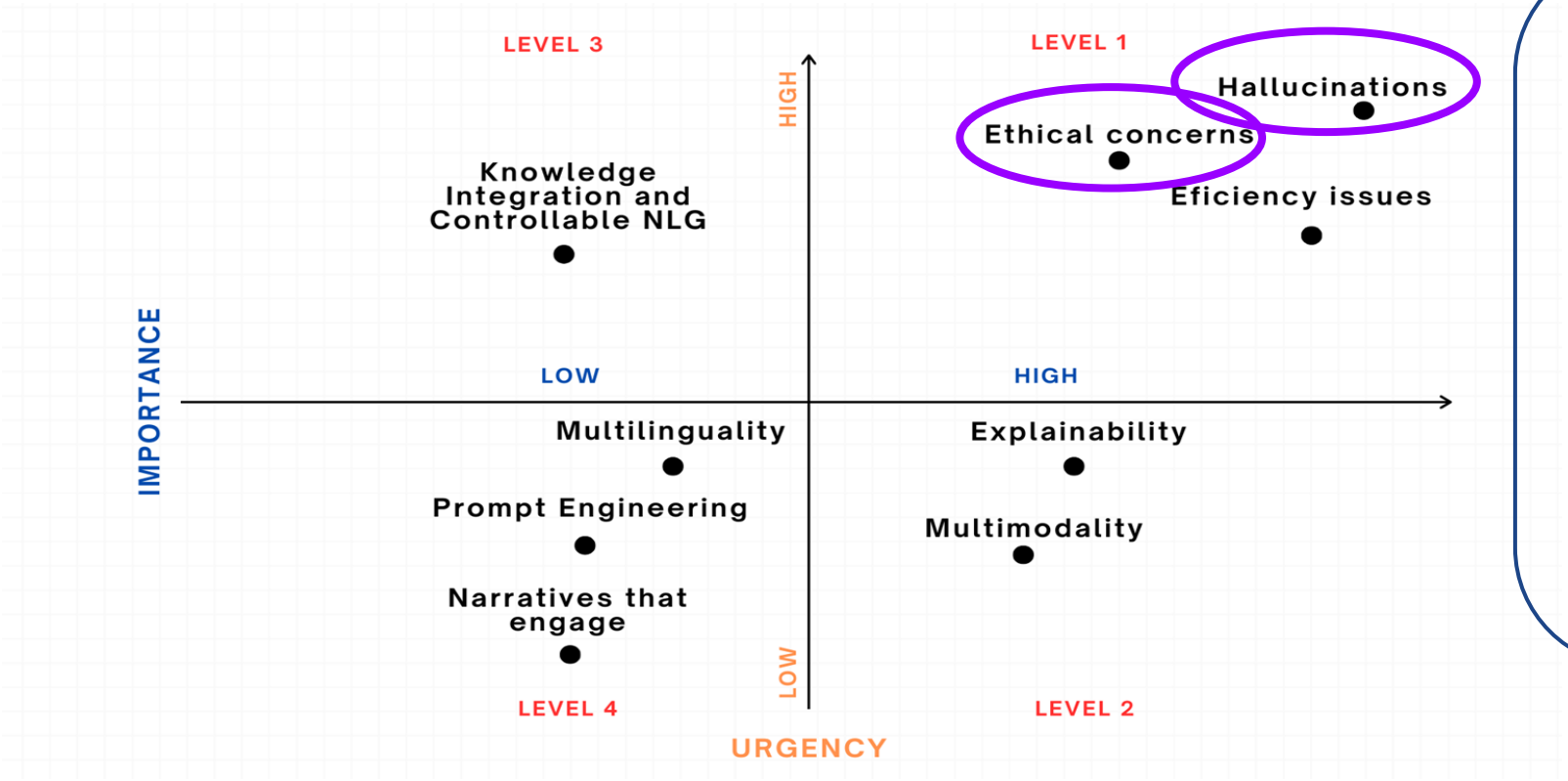


Maestre, M. M., Martínez-Murillo, I., Martin, T. J., Navarro-Colorado, B., Ferrández, A., Cueto, A. S., and Lloret, E. (2025). ***Roadmap for Natural Language Generation: Challenges and Insights***. Procesamiento del Lenguaje Natural, 74.

CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Main and current NLG problems**



- Hallucinations
  - Lack of commonsense
- Ethical concerns
  - Detect bias in NLG
  - Avoid misinformation

Maestre, M. M., Martínez-Murillo, I., Martin, T. J., Navarro-Colorado, B., Ferrández, A., Cueto, A. S., and Lloret, E. (2025). ***Roadmap for Natural Language Generation: Challenges and Insights***. Procesamiento del Lenguaje Natural, 74.

CORTEX project

CORTEX: Conscious Text Generation
(PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Detecting gender bias in NLG datasets and models**

a. CommonGen y C2Gen

b. SimpleNLG y T5

Controlled generation: Generate a sentence from 3-4 words

Generate a sentence with the words: "**baby**", '**diaper**', "**change**".

"The mother changed the baby's diaper"

Consuegra-Ayala, J. P., Martínez-Murillo, I., Lloret, E., Moreda, P., & Palomar, M. (2024). ***A multifaceted approach to detect gender biases in Natural Language Generation***. *Knowledge-Based Systems*, *303*, 112367.

CORTEX project

CORTEX: Conscious Text Generation
(PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

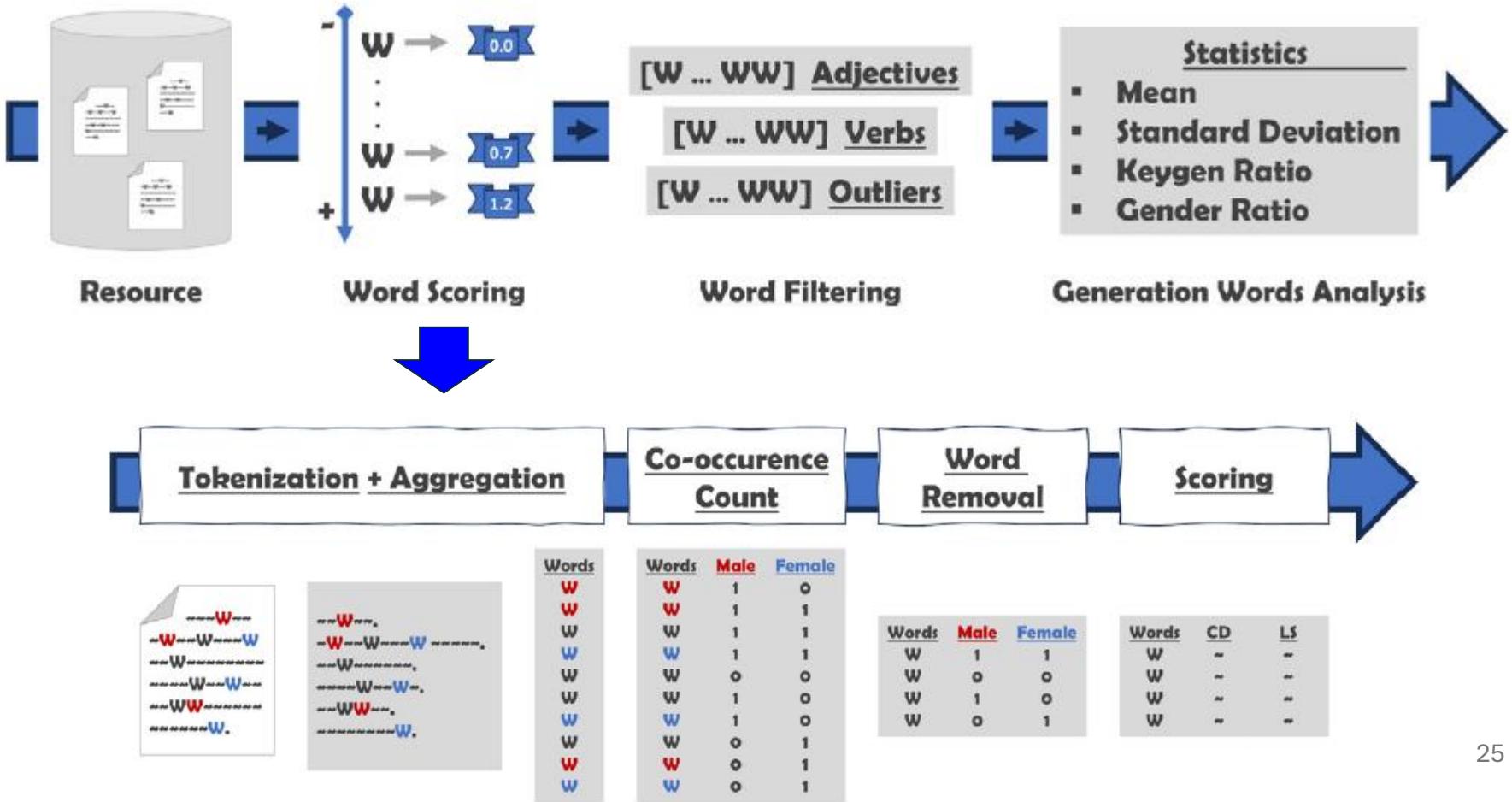
- **Detecting gender bias in NLG datasets and models**

Consuegra-Ayala, J. P., Martínez-Murillo, I., Lloret, E., Moreda, P., & Palomar, M. (2024). ***A multifaceted approach to detect gender biases in Natural Language Generation***. *Knowledge-Based Systems*, *303*, 112367.

CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Detecting gender bias in NLG datasets and models**



CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Detecting gender bias in NLG datasets and models**

Examples of biased words in NLG datasets

	CommonGen	**C2Gen**
Most predominant words in **feminine contexts** (top 5 - adjectives and verbs)	vibrant; **adorable**; positive; defiant;parental darken; **contemplate**; hire; borrow; strengthen	Effective; healthy; **attractive**; sad; fashionable **Understand**, carry, feel, ordering, exasperate
Most predominant words in **masculine contexts** (top 5 - adjectives and verbs)	incorrect; **angy**; moderate; **alcoholic**; ecumenical; accept; illustrate; evaluate; compare; **earn**	intensed; pleased; **fearsome**; tough;restful **Inspire**, operate, carve, sail, divide

CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Detecting gender bias in NLG datasets and models**

Take-home messages

- Methodologies based on word co-occurrence are useful for measuring intrinsic bias in NLG data and models.
  - CommonGen and C2Gen datasets contain gender bias → main causes: stereotypes encoded in the representation of the phrases/words
  - Traditional NLG algorithms (SimpleNLG) only present gender bias produced by the keygens, i.e., list of words used as seeds by the model to produce a sentence.
  - NLG algorithms based on Transformers (T5) present gender bias both produced by the keygens and by the stereotypes encoded in the representation.

CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Knowledge integration in NLG models**

a. Pragmatic knowledge (communicative intent)

Controlled generation: Generate a sentence from 3-4 words

Generate a **directive** sentence with the words: "**coffee**", '**pour**', "**cup**".

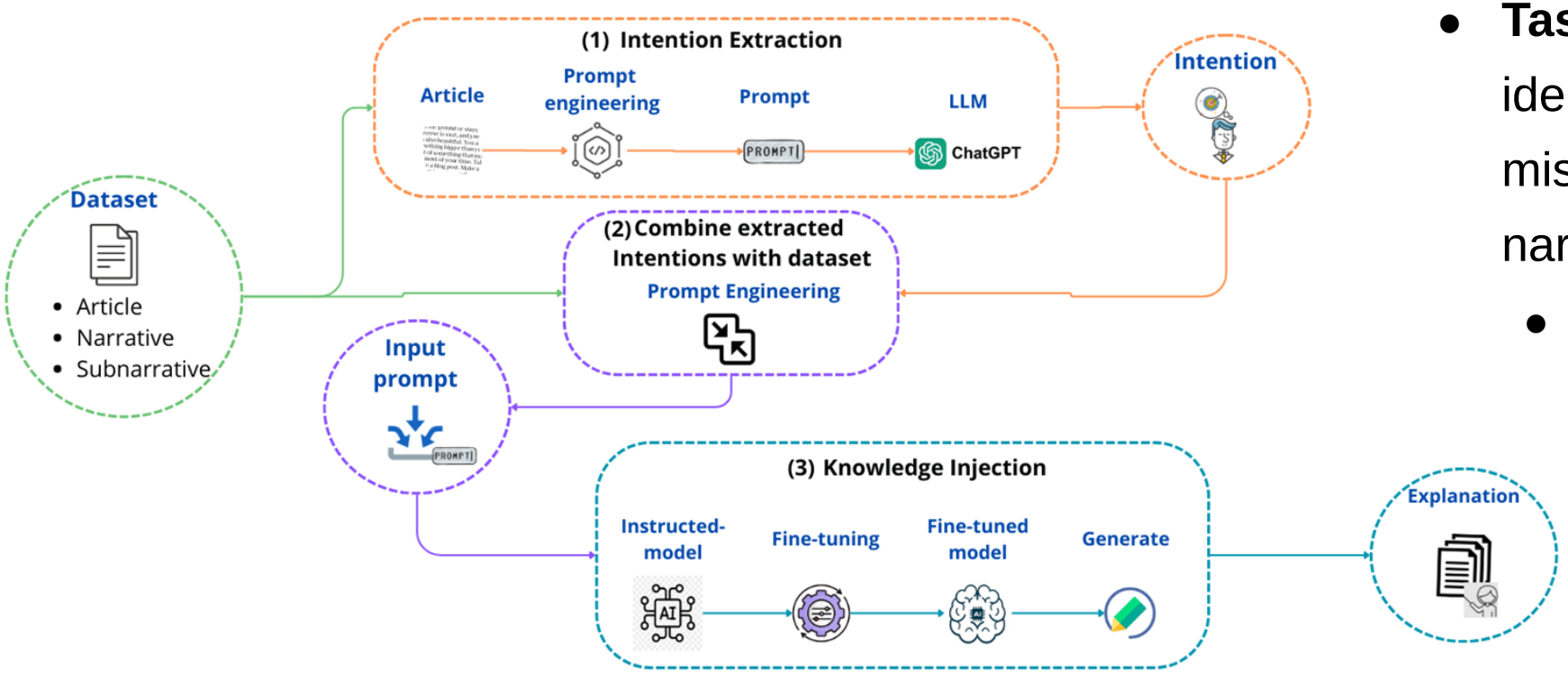
"Pour the coffee into the cup."

Iván Martínez-Murillo, María Miró Maestre, Aitana Morote Martínez,Snorre Ralund, Elena Lloret, Paloma Moreda, Armando Suárez Cueto. **GPLSICORTEX at SemEval-2025 Task 10: Leveraging Intentions for Generating Narrative Extractions**. Semeval 2025 (Accepted. In press)

CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Knowledge integration in NLG models**



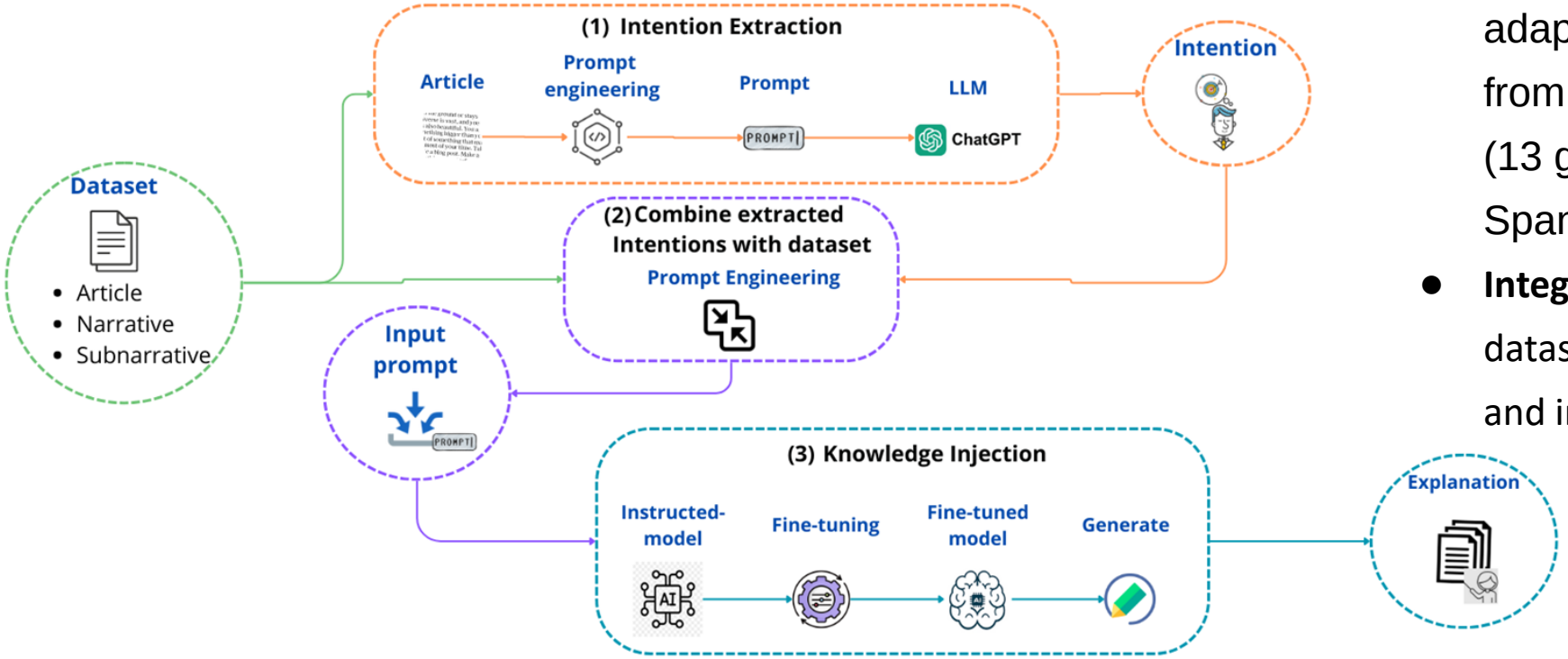
- **Task 10 Semeval-25**: identification and analysis of misinformation in news through narratives:
  - **news + narrative** → generate a free text explanation covering the most important part of the news according to the given narrative = **SUMMARY**.

Iván Martínez-Murillo, María Miró Maestre, Aitana Morote Martínez,Snorre Ralund, Elena Lloret, Paloma Moreda, Armando Suárez Cueto. **GPLSICORTEX at SemEval-2025 Task 10: Leveraging Intentions for Generating Narrative Extractions**. Semeval 2025 (Accepted. In press)

CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Knowledge integration in NLG models**



- **Intent detection and extraction** → adaptation of the method and prompt from (Maestre et al., 2025) to English (13 global intents; 84% of macroF1 in Spanish for ChatGPT)
- **Integration of intent knowledge** into dataset and pre-trained model training and inference (Flan T5-Large).

Iván Martínez-Murillo, María Miró Maestre, Aitana Morote Martínez,Snorre Ralund, Elena Lloret, Paloma Moreda, Armando Suárez Cueto. **GPLSICORTEX at SemEval-2025 Task 10: Leveraging Intentions for Generating Narrative Extractions**. Semeval 2025 (Accepted. In press)

CORTEX project

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Knowledge integration in NLG models**

Position	System	Precision	Recall	Macro F1
1	KyuHyunChoi	0.77686	0.73517	0.75040
2	WordWiz	0.75464	0.73705	0.74551
3	**GPLSICORTEX**	**0.75175**	**0.73274**	**0.74280**
4	TechSSN	0.73886	0.74568	0.74203
5	NarrativeNexus	0.71991	0.74267	0.73085
14	Baseline	0.65144	0.68344	0.66690

- Dataset:
  - 203 texts (training);
  - 30 texts (validation);
  - 68 texts (test)
- Evaluation metric: BertScore
- NLG model: Flan T5-Large
- Language: English
- Total participants: 15

CORTEX project

CORTEX: Conscious Text Generation
(PID2021-123956OB-I00)
https://cortex.gplsi.es/en/home/
Start: 01/09/2022 → End: 31/08/2025

- **Knowledge integration in NLG models**

Take-home message

- Determining the **communicative intent** of the message **is crucial**
  - to know **what and how to generate information**
  - to effectively **detect** articles disseminating **false information**

On-going projects

CORTEX: Conscious Text Generation (PID2021-123956OB-I00)

https://cortex.gplsi.es/en/home/

Start: 01/09/2022 → End: 31/08/2025

How to generate language in reliable manner, without hallucinations?

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

How can we preserve languages in danger of digital extinction and develop LLMs that understand and speak such languages?

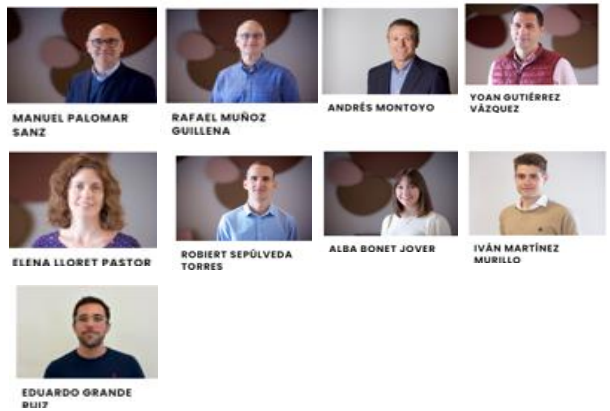
ILENIA project

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

- **Challenge/objective** → To generate multilingual resources that allow the development of applications in all the languages of Spain (Spanish, Catalan, Galician, Euskera and Valencian), promoting the new digital economy based on natural language.
- Expected results:
  - Creation of the first Iberian multilingual voice model
  - Creation of a text model that incorporates European languages
  - Automatic translation between all the languages of Spain
  - Have a corpus of almost 300 billion words (the largest), useful for training models



ILENIA project

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

ILENIA is a common and coordinated project between the following state centers: BSC-CNS (AINA project), CENID (VIVES project), HiTZ (NEL-GAITU project) and University of Santiago de Compostela (NÓS project), which aims to generate digital resources that allow the development of multilingual applications in the different languages of Spain.

AINA CATALÁN → Barcelona Supercomputing Center

GAITU EUSKERA → HiTZ (Basque Country)

NOS GALLEGO → CITIUS, University of Santiago de Compostela

VIVES VALENCIANO → Center of Digital Intelligence, University of Alicante

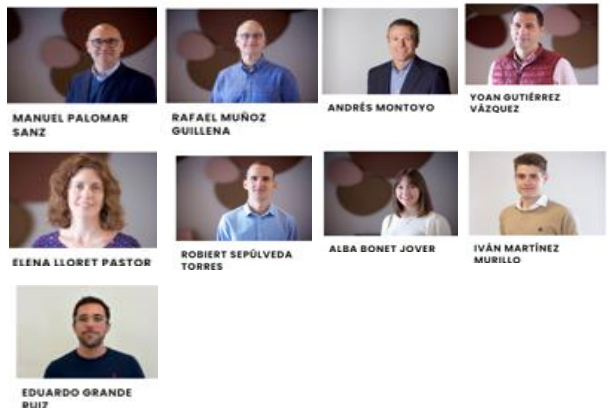
ILENIA project

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

- **VIVES subproject** → VIVES Language Technologies Plan: creation of linguistic resources and language models for Valencian language
- **Objectives:**
  - Identification of Valencian varieties.
  - Create corpora (text+voice) and develop language models for the different varieties of Valencian language
  - Create language models for Valencian language

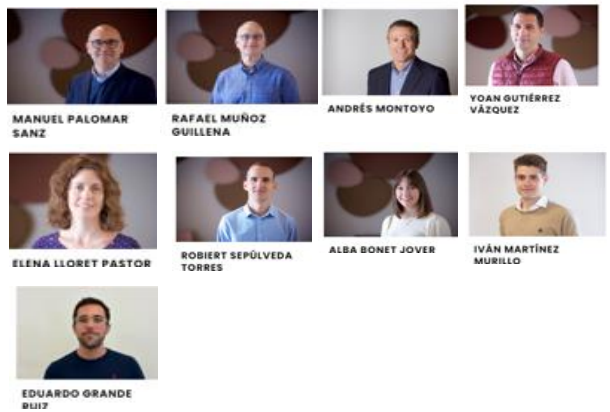


ILENIA project

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025



- **Valencian vs. Catalan**
  - GLOTTOLOG (https://glottolog.org/glottolog/language)
    - Catalán CAT (ISO 639-3)
    - Valencian VSV (ISO 639-3)
  - Both are co-oficial languages. Similar/same language but lexical and syntactic differences
    - *Seva (cat) - seua (vsv) <hers>*
    - *aquest (cat) - este (vsv) <this>*
    - *eina (cat) - ferramenta (vsv) <tool>*
    - *sortir (cat) – eixir (vsv) <go out>*
    - *Hagi vingut (cat) – haja vingut (vsv) <have come>*

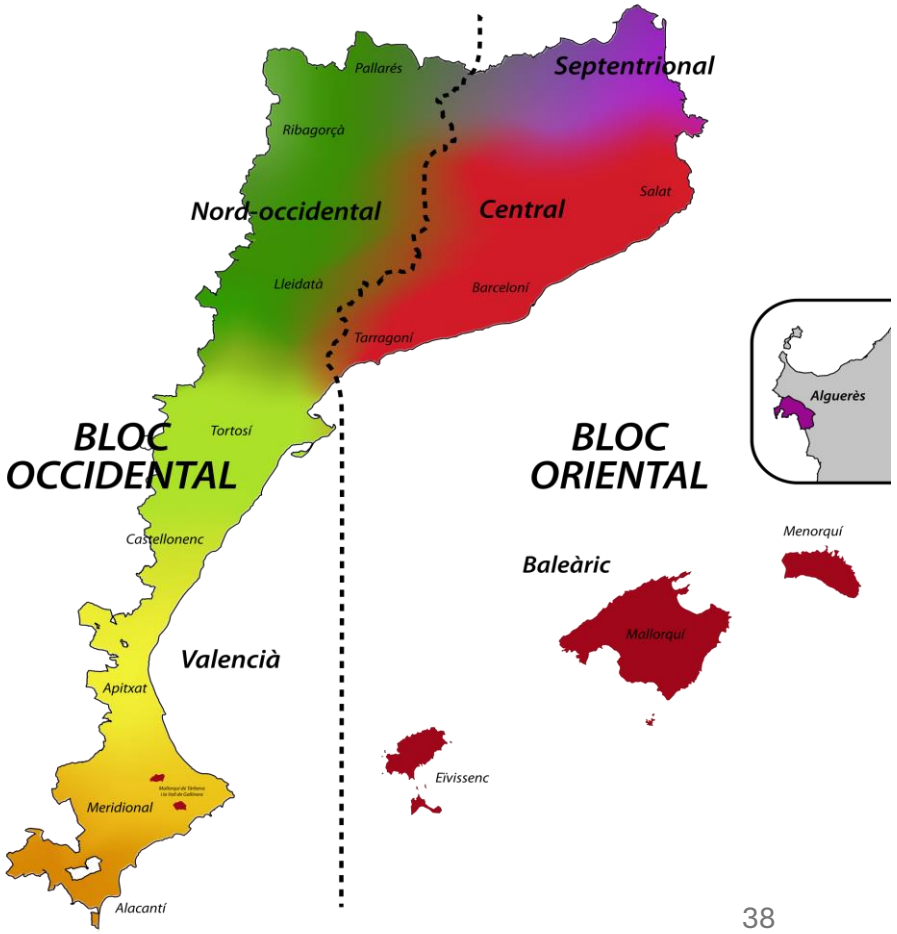
On-going projects

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

- **Catalan and Valencian language varieties**
  - AINA project
    - **Occidental block (West side)**
      - Nord-occidental (Noth-west)
    - Bloc oriental (East side)
  - VIVES project
    - **Bloc occidental (West side)**
      - València de transició o tortosí
      - València septentrional o castellonenc
      - València central o apitxat
      - València meridional
      - València alacantí



ILENIA project

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

- **Language Model →** AITANA-6.3B
  - First LLM for Valencian language
  - Developed using continual pre-training based on FLOR model [FLOR-6.3B] (https://huggingface.co/projecte-aina/FLOR-6.3B) with data in valencian language
  - Current versión →1.304 milion tokens
  - It It is bcan be used for NLG
  - eing fine-tuned and instruction-tuned for specific tasks (e.g. summarization, information extraction, machine translation)

AITANA

ILENIA project

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

- **Corpora**

- VIVES.TEXT_dogv: Valencian region oficial diary (DOGV). 283 millions of tokens.

- VIVES.TEXT_boua: UA oficial diary(BOUA). 7 millions of tokens.

- VIVES.TEXT_Les_Corts: Valencian Parliament. 57.05 millions of tokens.

- VIVES.TEXT_amics: newspapers/blogs (AMIC repository). 30 millions of tokens.

ILENIA project

ILENIA: Impulse of Languages in Artificial Intelligence (2022/TL22/00215337)

https://proyectoilenia.es/en/

Start: 01/01/2023 → End: 31/12/2025

- VIVES resources available at: https://gplsi.gitbook.io/vives-kit/

- ILENIA resources available at: https://proyectoilenia.es/en/resources-models-datasets/

Future projects

- **THE PUBLIC AI INFRASTRUCTURE IN SPANISH AND CO-OFFICIAL LANGUAGES (**https://alia.gob.es/eng/)
- The aim is to facilitate the creation of a new generation of innovative technological resources and services enriched by the immense **linguistic heritage of Spanish**, spoken by 600 million people in the world**, and the co-official languages.**
- ALIA kit: https://langtech-bsc.gitbook.io/alia-kit

Future projects

QUMLAUDE

- **QUANTUM MECHANICS FOR LANGUAGE UNDERSTANDING AND GENERATION**
- **Main hypothesis →** the employment of quantum information and computing theory would be beneficial for addressing complex NLP tasks and applications including language understanding and generation, such as text summarisation and simplification.
- **Goal →** to explore questions related to the application of quantum theory to text encoding and representation, the development of pure-quantum or hybrid language models, and the proposal of new approaches for NLP relying on quantum theory.

THANK YOU VERY MUCH FOR YOUR ATTENTION!

QUESTIONS?

Elena Lloret Pastor

elloret@dlsi.ua.es

Wednesday 28th May 2025

Natural Language Processing group @ Utrecht University

This research work is part of the R\&D projects ``CORTEX: Conscious Text Generation'' (PID2021-123956OB-I00), funded by MCIN/ AEI/10.13039/501100011033/ and by ``ERDF A way of making Europe'' and through the "ILENIA" project (grant number 2022/TL22/00215337) and "VIVES" subproject (grant number 2022/TL22/00215334)

Universitat d'Alacant
Universidad de Alicante

GPLSI

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Natural Language Generation research at GPLSI group

Elena Lloret Pastor

elloret@dlsi.ua.es

Wednesday 28th May 2025

Natural Language Processing group @ Utrecht University

Universitat d'Alacant
Universidad de Alicante

GPLSI

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES