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

Ressource pour l'analyse des erreurs en simplification de textes : nouvelle taxonomie et collection de test

Benjamin Vendeville[1, 2,3] Liana Ermakova[1, 3] Pierre De Loor[2, 4]
(1) Université de Bretagne Occidentale, Brest, 29200, France
(2) Lab-STICC, Brest, 29200, France (3) HCTI, Brest, 29200, France (4) ENIB, Brest, 29200, France
benjamin.vendeville@univ-brest.fr, liana.ermakova@univ-brest.fr,
deloor@enib.fr

RÉSUMÉ

Le grand public est souvent confronté à des textes complexes qu'il n'a pas le temps ni l'expertise de comprendre pleinement. La simplification automatique de textes (SAT) contribue à rendre l'information plus accessible, mais ses méthodes d'évaluation n'ont pas suivi les avancées récentes en génération de texte. Des études ont montré que les métriques actuelles ne corrèlent pas avec la présence d'erreurs, soulignant le besoin d'un meilleur cadre d'évaluation. Cet article réduit ce manque en proposant une taxonomie des erreurs en simplification, un jeu de données parallèle de textes scientifiques simplifiés annotés manuellement selon cette taxonomie, et une analyse des modèles existants pour détecter et classifier ces erreurs. Ces contributions fournissent aux chercheurs les outils nécessaires pour mieux évaluer les erreurs en SAT et améliorer la qualité des textes simplifiés automatiquement.

ABSTRACT

Resource for Error Analysis in Text Simplification : New Taxonomy and Test Collection

The general public often encounters complex texts that they do not have the time or expertise to fully understand. Automatic Text Simplification (ATS) helps make information more accessible, but its evaluation methods have not kept up with recent advances in text generation. Studies have shown that current metrics do not correlate with the presence of errors, highlighting the need for a better evaluation framework. This paper addresses this gap by proposing a taxonomy of errors in simplification, a parallel dataset of automatically simplified scientific texts manually annotated according to this taxonomy, and an analysis of existing models for detecting and classifying these errors. These contributions provide researchers with the tools needed to better evaluate errors in ATS and improve the quality of automatically simplified texts.

MOTS-CLÉS : Simplification de texte ; Classification d'Erreurs ; Hallucinations ; TALN.

KEYWORDS: Text Simplification ; Error Classification ; Hallucinations ; NLP.

ARTICLE ACCEPTÉ À : SIGIR 2025 : The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, Padua, Italy, July 13-17, 2025..
URL : https://dl.acm.org/doi/10.1145/3726302.3730304

Benjamin Vendeville et Liana Ermakova sont financés par l'ANR (ANR-22-CE23-0019-01) et MaDICS CNRS (https://www.madics.fr/ateliers/simpletext/). Les opinions exprimées n'engagent pas les financeurs.

Cette œuvre est mise à disposition sous licence Attribution 4.0 International.