Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

GDN-CC : un jeu de données pour la clarification automatique de corpus de consultations citoyennes assistées par l'IA

Pierre-Antoine Lequeu[1], Léo Labat[1,3], Laurène Cave[2], Gaël Lejeune[2]
François Yvon[1] et Benjamin Piwowarski[1]
(1) Sorbonne Université, CNRS, ISIR, Paris, France
(2) Sorbonne Université, STIH/CERES, Paris, France
(3) Institut Polytechnique de Paris, CNRS, CREST, Paris, France
lequeu *(at)* isir.upmc.fr

RÉSUMÉ

Les LLMs sont omniprésents dans le TAL moderne, et bien que leur applicabilité s'étende aux textes produits pour des activités démocratiques telles que les délibérations en ligne ou les consultations citoyennes, des questions éthiques ont été soulevées quant à leur utilisation comme outils d'analyse. Ce travail a deux objectifs : standardiser les contributions au **niveau pragmatique** pour faciliter l'analyse politique, et évaluer la fiabilité de petits LLM à *poids ouverts* (exécutables localement) pour cette tâche. Nous introduisons la tâche de **Clarification de Corpus**, un cadre de prétraitement transformant des données brutes et multi-thématiques en unités argumentatives structurées et autonomes. À cette fin, nous présentons **GDN-CC**, un jeu de données issu du Grand Débat National, comprenant 2 285 unités, clarifiées et annotées manuellement selon leur structure argumentative. Nous partageons aussi **GDN-CC-large**, comprenant 300,000 unités annotées automatiquement.

ABSTRACT

The GDN-CC Dataset : Automatic Corpus Clarification for AI-enhanced Democratic Citizen Consultations

LLMs are ubiquitous in modern NLP, and although their applicability extends to texts produced for democratic activities such as online deliberations or citizen consultations, ethical questions have been raised regarding their use as analysis tools. This work has two objectives : to standardize contributions at the pragmatic level to facilitate policy analysis, and to evaluate the reliability of small, open-weight LLMs (executable locally) for this task. We introduce the task of **Corpus Clarification**, a preprocessing framework that transforms raw, multi-thematic data into structured, self-contained argumentative units. To this end, we present **GDN-CC**, a dataset derived from the Grand Débat National (Great National Debate), comprising 2,285 units, clarified and manually annotated according to their argumentative structure. We also share **GDN-CC-large**, comprising 300,000 automatically annotated units.

MOTS-CLÉS : IA Démocratique, Clarification de Corpus, Minage d'Argument.

KEYWORDS: Democratic AI, Corpus Clarification, Argument Mining.

ARTICLE ACCEPTÉ À : The 64th Annual Meeting of the Association for Computational Linguistics.
URL : https://arxiv.org/abs/2601.14944

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 22–22.