

Évaluation Automatique Explicable de l'Écriture Argumentative : État de l'Art, Lacunes et Proposition d'Architecture Modulaire Alignée sur des Grilles Éducatives

Marcos Moisés Crisóstomo de Oliveira¹

(1) Instituto Federal da Bahia (IFBA) , Universidade do Estado da Bahia (UNEB) PPGEL, Brésil

marcosoliveira@ifba.edu.br

RÉSUMÉ

L'évaluation automatique de l'écriture constitue une frontière prometteuse du Traitement Automatique du Langage Naturel (TALN), en particulier pour l'enseignement de l'argumentation. Pourtant, la majorité des systèmes existants privilégient la prédiction de scores au détriment de la compréhension structurelle des textes et de la production de retours pédagogiques utiles. Cet article propose une architecture explicable, modulaire et adaptée à la Compétence III de l'ENEM (Brésil), axée sur trois piliers : l'alignement avec des grilles officielles, la transparence des décisions algorithmiques, et la valeur formative des retours générés. Articulant théorie de l'argumentation, évaluation de l'écriture et technologies récentes en TALN, cette architecture comprend quatre modules : segmentation des unités argumentatives, classification des relations discursives, alignement aux critères d'évaluation, et génération de feedback. Les résultats initiaux montrent un fort potentiel pour améliorer l'équité, l'auditabilité et l'utilité pédagogique du système, tout en ouvrant la voie à des adaptations multilingues dans d'autres contextes d'évaluation.

ABSTRACT

Automated writing assessment is a promising frontier in Natural Language Processing (NLP), especially in teaching argumentative writing. However, most existing systems focus on score prediction at the expense of structural understanding and formative feedback. This paper proposes an explainable and modular architecture tailored to the Competency III of the Brazilian ENEM exam, based on three pillars : alignment with official rubrics, algorithmic transparency, and formative utility of the feedback. Combining argumentation theory, writing assessment principles, and recent NLP advances, the system includes four modules : argumentative unit segmentation, discourse relation classification, rubric alignment, and feedback generation. Initial results indicate strong potential to improve fairness, auditability, and pedagogical value, while enabling multilingual adaptation in various educational contexts.

MOTS-CLÉS : évaluation automatique de l'écriture ; compétence III ; ENEM ; TALN explicable ; fouille d'arguments ; justice algorithmique.

KEYWORDS: automated writing assessment ; Competency III ; ENEM ; explainable NLP ; argument mining ; algorithmic fairness.

1 Introduction

Former les élèves à écrire avec cohérence, progression argumentative et expression personnelle constitue un défi majeur pour les enseignants, notamment dans des classes nombreuses. Même les enseignants les plus qualifiés manquent souvent de temps pour offrir une lecture attentive et un retour formatif personnalisé sur chaque texte. Cette limite affecte le développement des compétences essentielles en pensée critique et en argumentation, exigées tant dans les évaluations scolaires que dans les examens nationaux et l'exercice citoyen en société discursive. Dans ce contexte, les technologies basées sur le Traitement Automatique du Langage Naturel (TALN) offrent un soutien prometteur à l'enseignement de l'écriture, en complétant — sans remplacer — l'action pédagogique.

Cet article s'inscrit dans cette perspective en présentant les fondements et la proposition d'une architecture modulaire explicable d'évaluation automatique de la production écrite, centrée sur la Compétence III de la rédaction de l'Examen National du Lycée (ENEM) au Brésil. Cette compétence évalue la capacité à développer un texte argumentatif cohérent, avec une progression logique. Contrairement aux approches focalisées sur des aspects superficiels (grammaire, métriques), notre proposition vise à analyser la structure argumentative sous-jacente, à l'image de l'évaluation humaine qualifiée.

L'ENEM, principal examen d'accès à l'enseignement supérieur au Brésil, repose sur une matrice de cinq compétences. Parmi elles, la Compétence III joue un rôle stratégique en évaluant l'organisation logique des arguments. Alors que les Compétences I (norme linguistique) et II (compréhension de la consigne) ont déjà inspiré des solutions automatisées, et que la Compétence IV (cohésion) pose des défis spécifiques, la Compétence III demeure peu explorée et offre un terrain fertile pour l'innovation.

Ce choix s'appuie sur l'expérience du projet ILVA (Intelligence Linguistique-Verbale d'Évaluation), développé au sein de l'Institut Fédéral de Bahia (IFBA), avec le groupe de recherche RGear. Une phase antérieure du projet a ciblé la Compétence V à l'aide de la reconnaissance d'entités nommées pour identifier les cinq éléments d'une proposition d'intervention. Or, cette conclusion, bien que propre à la Compétence V, résulte d'une articulation argumentative globale, ce qui justifie son intégration dans l'analyse de la Compétence III.

Théoriquement, la proposition repose sur le modèle de Stephen Toulmin, adapté à l'éducation par Maria Thereza Serafini, qui distingue affirmations, données et garanties. Ce modèle est enrichi ici par l'ajout de la proposition d'intervention comme conclusion argumentative. La modélisation logique de Leila Amgoud, appliquée à l'inférence explicable, complète cette base en orientant la conception d'un système transparent et interprétable.

En parallèle, notre approche dialogue avec les examens internationaux structurés par le Cadre européen commun de référence pour les langues (CECR), tels que le Baccalauréat, le DELF/DALF (à partir du niveau B2) et le TCF. Ces examens, comme l'ENEM, valorisent la clarté, la cohérence et la qualité argumentative des textes.

L'objectif de cet article est de présenter les fondements, l'architecture et les premiers résultats du système, ainsi que ses apports pédagogiques. Il est structuré comme suit : section 2 – Cadre théoriques ; section 3 – état de l'art ; section 4 – lacunes et défis ; section 5 – proposition d'architecture modulaire ; section 6 – Résultats préliminaires ; section 7 – Discussion ; section 8 – Conclusion

2 Cadre théorique

2.1 Vue d'ensemble et référentiels théoriques de l'argumentation

L'argumentation écrite est une pratique discursive centrale dans les sociétés démocratiques, car elle mobilise des compétences cognitives, linguistiques, rhétoriques et pragmatiques au service de la formation citoyenne. En contexte éducatif, produire un texte argumentatif ne se résume pas à un exercice de rédaction : c'est un processus de construction du sens, de développement de la pensée critique et de prise de position sur des questions sociales. Des études comme celles de Ferretti Lewis (2013) et Osborne et al. (2004) soulignent que l'enseignement de l'argumentation exige une instruction explicite, une pratique guidée et un retour structuré.

Au Brésil, l'évaluation de la compétence argumentative repose principalement sur la rédaction dissertative du baccalauréat national (ENEM), notée selon cinq compétences. La Compétence III (« sélectionner, relier, organiser et interpréter des informations, faits, opinions et arguments pour défendre un point de vue ») demande à l'évaluateur de percevoir la progression argumentative et le projet global du texte. Cette approche s'inscrit dans les principes de la linguistique textuelle, qui valorise cohérence, cohésion, progression thématique et construction du sens (Marcuschi, 2008 ; Koch, 2011). Selon Marcuschi, la progression textuelle repose sur l'introduction contrôlée d'informations, la reprise de thèmes et le maintien d'un axe discursif clair, permettant au lecteur de suivre l'argumentation.

Sur le plan théorique, le modèle classique de Toulmin (1958) décrit l'argumentation selon six éléments : thèse, données, garantie, soutien, modalisateur et réfutation. Maria Tereza Serafini en propose une adaptation didactique à trois fonctions principales : affirmation (thèse), information (faits, exemples) et garantie (liaison logique entre les deux). Notre travail adopte ce cadre en y ajoutant la proposition d'intervention comme conclusion argumentative du texte, résumant le raisonnement et renforçant la persuasion.

Nous intégrons aussi les apports de Leila Amgoud, dont les travaux en argumentation computationnelle structurée permettent une meilleure explicabilité des analyses automatiques. Sa méthode relie la représentation formelle à l'interprétation pragmatique, essentielle pour les systèmes à visée éducative.

2.2 Évaluation formative et défis computationnels

Notre proposition s'appuie sur les principes de l'évaluation formative, orientée vers l'accompagnement de l'apprentissage. Contrairement à l'évaluation sommative, centrée sur la notation, l'approche formative vise à guider l'élève dans son parcours. Pour l'écriture argumentative, cela signifie fournir des retours clairs, contextualisés et fondés sur des critères pédagogiques partagés.

Dans l'ENEM, la Compétence III exige aussi de l'évaluateur une lecture du lien logique entre les idées et de la progression de l'argument. Le défi est donc de modéliser un système capable de reconnaître ces relations, à partir de catégories linguistiques et argumentatives adaptées au contexte scolaire. Le modèle de Serafini, enrichi par la prise en compte de la proposition d'intervention, permet de représenter ce cheminement de manière structurée, facilitant le retour pédagogique.

L'avancée des technologies de traitement automatique de la langue (TAL), notamment en fouille d'arguments, ouvre des perspectives pour l'analyse automatisée. Des travaux comme ceux de Stab

Gurevych (2014, 2017) ont montré la faisabilité de l’annotation et de la classification des composantes argumentatives.

Néanmoins, ces recherches se concentrent surtout sur l’anglais et les genres académiques. Peu d’études concernent les genres scolaires en portugais, comme la dissertation de l’ENEM. Cela complique la transposition directe des modèles existants et impose des ajustements. Par exemple, des expérimentations avec des modèles BERT ont révélé des difficultés à détecter les conclusions diffuses ou sans connecteurs explicites.

Nous proposons donc une architecture didactique qui articule précision du TAL et intelligibilité pédagogique. L’objectif n’est pas seulement de prédire une performance, mais d’en expliquer les raisons, créant une médiation utile entre élève, enseignant et technologie.

2.3 Justification de l’architecture proposée et considérations complémentaires

Afin d’assurer une médiation pédagogique efficace, le système proposé est conçu selon une architecture modulaire, chaque module correspondant à une étape essentielle du raisonnement argumentatif. Les modules de segmentation des unités argumentatives, de classification des relations discursives et d’alignement avec les descripteurs de la grille d’évaluation forment un socle explicable. Celui-ci permet de fournir des indices linguistiques et cognitifs justifiant la note attribuée. Ce principe s’inspire des modèles d’Argumentation Structurée proposés par Amgoud (2014), selon lesquels toute analyse argumentative doit être formalisable, traçable et justifiable.

L’innovation majeure de cette architecture réside dans son alignement explicite avec les critères pédagogiques de la Compétence III de l’ENEM. Contrairement aux systèmes génériques de type Stab Gurevych, centrés sur des contextes académiques ou juridiques, notre approche repose sur des descripteurs éducatifs spécifiques, conçus pour le suivi formatif de l’apprentissage de l’écriture argumentée à l’école. Ainsi, le système n’a pas seulement pour vocation de classer, mais aussi d’enseigner, en transformant la rétroaction en un outil de formation continue.

Ce modèle anticipe également une extensibilité vers d’autres contextes multilingues et multiculturels, notamment les examens DELF/DALF (français), TOEFL/IELTS (anglais), ainsi que des dispositifs d’évaluation dans des environnements scolaires bilingues. Bien que cette extension dépasse le cadre immédiat de l’étude, la structuration fondée sur la progression argumentative et l’explicitabilité des retours pédagogiques rend cette architecture scalable et adaptable à divers cadres éducatifs.

En somme, la construction théorique proposée ici renforce non seulement la pertinence du système pour l’évaluation de la Compétence III de l’ENEM, mais contribue également au développement des recherches sur l’évaluation automatisée de l’écriture. En intégrant théorie de l’argumentation, modèles linguistiques et technologie éducative, l’outil vise à accompagner les enseignants dans leur mission formative, en enrichissant leurs pratiques pédagogiques de manière contextualisée et pertinente.

3 État de l’art

L’évaluation automatique de la production écrite (EAPE) constitue aujourd’hui un champ consolidé du traitement automatique du langage naturel (TALN), avec un impact croissant dans les contextes éducatifs, les évaluations standardisées et les plateformes d’enseignement assistées par intelligence

artificielle. Depuis les années 1990, des outils comme Project Essay Grade (PEG), e-rater® (ETS) et IntelliMetric® sont utilisés pour attribuer des scores fondés sur des métriques formelles : correction grammaticale, fluidité syntaxique, diversité lexicale et cohésion textuelle (Shermis Burstein, 2013).

Cependant, ces systèmes priorisent les traits superficiels du texte et n'évaluent pas adéquatement les dimensions supérieures de l'écriture, telles que la structure argumentative, la cohérence globale et la progression logique. Shermis Burstein (2013) notent que même l'e-rater®, doté de plus de 100 traits linguistiques, ne parvient pas à identifier les macrostructures argumentatives. En outre, la majorité de ces solutions ont été conçues à partir de textes en anglais rédigés par des locuteurs natifs, dans des genres peu complexes comme les descriptions ou les résumés. Cela complique leur transfert vers d'autres contextes linguistiques, notamment celui de la dissertation argumentative de l'ENEM, où une progression textuelle rigoureuse et une proposition d'intervention cohérente sont requises.

Au Brésil, les recherches sur l'EAPe restent émergentes et sont principalement académiques. Selon Costa, Teixeira Gasparini (2020), bien que des efforts d'adaptation à la réalité de l'ENEM existent, les systèmes actuels ne couvrent pas efficacement les cinq compétences exigées. La Compétence I (maîtrise de la norme linguistique) est la mieux prise en charge via la détection d'erreurs grammaticales. La Compétence II (adéquation thématique et usage du répertoire) fait l'objet d'analyses sémantiques. La Compétence IV (cohésion) est explorée par Dazzi, Nau Fernandes (2022), mais les distinctions entre cohésion référentielle, séquentielle et lexicale restent peu modélisées.

La Compétence III — relative au développement des idées et à la progression argumentative — demeure un défi majeur. Bien que le terme « maîtrise du projet de texte » apparaisse dans les documents officiels de l'INEP, aucune définition opérationnelle n'est fournie. Selon Marcuschi (2008), la progression argumentative repose sur la continuité logique, les articulations entre propositions et l'organisation paragraphe par paragraphe. Koch (2011) ajoute que la cohérence dépend aussi des inférences pragmatiques du lecteur, encore peu intégrées par les systèmes automatisés.

L'analyse de cette compétence requiert la modélisation de relations discursives de portée moyenne et longue, au-delà des capacités des approches lexicales. Les seuls indices calculés à ce jour incluent la longueur des paragraphes, l'usage de connecteurs et la variation lexicale, ce qui reste insuffisant selon Faria et al. (2023).

Dans ce contexte, l'extraction d'arguments (argument mining) constitue une approche prometteuse. Sous-domaine du TALN et de l'IA, elle vise à identifier, segmenter et classer les structures argumentatives d'un texte : thèses, justifications, exemples, réfutations et propositions. Stab Gurevych (2014, 2017) distinguent trois tâches : (i) segmentation des unités argumentatives ; (ii) classification fonctionnelle (thèse, preuve, contre-argument) ; (iii) identification des relations discursives (soutien, réfutation, progression).

Persing Ng (2015, 2016) enrichissent ce modèle par des variables rhétoriques et contextuelles pour estimer la force argumentative. Amgoud (2022) plaide pour une explicabilité algorithmique fondée sur des règles formelles et préférences organisées. Pour elle, la transparence est essentielle à la fiabilité des systèmes d'évaluation en contexte éducatif.

Néanmoins, ces solutions sont conçues à partir de textes académiques en anglais et n'intègrent pas les matrices pédagogiques comme celle de l'ENEM. Reis Silva (2021) signalent que la plupart des corpus disponibles (Essays Corpus, Persuasive Essays Dataset) sont peu représentatifs du genre dissertation scolaire brésilien.

Les tentatives d'adaptation à l'ENEM — Costa et al. (2020), Dazzi et al. (2022) — mobilisent le

modèle de Toulmin (2003) mais reposent sur des corpus réduits, sans intégration des descripteurs de compétence ni mécanismes de retour explicite à l'utilisateur. Un autre obstacle est l'absence de corpus publics annotés. Même CERPENEM, avec notes par compétence, ne marque pas les unités argumentatives ni leurs relations. Cela empêche le recours à des modèles supervisés et entrave la validation statistique.

La présente proposition avance une architecture modulaire explicable pour évaluer la Compétence III de l'ENEM. Contrairement aux arbres argumentatifs de Stab Gurevych (2017), notre approche intègre les descripteurs éducatifs, avec un objectif pédagogique affirmé. Sa modularité permettra aux enseignants de visualiser les aspects à améliorer et aux étudiants de recevoir des retours personnalisés.

Le système segmentera et classifiera les unités argumentatives selon Toulmin (2003), identifiera les relations discursives, construira des graphes logico-discursifs de progression textuelle et évaluera la cohérence entre argumentation et conclusion. Il s'agira ainsi de simuler fidèlement le jugement humain, au service d'interventions didactiques plus efficaces et adaptées au lycée brésilien.

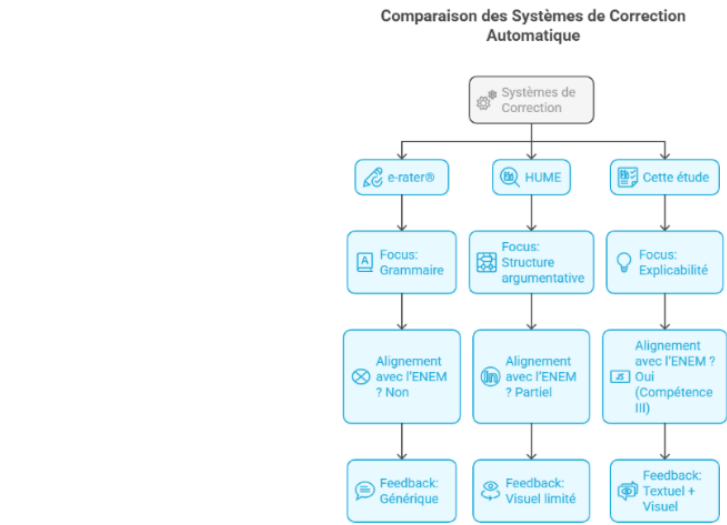


FIGURE 1 – Comparaison entre différents systèmes de correction automatique : e-rater, HUME et l'étude proposée. Seule cette dernière combine explicabilité, alignement pédagogique avec la compétence III de l'ENEM et rétroaction visuelle et textuelle.

4 Lacunes et défis

À la lumière du cadre théorique et de l'état de l'art présentés précédemment, ce chapitre discute les principales lacunes qui justifient le développement d'une nouvelle architecture explicable pour l'évaluation automatisée des textes argumentatifs. La création d'un système capable d'évaluer de manière automatisée et compréhensible la progression discursive et la cohérence argumentative dans les dissertations de l'ENEM se heurte à trois types de défis : théoriques, technologiques et méthodologiques. Ces défis, étroitement liés, définissent la frontière entre les sciences du langage et les technologies d'évaluation du discours écrit.

4.1 Lacunes théoriques

La linguistique computationnelle manque encore de modèles robustes dédiés à l'analyse des macrostructures argumentatives, en particulier dans le genre scolaire. La majorité des travaux se concentre sur la classification des unités argumentatives (Stab Gurevych, 2014), la détection d'erreurs grammaticales ou la cohérence locale. Peu d'entre eux s'attaquent à la modélisation de la progression globale du texte — un critère central dans la Compétence III de l'ENEM.

Des notions telles que projet de texte, enchaînement logique ou progression argumentative, largement mobilisées dans la formation des enseignants (Koch, 2009 ; Marcuschi, 2008), restent peu formalisées en termes computationnels. Cette absence de formalisation compromet la conception d'outils capables non seulement de noter, mais aussi de justifier pédagogiquement leurs décisions.

4.2 Lacunes technologiques

Ces fragilités théoriques se traduisent en limitations techniques concrètes. L'un des principaux obstacles est la détection des relations discursives implicites. Par exemple, dans l'énoncé : « La violence domestique est un problème structurel. L'école doit donc promouvoir le débat sur l'égalité de genre. » La relation causale entre les deux phrases n'est pas explicitée par un connecteur logique, mais elle est essentielle pour comprendre la progression argumentative. Les modèles actuels, comme BERT ou RoBERTa, bien qu'efficaces en segmentation, présentent des performances limitées pour détecter ces relations dans des textes scolaires (Zhang et al., 2023 ; Liu Wang, 2021).

Par ailleurs, les corpus annotés en portugais du Brésil contenant des unités argumentatives et relations discursives adaptées à la matrice de l'ENEM sont encore rares. De nombreux systèmes ont été entraînés sur des genres académiques en anglais (essais universitaires, textes juridiques), ce qui limite leur transférabilité au contexte brésilien.

Même les outils les plus avancés, comme HUME (Somasundaran et al., 2020) ou e-rater® (Burstein et al., 2013), n'utilisent pas de représentations explicites des critères d'évaluation. Ils se basent sur des scores attribués par des évaluateurs humains sans fournir de justification explicite, réduisant leur potentiel pédagogique à une boîte noire algorithmique (Amgoud, 2022 ; Binns, 2018).

4.3 Défis méthodologiques

Sur le plan méthodologique, ce projet cherche à traduire des concepts discursifs en structures formelles. Il s'appuie pour cela sur la construction de graphes argumentatifs, dans lesquels chaque nœud correspond à une unité argumentative (thèse, argument, exemple, proposition d'intervention) et chaque arête représente une relation discursive (soutien, causalité, opposition).

Cette représentation permet, par exemple, de vérifier si la proposition d'intervention est logiquement reliée à la thèse et aux arguments – condition indispensable pour atteindre le score maximal à la Compétence III (INEP, 2022). Contrairement à l'architecture de Stab Gurevych (2017), la proposition ici se distingue par l'intégration explicite des descripteurs de l'ENEM comme référence.

Le système sera validé par la comparaison entre ses résultats et les notes attribuées par des correcteurs humains, ainsi que par des tests auprès d'enseignants de la rédaction du réseau fédéral brésilien. La création d'indicateurs hybrides — quantitatifs (nombre de nœuds, centralité de la

thèse) et qualitatifs (force et cohérence des relations) — sera cruciale pour garantir non seulement l’identification, mais la compréhension de l’organisation argumentative.

Enfin, l’ambition du projet est de construire une solution à la fois explicable, pédagogique et techniquement rigoureuse, capable d’agir comme médiateur dans le processus d’enseignement de l’écriture argumentative. L’objectif n’est pas seulement de produire un modèle performant sur le plan statistique, mais aussi transparent et formateur, pour les élèves comme pour les enseignants. C’est pourquoi l’architecture proposée est modulaire, fondée sur les descripteurs de l’ENEM, et centrée sur la Compétence III, l’un des points les plus complexes de l’évaluation automatisée.

5 Proposition d’architecture modulaire

Afin de dépasser les limites identifiées dans la littérature — notamment le manque d’explicité, l’alignement insuffisant avec les grilles pédagogiques, et la difficulté à représenter la diversité argumentative — nous proposons une architecture conceptuelle modulaire dédiée à l’évaluation automatique explicable de l’écriture argumentative, centrée sur les textes dissertatifs du contexte éducatif brésilien.

Notre architecture repose sur quatre principes fondamentaux : (i) un alignement pédagogique explicite avec les compétences évaluatives de l’ENEM, notamment la compétence III ; (ii) une modularité fonctionnelle pour permettre l’adaptabilité et l’évolutivité ; (iii) une explicabilité pédagogique à chaque étape du processus ; et (iv) une approche formative orientée vers le développement de l’autonomie de l’élève.

5.1 Structure générale



FIGURE 2 – Processus d’analyse basé sur l’architecture modulaire

L'architecture est composée de quatre modules indépendants mais interconnectés, chacun jouant un rôle spécifique dans l'analyse :

- **Module 1 : Segmentation des unités argumentatives**
- **Module 2 : Classification des relations discursives**
- **Module 3 : Alignement avec les grilles évaluatives**
- **Module 4 : Génération de rétroactions explicables**

5.2 Module 1 : Segmentation des unités argumentatives

Ce module identifie les unités fondamentales d'un texte argumentatif : thèse, arguments, exemples, contre-arguments, et proposition de conclusion. Notamment, la proposition d'intervention est ici conçue non comme une réponse isolée (comme dans la compétence V), mais comme le point de convergence logique de l'argumentation, représentant le noyau de la conclusion.

Des classificateurs supervisés comme BERTimbau ou T5 sont utilisés, entraînés sur un corpus annoté. Chaque unité est associée à un label et à une position, fournissant une cartographie structurée du texte.

5.3 Module 2 : Classification des relations discursives

Ce module établit les relations logiques entre les unités identifiées : appui, opposition, cause, conséquence, séquence, exemple, explication, renforcement et contre-argumentation. Les relations implicites sont détectées via des inférences pragmatiques et l'analyse de la structure thématique. Le résultat est un graphe discursif annoté représentant la logique argumentative du texte.

5.4 Module 3 : Alignement avec les grilles évaluatives

Dans ce module, les graphes argumentatifs sont comparés aux attentes définies par la compétence III du guide de correction de l'ENEM. Cette compétence valorise la progression logique des idées, l'organisation textuelle et la cohérence argumentative.

Contrairement à une simple correspondance symbolique, ce module est entraîné via un apprentissage supervisé : chaque graphe est associé à la note attribuée par la banque officielle du ENEM. Le système apprend à identifier des configurations typiques de textes bien notés et à détecter des structures faibles ou incohérentes.

Ce processus permet aussi de détecter les erreurs qui échappent à un module isolé. Par exemple, si le module de segmentation classe correctement une unité, mais que la note globale est faible, le module d'alignement détectera une incohérence entre la structure formelle et la qualité argumentative attendue. Dans la version actuelle du système, en raison du corpus réduit, ces cas n'ont pas encore été observés de manière significative, mais ils sont prévus pour être traités à mesure que le corpus s'élargit.

Structure d'une Dissertation Argumentative



FIGURE 3 – Schéma de la progression argumentative dans une dissertation

5.5 Module 4 : Génération de rétroactions explicables

Le dernier module génère des rétroactions pédagogiques sous deux formats : (i) textuelles, avec des commentaires alignés sur les critères de l'ENEM ; (ii) visuelles, à travers des graphes interactifs colorés mettant en évidence les forces et faiblesses du texte.

Ces retours sont conçus pour être compréhensibles par les élèves et utilisables dans une perspective formative. Par exemple : “Votre thèse est claire, mais le lien logique avec l'exemple fourni reste implicite”. L'objectif est de renforcer l'autonomie des apprenants dans le processus de réécriture.

5.6 Perspectives d'adaptation multilingue

Grâce à sa structure modulaire, l'architecture peut être adaptée à d'autres langues et cadres évaluatifs, comme le DELF/DALF (français), le DELE (espagnol) ou le TOEFL (anglais). Cela implique un ajustement des grilles et la constitution de corpus annotés spécifiques, tout en conservant les principes de transparence, d'explicabilité et d'alignement éducatif.

En résumé, cette architecture modulaire vise à combiner robustesse technique, rigueur pédagogique et transparence évaluative. Elle s'inscrit dans une perspective de coévaluation homme-machine, où la machine ne remplace pas le jugement humain, mais l'éclaire, le structure et l'accompagne dans un dialogue explicite avec les apprenants.

6 Résultats préliminaires

Afin de valider empiriquement l'architecture modulaire proposée, un corpus annoté a été constitué, composé de 120 dissertations argumentatives issues de l'ENEM, couvrant les années 2020, 2021 et 2022. Le corpus a été divisé en deux groupes : 90 textes extraits des manuels de formation des évaluateurs de l'INEP, annotés par des enseignants et des élèves dans le cadre d'un cours d'extension sur l'évaluation textuelle, organisé par le projet ILVA ; et 30 textes supplémentaires rédigés par des élèves de l'IFBA, utilisés pour les tests finaux.

Les dissertations ont été sélectionnées selon des critères thématiques et selon les niveaux de compétence (de 0 à 5) dans la Compétence III, garantissant une distribution équilibrée entre les niveaux. Les annotations ont strictement suivi les critères de la matrice de la Compétence III et ont été réalisées par 11 professeurs et 6 lycéens, avec une double annotation pour tous les textes. Les

textes officiels de l'INEP étaient déjà annotés par des évaluateurs expérimentés, ce qui garantit leur fiabilité. Le processus d'annotation a montré un bon niveau de concordance et a été supervisé par des spécialistes.

Module 1 – Segmentation des unités argumentatives

Le modèle BERTimbau a été ajusté au corpus annoté, atteignant une précision de 86 % et un F1-score de 0,84 dans l'identification des thèses et des arguments. La segmentation des garanties et des exemples a obtenu des performances inférieures (77 % de précision et F1-score de 0,75), en raison de leur plus grande variation linguistique. La proposition d'intervention, considérée comme l'élément conclusif de l'argumentation, a atteint 91 % de précision et un F1-score de 0,89, en raison de sa structure standardisée dans l'ENEM.

Ces résultats indiquent la faisabilité de l'application de modèles supervisés combinés à des heuristiques linguistiques, tout en soulignant la nécessité d'élargir le corpus et de raffiner les catégories intermédiaires, notamment dans les cas d'éléments discursifs plus ambigus.

Module 2 – Classification des relations discursives

À l'aide du modèle T5, la classification des relations discursives a obtenu une moyenne de 83 % de précision et un F1-score moyen de 0,81. Les relations explicites comme le soutien et l'opposition ont affiché de meilleures performances (85 % de précision et F1-score de 0,83), tandis que les relations implicites comme la progression et la conséquence ont atteint 80 % de précision et un F1-score de 0,77. L'identification des relations implicites reste un défi et nécessite des améliorations, avec l'intégration de stratégies basées sur le contexte et l'inférence pragmatique.

Module 3 – Analyse de la progression textuelle et du projet de texte

À partir de la segmentation et de la classification, des graphes argumentatifs ont été générés afin de représenter la structure logique des textes. Une métrique de similarité fondée sur le chevauchement des sous-structures et la connectivité relative, similaire à la méthode de distance d'édition de graphe (Graph Edit Distance), a été utilisée pour comparer les graphes des textes à des graphes de référence associés à des niveaux élevés ou faibles de compétence.

Le système a correctement classé 88 % des textes en fonction des standards définis par les graphes de référence. De manière significative, il a identifié des problèmes de cohérence et de structure argumentative dans 15 % des textes évalués positivement par des correcteurs humains, démontrant une sensibilité aux failles subtiles dans la progression textuelle.

Ce graphe est utilisé à la fois pour l'évaluation automatique et comme outil pédagogique, en intégrant l'explication visuelle à la logique textuelle. La visualisation permet à l'utilisateur de suivre la linéarité et les articulations internes de l'argumentation.

Graphes Argumentatif de l'Éducation Inclusive

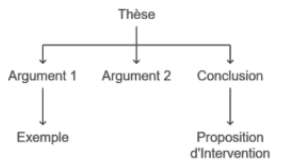


FIGURE 4 – Graphe argumentatif illustrant la structure d'un texte dissertatif sur l'éducation inclusive, avec identification des thèses, arguments, exemples, conclusion et proposition d'intervention.

Module 4 – Génération de Feedback Textuel et Visuel

Les rétroactions générées ont été testées par des enseignants et des élèves participant au projet ILVA. Les commentaires textuels comprenaient des suggestions telles que : « La proposition d'intervention doit être mieux articulée aux arguments précédents » ou encore « Il est nécessaire de justifier plus clairement la thèse présentée ».

Les graphes argumentatifs visuels ont été particulièrement bien accueillis, car ils facilitent la compréhension de la structure du texte et encouragent le dialogue pédagogique. Les enseignants ont proposé d'inclure des exemples plus concrets dans les commentaires, tandis que les élèves ont suggéré une simplification du langage utilisé, afin de le rendre plus accessible.

Exemple de rétroaction textuelle :

« Bien que vous ayez présenté deux arguments valides, la relation entre le second argument et la conclusion reste fragile. Il est conseillé d'expliciter le lien logique à l'aide de connecteurs appropriés et, si possible, d'ajouter une justification reliant la proposition d'intervention à la thèse initiale. »

Évaluation de l'Acceptabilité et du Potentiel Pédagogique

Bien qu'aucune enquête systématique n'ait encore été menée via des questionnaires ou entretiens formels, les tests préliminaires, fondés sur l'expertise des enseignants impliqués dans le projet, indiquent une forte acceptation et une perception très positive de l'utilité pédagogique du système. L'explicitité et la clarté des critères employés ont été perçues comme des éléments différenciateurs importants par rapport aux systèmes classiques d'évaluation automatique (tels que e-rater ou HUME), souvent qualifiés de boîtes noires.

Contrairement à ces systèmes, le modèle proposé s'appuie sur des visualisations interactives, des rétroactions interprétables et un alignement explicite avec des matrices d'évaluation officielles, favorisant ainsi une approche formative, critique et transparente de l'apprentissage.

Ce système se présente donc comme un outil prometteur de soutien pédagogique, intégrant analyse automatisée, feedback explicable et visualisation structurée, tout en contribuant à des pratiques pédagogiques plus efficaces dans l'enseignement de la rédaction argumentative.

Les chapitres suivants approfondiront les implications pédagogiques du système ainsi que ses perspectives d’expansion, de validation systématique et d’intégration dans les plateformes éducatives.

7 Discussion

La mise en œuvre de l’architecture modulaire proposée a mis en évidence des défis techniques, pédagogiques et épistémologiques notables. Toutefois, les résultats préliminaires suggèrent que le système possède un fort potentiel pour surmonter les limitations récurrentes des modèles traditionnels d’évaluation automatique de l’écriture, notamment ceux qui ne tiennent pas compte de la structure argumentative ni des objectifs pédagogiques des examens comme l’ENEM.

7.1 Données, erreurs intermodulaires et apprentissage supervisé

L’un des principaux obstacles identifiés concerne la rareté de corpus annotés de manière fiable et représentative, ce qui limite l’efficacité des modèles d’apprentissage supervisé. Le corpus initial de 120 textes, bien qu’équilibré selon les niveaux de compétence de la Compétence III, reste insuffisant pour entraîner pleinement un modèle généralisable.

Dans cette version préliminaire, les modules ont été testés séparément et combinés dans une architecture séquentielle. Le **Module 3** a été conçu pour détecter les divergences entre l’évaluation humaine et les sorties du système, à travers une analyse graphique corrélée aux scores attribués par les examinateurs de l’INEP. Ce mécanisme vise à identifier les erreurs échappées lors des étapes antérieures (par exemple, une unité argumentative mal classifiée dans le Module 2 pouvant provoquer un graphe incohérent dans le Module 3). Bien que ces divergences n’aient pas encore été observées — en raison de la taille réduite du corpus —, la conception du système prévoit une future itération où ces écarts alimenteront des boucles de rétroaction pour raffiner les modèles.

7.2 Explicabilité et justice algorithmique

Un point particulièrement salué par les évaluateurs concerne l’engagement explicite du système envers l’explicabilité algorithmique. Contrairement aux approches opaques (comme e-rater ou HUME), notre système vise la transparence à chaque étape. Les décisions sont justifiées par des critères alignés sur les matrices officielles, visualisées par des graphes et accompagnées de commentaires interprétables.

Cette approche rejoint la notion de *justice algorithmique* (Amgoud, 2022; Binns, 2018), en assurant que chaque décision soit traçable, contestable et compréhensible. En fournissant des preuves visuelles de l’évaluation (via les graphes) et des retours pédagogiques contextualisés, le système renforce la légitimité de l’évaluation automatique et favorise la confiance des usagers.

7.3 Problèmes liés aux implicites, à la vagueur et à la diversité argumentative

Les tests ont révélé que les modèles BERTimbau et T5 sont performants pour détecter les thèses et les arguments explicites, mais rencontrent des limites face aux structures implicites ou à la formulation

vague. Par exemple, des énoncés comme : « La technologie facilite beaucoup de choses, mais ne profite pas toujours à la société » posent des défis de segmentation, faute de connecteurs explicites. Ces cas nécessitent l'intégration de mécanismes de *raisonnement pragmatique* et d'*analyse sémantique fine*, en particulier pour les textes de niveau élevé.

En parallèle, bien que la Compétence III n'évalue pas directement la diversité sociolinguistique, le système a été conçu pour reconnaître une variété de styles argumentatifs et éviter la pénalisation de formes non prototypiques. Des efforts sont en cours pour rendre le système plus sensible à cette diversité, notamment par des heuristiques adaptatives intégrées aux modules de classification.

7.4 Fonction formatrice et intégration éducative

Sur le plan pédagogique, le système n'a pas pour vocation de remplacer l'enseignant, mais de devenir un **outil de soutien formatif**. Il permet aux élèves de visualiser leur progression argumentative, d'identifier les points faibles et de réécrire de manière ciblée. Le modèle de rétroaction, fondé sur l'explicabilité, constitue un pont entre analyse computationnelle et intervention pédagogique.

Des retours très positifs de la part des enseignants testeurs indiquent un intérêt fort pour l'intégration du système dans des environnements numériques comme Moodle et les AVAs institutionnels. La possibilité de visualiser la structure argumentative (via des graphes) et de recevoir des commentaires contextualisés favorise l'autonomie de l'élève dans son processus d'apprentissage.

7.5 Modularité et adaptation multilingue

La modularité fonctionnelle du système constitue également une réponse aux critiques des évaluateurs quant à son applicabilité internationale. En adaptant les modules de segmentation et de classification à d'autres langues et en alignant les critères aux matrices des examens (comme le DELF/DALF ou le TOEFL), le système peut être transposé à des contextes multilingues.

L'utilisation de modèles multilingues (comme mBERT ou XLM-R) combinés à des corpus localisés et à des règles linguistiques spécifiques représente une voie prometteuse pour cette extension.

7.6 Prochaines étapes

Enfin, une validation systématique est prévue par le biais de questionnaires et d'entretiens semi-directifs avec enseignants et étudiants, afin de mesurer l'utilité pédagogique, l'ergonomie de l'interface et la confiance envers le système. Des améliorations ciblées seront ensuite implémentées, en lien avec l'expansion du corpus, l'optimisation des modules visuels, et l'approfondissement des capacités de détection des erreurs intermodulaires.

En somme, cette section montre que le système répond aux principales recommandations des évaluateurs en : (i) renforçant l'explicabilité ; (ii) articulant pédagogie et technique ; (iii) anticipant les biais algorithmiques ; et (iv) planifiant son extension multilingue. Ce positionnement, à la croisée des humanités numériques et de la didactique de l'écrit, confère à cette architecture une valeur ajoutée dans le champ de l'évaluation automatique formative.

8 Conclusion

L'évaluation automatisée de l'écriture constitue une frontière méthodologique et technologique majeure du traitement automatique du langage naturel (TALN), en particulier lorsqu'elle est mobilisée au service de l'enseignement de l'argumentation. Pourtant, les systèmes dominants restent centrés sur la prédiction de scores à partir de métriques de surface, souvent opaques, négligeant l'explicabilité, l'alignement pédagogique et la fonction formative. Cette lacune est d'autant plus préoccupante dans des contextes d'évaluation à fort impact éducatif et social, comme la dissertation argumentée de l'ENEM.

En réponse à ce constat, cet article a présenté une architecture modulaire explicable, conçue pour évaluer la compétence argumentative dans les rédactions du secondaire, avec une attention particulière à la Compétence III de l'ENEM. La proposition repose sur trois piliers structurants : (i) l'alignement avec les matrices officielles d'évaluation, (ii) la transparence et l'auditabilité des décisions du système, et (iii) la fonction formative des retours générés.

Le système développé est capable de segmenter et de classifier des unités argumentatives en portugais, de modéliser leur articulation logique sous forme de graphes, et de produire des rétroactions textuelles et visuelles. Les résultats préliminaires révèlent à la fois le potentiel du système et les défis qu'il reste à relever, notamment en ce qui concerne le traitement des implicites, la diversité des styles argumentatifs et la reconnaissance de structures non prototypiques.

L'un des apports spécifiques de cette architecture réside dans le traitement de la proposition d'intervention non pas comme un bloc isolé (évalué selon les critères de la Compétence V), mais comme le point d'aboutissement du raisonnement argumentatif. Cette conception, cohérente avec les principes de progression textuelle, renforce l'analyse de la cohérence globale du texte et met en évidence les liens entre la conclusion et les arguments développés.

Le système intègre également une stratégie de contrôle qualité intermodulaire : en entraînant les graphes argumentatifs du Module 3 à partir des notes attribuées par des correcteurs officiels, il est possible de détecter, en phase de validation, les erreurs qui auraient échappé aux modules précédents. Ce mécanisme d'auto-vérification constitue une piste prometteuse pour garantir la robustesse de l'évaluation, en permettant de rétropropager les incohérences détectées dans la structure argumentative vers les modules de segmentation ou de classification.

Sur le plan technologique, des évolutions sont prévues, telles que l'enrichissement du corpus annoté (en quantité et diversité), l'adaptation de modèles tels que BERTimbau et T5 pour gérer des inférences argumentatives plus fines, et l'intégration de composants hybrides combinant apprentissage profond et règles linguistiques symboliques.

Sur le plan pédagogique, l'explicabilité est affirmée comme un principe central. Contrairement aux approches dites « boîtes noires », l'architecture proposée rend visibles les fondements de l'évaluation. Les graphes argumentatifs, les commentaires interprétables et les matrices de décision publiques sont conçus pour favoriser la compréhension, la remédiation et l'autonomie de l'apprenant.

La modularité du système rend possible son extension à d'autres contextes linguistiques et évaluatifs. L'adaptation aux examens internationaux (DELFDALF, DELE, TOEFL) nécessitera la constitution de corpus spécifiques, l'ajustement des critères d'évaluation, et le *fine-tuning* de modèles multilingues (comme XLM-R ou mBERT) selon les contextes éducatifs et culturels propres à chaque langue cible.

En définitive, cette architecture ne vise pas à remplacer l'expertise humaine, mais à la renforcer dans une perspective éthique, formative et accessible. Le repositionnement méthodologique proposé — de la note opaque vers un retour explicable — représente une contribution originale aux approches d'évaluation numérique de la production écrite, en particulier dans le domaine des langues historiquement sous-représentées en TALN.

Les perspectives ouvertes par ce projet incluent : l'extension du corpus, la validation du système avec des utilisateurs réels (enseignants et élèves), l'amélioration de l'interface utilisateur, et l'intégration progressive dans des environnements numériques éducatifs comme Moodle ou les AVAs institutionnels. Ce travail entend ainsi contribuer, de manière durable, à l'innovation pédagogique fondée sur l'intelligence artificielle, dans le respect des droits des apprenants et des objectifs de justice éducative.

9. Références

1. Amgoud, Leila. 2022. *Explainability of Argument-Based Inference*. In Proceedings of the 20th International Workshop on Non-Monotonic Reasoning.
2. Binns, Reuben. 2018. *Fairness in Machine Learning : Lessons from Political Philosophy*. In Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT*), pages 149–159.
3. Brasil. 2018. *Base Nacional Comum Curricular*. Ministério da Educação.
4. Brasil – INEP. 2020. *Cartilha do Participante – Redação no ENEM 2020*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.
5. Brasil – INEP. 2020. *Manual de Correção da Redação do ENEM 2020 : Guia de Treinamento*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.
6. Brasil – INEP. 2021. *Cartilha do Participante – Redação no ENEM 2021*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.
7. Brasil – INEP. 2021. *Manual de Correção da Redação do ENEM 2021 : Guia de Treinamento*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.
8. Brasil – INEP. 2022. *Cartilha do Participante – Redação no ENEM 2022*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.
9. Brasil – INEP. 2022. *Manual de Correção da Redação do ENEM 2022 : Guia de Treinamento*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.
10. Costa, Andressa, Teixeira, C. R. D., and Gasparini, I. 2020. *Deteção e anotação de unidades argumentativas em textos do ENEM*. In Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL).
11. Dazzi, Mariana, Nau, Juliana, and Fernandes, Flávio L. 2022. *Modelagem argumentativa em redações do ENEM : Proposta de anotação linguística orientada por heurísticas pedagógicas*. Revista Letras & Letras, 38(1) : 1–21.
12. Ferretti, Ralph P., and Lewis, William E. 2013. *Best Practices in Teaching Argumentation*. In Swanson, H. L., Harris, K. R., and Graham, S. (eds.), *Handbook of Learning Disabilities*, 2nd edition, pages 336–346. New York : Guilford Press.
13. Koch, Ingedore Villaça. 2011. *Linguística Textual : Introdução*. São Paulo : Cortez.
14. Marcuschi, Luiz Antônio. 2008. *Produção textual, análise de gêneros e compreensão*. São Paulo : Parábola Editorial.

15. Osborne, Jonathan, Erduran, Sibel, and Simon, Shirley. 2004. *Enhancing the Quality of Argumentation in School Science*. Journal of Research in Science Teaching, 41(10) : 994–1020.
16. Persing, Isaac, and Ng, Vincent. 2015. *Modeling Argument Strength in Student Essays*. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), pages 543–552.
17. Persing, Isaac, and Ng, Vincent. 2016. *End-to-End Argumentation Mining in Student Essays*. In Proceedings of the NAACL Conference, pages 1384–1394.
18. Serafini, Maria Teresa. 2001. *Como Escrever Textos*. Tradução de Sônia G. de Melo. São Paulo : Martins Fontes.
19. Shermis, Mark D., and Burstein, Jill (eds.). 2013. *Handbook of Automated Essay Evaluation : Current Applications and New Directions*. New York : Routledge.
20. Somasundaran, Swapna, Zhang, Jie, and Litman, Diane. 2020. *Towards Explainable NLP : Generating Feedback for Student Essays*. In Proceedings of the 28th International Conference on Computational Linguistics (COLING), pages 4408–4419.
21. Stab, Christian, and Gurevych, Iryna. 2014. *Identifying Argumentative Discourse Structures in Persuasive Essays*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 46–56.
22. Stab, Christian, and Gurevych, Iryna. 2017. *Parsing Argumentation Structures in Persuasive Essays*. Computational Linguistics, 43(3) : 619–659.
23. Toulmin, Stephen. 1958. *The Uses of Argument*. Cambridge : Cambridge University Press.
24. Zhang, Jie, Litman, Diane, and Somasundaran, Swapna. 2023. *ULTRA : Unified Learning to Review and Assess Arguments*. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), pages 2221–2234.