The NL4XAI program: A retrospective

Jose M. Alonso-Moral
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela,
15782 Santiago de Compostela, Spain
josemaria.alonso.moral@usc.es

Abstract

The "Interactive Natural Language Technology for Explainable Artificial Intelligence" (NL4XAI) program was developed from 2019 to 2024 as a Marie Curie Doctoral Network. It was an interdisciplinary program for training a new generation of researchers ready to make Artificial Intelligence (AI) self-explaining, i.e., with the focus on generation and evaluation of interactive explanations in natural language. The consortium comprised 20 institutions, including 10 beneficiaries and 10 partner organizations. A total of 16 early-stage researchers (ESRs) from 11 countries worldwide were involved in 11 doctoral projects, with 7 researchers having successfully completed their PhD degree to date (and 3 more PhD theses expected to be published in 2026).

1 Introduction

The conception of the NL4XAI program dates back to 2016, when I was a visiting researcher at the University of Aberdeen (Scotland, UK) under the supervision of Prof. Ehud Reiter and Prof. Kees van Deemter. At that time, the DARPA Challenge on Explainable AI (XAI) was published (Gunning et al., 2021). In addition, I attended a talk on Marie Curie Training Networks in Aberdeen and began developing a training program focused on XAI and Natural Language Generation (NLG) (Reiter, 2025). Two years later, during the INLG2018 conference, a preliminary program was discussed with Professors Reiter and van Deemter, but also with other colleagues who are well-known in the NLG community, such as Albert Gatt, Claire Gardent, and Mariët Theune. They all joined the program proposal that was successfully evaluated and granted in 2019.

The NL4XAI research and training program was organized with the aim of covering the following four technical pillars: (i) designing and developing XAI models; (ii) enhancing NLG for XAI; (iii) exploiting argumentation technology for XAI; and (iv) developing interactive interfaces for XAI. Training, dissemination, management, and ethics were also handled. The program included 11 doctoral projects with cross-linked technical work packages (see Table 1 in the Appendix A for further details). It is worth noting that the recruited researchers acquired technical and soft skills through a training program that included network-wide events (e.g., specialized workshops, industry days, etc.) and complementary local activities at the host institutions.

2 The Role of NLG in NL4XAI

Quoting an informal conversation with Prof. Reiter when I was in Aberdeen: "Natural Language is the preferred modality for humans to convey explanations, with visualization as a complementary modality". This is the main motto behind NL4XAI (J.M. Alonso et al., 2020).

Accordingly, recruited researchers addressed complementary approaches to generating and evaluating explanations in natural language. For example, they produced novel benchmarks to assess the grounding capabilities of large, pretrained, deep neural models for both natural language understanding and generation (I. Kesen et al., 2024). They also published a new dataset consisting of verbal descriptions of different types of visual inputs, at different levels of abstraction (Cafagna et al., 2023b). In addition, they conducted in-depth studies of state-of-the-art explainability methods and their prospects for explanations to be verbalized through fluent narratives (E. Mariotti et al., 2024).

3 NLG Evaluations in NL4XAI

Prof. Reiter taught researchers in the NL4XAI program about the importance of conducting rigorous and reproducible evaluations (Belz and Reiter, 2006; A. Belz et al., 2023). Accordingly, some

ESRs proposed new automatic metrics, while most conducted several user studies to evaluate their work. For example, Ettore (ESR1) and Adarsa (ESR3) collaborated to develop a new metric for evaluating surrogate XAI models (Mariotti et al., 2023). The new metric, shapGAP, serves to identify reliable surrogate models, paving the way for more robust applications. In addition, Eduardo (ESR4) collected human judgments on naturalness and examined how incorporating formulaicness into existing metrics affects alignment with these judgments (E. Calò et al., 2025). Other researchers paid attention to how to communicate explanations in user studies on tabular data (Sivaprasad and Reiter, 2024) as well as on vision and language (Cafagna et al., 2023a). They also conducted user studies to determine the impact of cognitive bias mitigation approaches (Rieger et al., 2024).

4 Conclusions

To sum up with, the NL4XAI network produced a total of 58 publications, including 11 journal articles, 36 conference papers, 4 book chapters, and 7 doctoral dissertations to date (with three more expected to be published this year). Also, 9 datasets were released. Moreover, the project's members delivered more than 30 presentations at prominent scientific conferences, highlighting their innovative research findings. For the sake of reproducibility and open science, outcomes are openly available at `https://zenodo.org/communities/nl4xai/`.

Acknowledgments

The NL4XAI project received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. The author recognizes the support of the Galician Ministry of Education, Science, Universities and Professional Training (grant CiTIUS 2024-2027 ED431G2023/04) and the European Union (European Regional Development Fund - ERDF). This work is also supported by the Spanish Ministry of Science, Innovation and Universities (MCIN/AEI/10.13039/501100011033/) through the MAIXAI4TRUST grant (PID2024-157680NB-I00).

References

A. Belz et al. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10. ACL.

A. Belz and E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320. ACL.

M. Cafagna, L.M. Rojas-Barahona, K. van Deemter, and A. Gatt. 2023a. Interpreting vision and language generative models with semantic visual priors. *Frontiers in Artificial Intelligence*, 6.

M. Cafagna, K. van Deemter, and A. Gatt. 2023b. HL dataset: Visually-grounded description of scenes, actions and rationales. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 293–312, Prague, Czechia. ACL.

E. Calò et al. 2025. Incorporating formulaicness in the automatic evaluation of naturalness: A case study in logic-to-text generation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 352–365. ACL.

E. Mariotti et al. 2024. TextFocus: Assessing the faithfulness of feature attribution methods explanations in natural language processing. *IEEE Access*, 12:138870–138880.

D. Gunning, E. Vorm, J.Y. Wang, and M. Turek. 2021. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4):e61.

I. Kesen et al. 2024. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. In *International Conference on Learning Representations (ICLR)*.

J.M. Alonso et al. 2020. Interactive natural language technology for explainable artificial intelligence. In *Trustworthy AI - Integrating Learning, Optimization and Reasoning: First International Workshop, TAILOR 2020, Virtual Event, September 4–5, 2020, Revised Selected Papers*, page 63–70, Berlin, Heidelberg. Springer-Verlag.

E. Mariotti, A. Sivaprasad, and J.M. Alonso-Moral. 2023. Beyond prediction similarity: ShapGAP for evaluating faithful surrogate models in XAI. In *Explainable Artificial Intelligence*, pages 160–173. Springer Nature Switzerland.

E. Reiter. 2025. *Natural Language Generation*. Springer Nature Switzerland.

A. Rieger, T. Draws, M. Theune, and N. Tintarev. 2024. Nudges to mitigate confirmation bias during web search on debated topics: Support vs. manipulation. *ACM Trans. Web*, 18(2).

A. Sivaprasad and E. Reiter. 2024. Linguistically communicating uncertainty in patient-facing risk prediction models. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 127–132, St Julians, Malta. ACL.

A NL4XAI Doctoral Projects

Doctoral Projects	Early-stage Researchers	Host Institutions	Supervisors	Doctoral Dissertations
ESR1: Explaining black-box models in terms of grey-box twin models	Ettore Mariotti (Italy)	CiTIUS, University of Santiago de Compostela (Spain)	J.M. Alonso-Moral A. Gatt	A holistic perspective on designing and evaluating explainable AI models: from white-box additive models to post-hoc explanations for black-box models (2024) https://zenodo.org/records/13748400
ESR2: From grey-box models to explainable models	Conor Henessy (Ireland) Nikolay Babakov (Russian Federation)	CiTIUS, University of Santiago de Compostela (Spain)	A. Bugarin E. Reiter	Novel Methods for Bayesian Networks Construction and Explanation using Natural Language (2025) https://hdl.handle.net/10347/45440
ESR3: Explaining Probabilistic Reasoning	Jaime Sevilla (Spain) Adarsa Sivaprasad (India)	University of Aberdeen (UK)	E. Reiter S. Pera	Expected to be published in 2026
ESR4: Explaining logical formulas	Alexandra Mayn (Russian Federation) Eduardo Caló (Italy)	University of Utrecht (The Netherlands)	K. van Deemter J. Levy	Automatically Expressing the Meaning of Logical Formulae in Natural Language (2026) https://doi.org/10.33540/3515
ESR5: Multimodal Semantic Grounding and Model Transparency	Michele Cafagna (Italy)	University of Malta (Malta)	A. Gatt K. van Deemter	Visually grounded language generation: data, models and explanations beyond descriptive captions (2024) https://zenodo.org/records/14052376
ESR6: Explainable Models for Text Production	Juliette Paule Alice Faille (France)	Lorraine Research Laboratory in Computer Science and its Applications, Centre National de la Recherche Scientifique (France)	C. Gardent A. Gatt	Data Based Natural Language Generation: Evaluation and Explanability (2023) https://zenodo.org/records/14231559
ESR7: Argumentation-based multi-agent recommender system	Qurat-ul-ain Shaheen (Pakistan) Jairo Alejandro Lefebre Lobaina (Cuba)	Artificial Intelligence Research Institute (IIIA), CSIC (Spain)	C. Sierra K. Budzynska	Expected to be published in 2026
ESR8: Customized interactive argumentation schemes for XAI	Martijn Demoulin (The Netherlands) He Zhang (China)	Warsaw University of Technology (Poland)	K. Budzynska J.M. Alonso-Moral	Not to be published
ESR9: Personalized explanations by virtual characters	Sumit Srivastava (India)	University of Twente (The Netherlands)	M. Theune A. Catala	Expected to be published in 2026
ESR10: Interactions to mitigate human biases	Alisa Rieger (Germany)	Technical University of Delft (The Netherlands)	S. Pera M. Theune	Striving for Responsible Opinion Formation in Web Search on Debated Topics (2024) https://zenodo.org/records/13768214
ESR11: A framework for explainability in process mining	Luca Nannini (Italy)	Indra (Spain)	S. Barro A. catala	Explainability in Process Mining: A Framework for Improved Decision-Making (2024) https://zenodo.org/records/14162735

Table 1: Summary of details on the doctoral projects that were developed in the NL4XAI program. It is worth noting that some of the early-stage researchers (ESRs) recruited initially resigned from their contracts before completing their projects. Accordingly, they had to be replaced by newly recruited researchers. This is why some rows in the table include two researchers for the same project. This also explains why some doctoral dissertations were published after the program ended, while others are still pending publication. For the *i-th* ESR, the row *i* in the table includes two supervisors. The first one (listed first in the column "Supervisors") served as the main supervisor at the host institution, and the second one served as a co-supervisor at another institution in the consortium. Prof. Ehud Reiter played a central role in this project. Despite significant challenges posed by the COVID-19 pandemic, Brexit-related requirements, and the resignation of some of the recruited researchers, all in all, the work carried out in the project contributed to strengthening European innovation capacity by advancing the state of the art in XAI and by creating new models and technologies.