ROBOT-TALK Project for the Recognition of the Robotic Origin of Texts: Methodologies and Results

The detection of automatically generated texts—commonly referred to as Text Machine Generation (TMG)—has gained increasing attention due to the growing impact of Large Language Models (LLMs) in producing texts of high grammatical and semantic quality. These texts may, in many cases, be indistinguishable from those written by humans (Casal & Kessler, 2023; Uchendu et al., 2023; Jones & Bergen, 2024, among others). While the high quality of such content offers clear benefits, its potential misuse for malicious or criminal ends poses substantial threats to public safety. Documented uses of these texts encompass disinformation campaigns, reputational damage, fraudulent impersonation, and threatening communications (Pavlyshenko, 2022, among others). In such contexts, the availability of reliable methods and tools to distinguish automatically generated texts from those authored by humans becomes essential (Maloyan et al., 2022).

The automatic or semi-automatic detection of texts produced by LLMs remains an unresolved challenge. Consequently, the primary objective of our research is to examine how the methods for identifying machine-generated texts can be improved. This investigation focuses specifically on the Spanish language and is conducted within the framework of the project *"ROBOT-TALK: Recognition of the Robotic Origin of Texts. Task Automation and Linguistic Knowledge,"* PID2022-140897OB-I00, funded by the Spanish Ministry of Science and Innovation.

The initial hypothesis supporting this research relies on observing significant linguistic differences between texts generated by Large Language Models (LLMs) and those written by humans (Alonso Simón et al., 2023). Additionally, our hypothesis proposes that each LLM possesses a unique writing style or an " idiolect".

Our participation in Red ATLAS will focus on presenting the results obtained from the analysis of potentially distinctive linguistic features of texts generated by Large Language Models (LLMs) using a specially created corpus, the ROBOT-TALK corpus. It the first comparable corpus of Spanish texts—including scientific linguistics articles, news reports, and film reviews—authored by humans and by four different Large Language Models (LLMs): OpenAI's ChatGPT-3.5-turbo, OpenAI's ChatGPT-4, Google's Gemini, and Mixtral AI's Mixtral-8x7B-Instruct-v0.1.

In addition, we will present the results obtained from the automatic baseline classification of human-generated versus machine-generated texts using machine learning classifiers without any explicit linguistic features, as well as our current lines of research. Finally, we will conclude the presentation by sharing and discussing with the attendees our preliminary conclusions regarding the challenges of identifying AI generated texts.

Keywords*: authorship identification, large language models, computational linguistics, forensic linguistics*

Ana Fernández-Pampillón Cesteros
Doaa Samy

Área de Lingüística General. Departamento de Lingüística, Estudios Árabes, Hebreos, Vascos y de Asia Oriental. Facultad de Filología. Universidad Complutense de Madrid