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

Vers l'évaluation automatique de résumés oraux d'enfants : étude exploratoire de métriques de compréhension

Agathe Wallet [1] Brice Brossette[2] Lucile Gelin[3, 4]
Stéphane Huet [1] Nathalie Camelin [1]
(1) LIA, Avignon Université, Avignon, France
(2) Laboratoire d'Etude des Mécanismes Cognitifs (EMC), Université Lumière Lyon 2, Lyon, France
(3) Lalilo by Renaissance Learning, France
(4) IRIT, Université Paul Sabatier, CNRS, Toulouse, France
prenom.nom@univ-avignon.fr, brice.brossette@univ-lyon2.fr,
lucile.gelin@renaissance.com

RÉSUMÉ

Apprendre à lire nécessite le développement de stratégies de compréhension, dont l'évaluation est un enjeu pédagogique majeur. Le projet CHICA-AI propose une plateforme numérique d'entraînement au résumé oral destinée aux élèves de CM1-CM2, intégrant un outil d'évaluation automatique. Cet article étudie la pertinence de l'application de métriques standard, initialement conçues pour l'écrit, à des résumés oraux d'enfants. Après une présentation de la plateforme, nous détaillons les caractéristiques principales du corpus collecté et proposons un protocole permettant de générer automatiquement une mesure de compréhension. Les scores ROUGE et BERTScore sont comparés aux évaluations humaines, permettant d'analyser les divergences et d'identifier les spécificités de ce type de données. Cette première étude ouvre la voie au déploiement complet de l'exercice.

ABSTRACT

Towards Automated Evaluation of Children's Oral Summaries : An Exploratory Study of Comprehension Metrics

Learning to read requires the development of comprehension strategies, whose assessment is a major pedagogical concern. The CHICA-AI project introduces a digital platform for practicing oral summarization, aimed at 4th and 5th grade students (CM1-CM2), featuring an automated evaluation tool. This paper examines the relevance of applying standard metrics, originally designed for written text, to children's oral summaries. After presenting the platform, we detail the main characteristics of the collected corpus and propose a protocol for automatically generating comprehension scores. ROUGE and BERTScore metrics are compared against human evaluations, allowing for an analysis of discrepancies and the identification of specificities inherent to this type of data. This preliminary paves the way for the full deployment of the exercise.

MOTS-CLÉS : résumé oral, évaluation automatique, compréhension de la lecture.

KEYWORDS: oral summary, automatic evaluation, reading understanding.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 655–668.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Le résumé oral de textes écrits à l'école primaire

1.1 L'intérêt de la pratique du résumé oral

À partir du cycle 3, l'apprentissage de la lecture marque une transition du « apprendre à lire » vers le « lire pour apprendre », où l'enseignement se concentre davantage sur les compétences de haut niveau, telle que la compréhension écrite (Chall, 1983; Brossette *et al.*, 2025). Cette compétence implique de sélectionner les informations pertinentes, d'établir des inférences, de contrôler la cohérence de son interprétation et d'ajuster sa lecture en cas d'incompréhension. Les élèves sont incités à mobiliser ces processus à travers l'enseignement de stratégies explicites (Elleman & Oslund, 2019). Les enseignants utilisent notamment l'activité de résumé, qui permet de solliciter conjointement plusieurs de ces leviers (Ophuis-Cox *et al.*, 2024).

La littérature s'est principalement intéressée aux résumés écrits produits à partir du collège et au-delà (Stevens *et al.*, 2019). Pourtant, chez les élèves les plus jeunes, le résumé est souvent réalisé à l'oral, cette modalité permettant de contourner les difficultés rédactionnelles encore présentes à l'école primaire. Toutefois, cette pratique impose à l'enseignant de veiller à répartir équitablement le temps de parole, tout en évaluant de manière objective une production qui, par nature, disparaît une fois énoncée. Pour lever ces freins, le développement d'un Environnement Informatique pour l'Apprentissage Humain (EIAH) dédié à cet exercice présente un intérêt évident pour l'enseignant. Ce dispositif permettrait la production simultanée de plusieurs résumés oraux, assurerait la conservation des productions et ouvrirait la possibilité d'une évaluation automatisée.

1.2 Développement d'un EIAH dédié au résumé oral

Dans le cadre du projet ANR CHICA-AI, un EIAH a été conçu sur la plateforme Lalilo pour soutenir la pratique autonome du résumé oral. Il permet à l'élève de lire un texte et d'en produire un résumé à l'oral, en s'appuyant sur des attentes explicites. Ainsi, nous avons conçu des leçons préliminaires intégrées à la plateforme, expliquant la construction d'un résumé oral, afin que les élèves partagent une représentation commune de l'activité.

Les productions orales diffèrent de celles à l'écrit sur plusieurs aspects. Dans un premier temps, la variabilité de représentation de la tâche, déjà observée chez des adolescents (Brown & Day, 1983) pourrait être plus marquée à l'école primaire. Une compréhension partielle des attentes peut amener à un résumé oral dont l'organisation diverge de la structure attendue, compliquant l'évaluation automatique basée sur des indices structurels ou macrostructurels. Par ailleurs, le vocabulaire en développement (Rice & Hoffman, 2015) entraîne approximations, substitutions ou omissions, fragilisant les méthodes fondées sur la seule similarité lexicale. Enfin, la syntaxe elle-même, caractérisée par des phrases qui peuvent être plus courtes, inachevées ou structurées différemment de l'écrit (O'Donnell *et al.*, 1967; Nippold *et al.*, 2005), rend l'analyse automatique difficile pour des modèles entraînés sur des corpus écrits.

En outre, la modalité orale impose des contraintes mnésiques spécifiques. Contrairement à l'écrit, le texte n'est plus disponible lors de la restitution et l'élève doit maintenir les informations pertinentes en mémoire tout en planifiant sa production. Cette charge cognitive, pouvant être accrue par l'environnement bruyant et distrayant de la classe, peut favoriser une restitution linéaire des idées au fil de leur récupération en mémoire plutôt que selon l'organisation du texte initial. Ce phénomène nuit à

l'efficacité des mesures fondées sur la correspondance séquentielle entre un texte source et un résumé.

1.3 Objectifs

Le projet vise à traiter automatiquement les productions des enfants afin de proposer un retour pertinent à l'élève et à son enseignant. L'annotation de chaque résumé selon une grille d'évaluation révèle une hétérogénéité des productions qui questionne l'applicabilité des mesures d'évaluation automatiques, traditionnellement dévolues à l'écrit. Dans cette première étude, nous explorons différentes méthodes de notation automatique et procédons à une analyse qualitative des erreurs observées. L'objectif est de mettre en évidence les difficultés liées au changement de modalité (*orale*) et de public (*enfants de CM1-CM2*) afin de mieux caractériser la nature du résumé oral d'enfant.

2 Données

2.1 Création des ressources pédagogiques

L'activité de résumé oral est proposée aux élèves dans le cadre de la progression pédagogique de la plateforme Lalilo. À cet effet, l'équipe pédagogique a créé cinq textes narratifs (*Story*), de difficulté croissante et ciblant différentes compétences. Ces supports permettent ainsi à chaque enfant de travailler sa compréhension à plusieurs reprises et en autonomie. Nous désignerons ces cinq récits par le terme mis en gras dans le titre. Leurs caractéristiques statistiques, obtenues après tokenisation par mots et lemmatisation par l'outil Stanza [1], sont détaillées dans la Table 1.

Titre	# mots	# lexique
Le prix de l'**indifférence**	1 163	355
Le **paon** aux plumes d'or	1 372	384
L'**anneau** du destin	1 491	440
Tyrannos et l'oracle de Dodone	1 372	417
Le procès de la **sorcière**	1 358	381
Total	6 756	1 256

TABLE 1 – Textes originaux proposés à la lecture pour résumer, présentés selon l'ordre pédagogique

Pour évaluer les productions, une première grille d'évaluation détaillée répertorie les éléments clés attendus : personnages (principaux et secondaires), lieux et idées principales. Parallèlement, trois types de résumés *écrits* servent de référence pour chaque texte :

— *Detailed* : une version détaillée réduisant la taille du texte original à environ 30% ;
— *Short* : une version courte, limitée à 10% de la taille originale ;
— *Child* : une version calibrée sur les productions d'élèves de Cycle 3 avec une taille finale réduite à 15%.

1. https://stanfordnlp.github.io/stanza/index.html

2.2 Collecte et annotation du corpus de résumé

Sur la plateforme Lalilo, l'élève accède d'abord aux leçons préparant à la production d'un résumé oral. Une fois celles-ci assimilées, l'activité de résumé oral est débloquée. Pour faciliter leur lecture, les textes sont présentés par pages de 300 mots maximum, parcourues à l'aide de flèches directionnelles. Afin d'assurer une lecture attentive, des questions sont posées à intervalles réguliers et conditionnent l'accès à la suite du récit. À l'issue de la lecture, l'enfant enregistre son résumé à voix haute, avec la possibilité de le réécouter et de le réenregistrer avant validation.

L'enregistrement se faisant en complète autonomie par l'élève dans sa classe, 50% des fichiers audio récupérés sont inexploitables, pour cause de bruits de fonds, de paroles incompréhensibles ou de fausse manipulation. Les fichiers valides sont alors transcrits automatiquement par `Whisper-large-v3` [2], puis corrigés manuellement par six annotatrices, étudiantes en orthophonie. Le taux d'erreur mots (WER) de la transcription automatique atteint 33%, en raison à la fois du caractère écologique du corpus et de l'inadaptation de Whisper à notre tâche et à la parole d'enfants. Le développement d'un tel système constitue un objectif du projet. Dans l'attente, nous utilisons les transcriptions de référence afin d'éviter que ces erreurs n'affectent l'étude. Celles-ci conservent les disfluences, répétitions et erreurs de prononciation (par ex. « pahon » pour « paon »).

Certains enregistrements ont fait l'objet d'annotations multiples. Nous n'avons conservé qu'un seul exemplaire des transcriptions manuelles strictement identiques. Ainsi, un même fichier audio peut être associé à plusieurs transcriptions lorsqu'elles diffèrent, même légèrement. À l'inverse, des fichiers audio différents très courts peuvent partager une même transcription (par exemple « Léonard »).

Une fois la transcription corrigée, les annotatrices ont procédé à l'analyse des résumés selon la grille d'évaluation fournie. Sur la base de ces critères, une note globale de compréhension de 0 à 10 est attribuée de manière subjective. D'autres notes sont également données concernant la structure du résumé ou encore sa fluidité. La tâche d'annotation a fait l'objet d'une publication détaillée (Labbé *et al.*, 2025). Les résultats de la dernière phase de test montrent que l'accord inter-annotatrices, mesuré par coefficient de corrélation intra-classe à deux facteurs mixtes (ICC3), est de 0,81 sur la note globale de compréhension. Ce niveau d'accord élevé justifie son utilisation comme score de référence pour notre étude.

2.3 Premières statistiques descriptives

Préalablement à l'étude des mesures d'évaluation automatique de la qualité, nous avons mené une première étude descriptive des données. La table 2 présente quelques statistiques sur le corpus recueilli. Du fait de l'annotation multiple d'un même fichier, le nombre de fichiers audio exploitables et celui des transcriptions manuelles uniques diffèrent. Pour ces cas, les notes attribuées par les annotatrices ont été agrégées en prenant leur moyenne.

On observe une hétérogénéité marquée dans les productions, tant sur le plan quantitatif que qualitatif, distinguant ce corpus des jeux de données classiquement utilisés pour calibrer les métriques automatiques. Sur le plan structurel, les résumés d'enfants présentent de fortes disparités de longueur, oscillant entre une phrase unique et des récits plus exhaustifs de plus de 350 mots.

Au-delà de ces écarts de format, la nature orale des productions induit une spécificité lexicale majeure :

2. https://huggingface.co/openai/whisper-large-v3

Titre	# fichiers audio exploitables	# transcriptions manuelles uniques	# mots moyenne	# mots écart-type
Indifférence	237	224	83,5	75,3
Paon	253	245	100,2	97,9
Anneau	169	162	66,9	75,3
Tyrannos	151	147	53,4	60,8
Sorcière	215	197	61,6	70,5
Total	1025	975	76,0	80,5

TABLE 2 – Détails du corpus récolté en fonction des histoires

alors que la densité lexicale [3] des références écrites excède les 60%, elle chute à environ 40% pour les productions orales de notre corpus. Cet écart s'explique par les caractéristiques de la parole spontanée : disfluences, répétitions et recours plus systématique aux marqueurs logiques pour guider l'auditeur. Cette charge cognitive de planification orale, couplée à l'effort mnésique de restitution d'informations pertinentes, complexifie la comparaison automatique avec des références écrites, par nature plus denses et synthétiques.

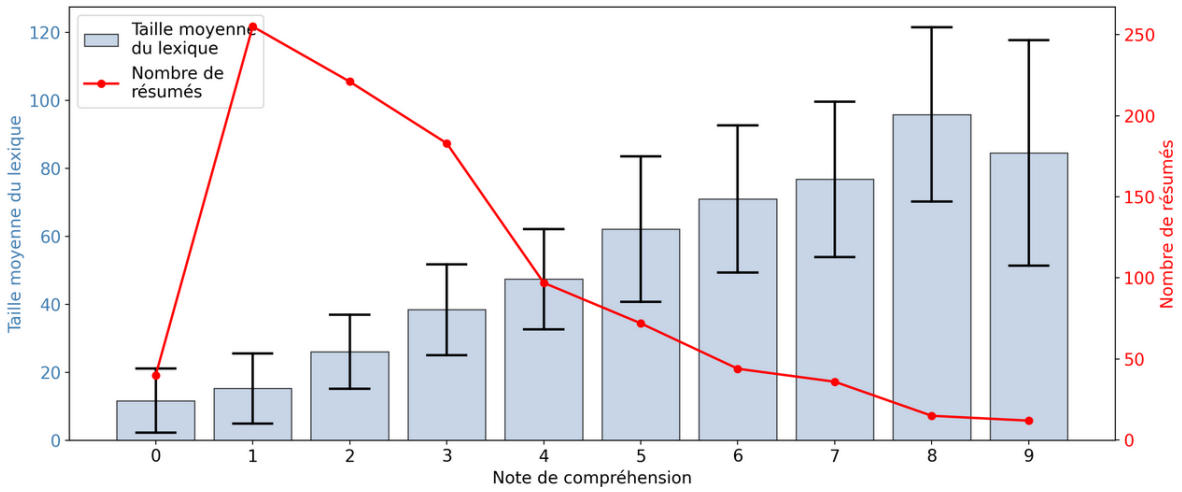

FIGURE 1 – Taille du lexique et distribution des résumés par note de compréhension

Concernant la qualité de compréhension, la Figure 1 présente la distribution des 975 notes attribuées par les annotatrices. L'hétérogénéité des résultats est marquée : 72% des résumés ont obtenu une note inférieure ou égale à 3/10. Cette forte concentration de notes basses, contrastant avec la rareté des notes supérieures ou égales à 7 (6,5%), constitue un biais structurel qu'il conviendra d'intégrer à l'analyse des résultats. Par ailleurs, la figure met en évidence une corrélation positive entre la richesse lexicale et la note attribuée. Ceci n'est pas étonnant puisqu'un résumé plus long et lexicalement dense permet de restituer davantage d'unités informationnelles, augmentant ainsi l'exhaustivité perçue par l'annotatrice et, par extension, la note de compréhension finale.

3. Proportion de mots lexicaux, c'est à dire ne figurant pas dans la liste des 200 mots outils de la bibliothèque python `stop-words` (https://github.com/Alir3z4/python-stop-words).

3 Mesures d'évaluation de la qualité d'un résumé

Après avoir décrit les caractéristiques des résumés de notre corpus, nous nous intéressons à leur évaluation.

3.1 Les mesures existantes

Depuis l'émergence des systèmes de résumé de texte, l'évaluation automatisée constitue une préoccupation majeure, les méthodes manuelles étant peu reproductibles. La définition de mesures corrélées au jugement humain a conduit à de nombreuses études et est devenu un domaine de recherche à part entière (Zhang *et al.*, 2025). La majorité de ces métriques s'appuient sur un ou plusieurs résumés de référence et évaluent leur similarité avec le résumé automatique. Les mesures basées sur le chevauchement lexical de n-grammes, dont les plus connues sont les variantes de ROUGE (Lin, 2004), ont longtemps constitué le standard pour évaluer la performance. Plus récemment, l'avènement des encodeurs neuronaux a conduit à l'apparition de métriques de similarité basées sur les plongements contextuels, telles que BERTScore (Zhang *et al.*, 2020), MoverScore (Zhao *et al.*, 2019) ou BARTScore (Yuan *et al.*, 2021). Parmi les évolutions récentes, rendues nécessaires par l'émergence des résumés abstractifs, des chercheurs ont exploré les métriques pour évaluer automatiquement la fidélité des résumés en les comparant avec des documents d'origine à l'aide de modèles dérivés de BERT (Kryściński *et al.*, 2020) ou de systèmes de questions réponses (Scialom *et al.*, 2021). Enfin, parmi les approches récentes, on peut citer l'utilisation de grands modèles de langues (LLM) pour évaluer les résumés en fonction de critères tels que la cohérence, la fluidité, l'exhaustivité et la pertinence. Le modèle reçoit alors une grille d'évaluation ou des indications sur les critères à prendre en compte afin de fournir une note d'évaluation (Fabbri *et al.*, 2021).

Bien que ces mesures aient été initialement conçues pour l'évaluation de résumés textuels et générés automatiquement, nous souhaitons tester leur performance sur des transcriptions de résumés oraux produits par des enfants. Dans ce cadre, deux facteurs sont susceptibles d'impacter la performance des mesures : le passage d'une production automatique à une production humaine, d'enfants qui plus est, et le passage de la modalité écrite à la modalité orale.

3.2 Les mesures évaluées

Parmi les métriques existantes, nous avons testé ROUGE et BERTScore, représentatives des approches lexicales et sémantiques.

ROUGE demeure la mesure la plus répandue pour l'évaluation de résumé. Elle compare des séquences de tokens entre résumés produits et de référence sans tenir compte de la position de la chaîne dans le résumé. Elle se décline en plusieurs variantes : une approche par n-grammes (ROUGE-n, recherche d'une suite de n tokens) ou l'identification de la plus longue séquence commune. Pour notre étude, nous avons retenu ROUGE-1, ROUGE-2, ROUGE-L.

Pour compléter ROUGE qui se limite à la recherche des séquences de mots exacts, nous avons testé BERTScore. Cette mesure privilégie la détection du sens et pénalise moins l'utilisation de synonymes. BERTScore convertit chaque mot de la production et de la référence en un vecteur puis mesure la similarité cosinus entre chaque vecteur du résumé et chaque vecteur de la référence. Pour chaque

mot du résumé produit, l'algorithme ne conserve que son meilleur correspondant dans la référence, c'est-à-dire le vecteur avec lequel la similarité cosinus est la plus élevée, et inversement. À partir de cela, une précision, un rappel et une f-mesure sont calculés. Le rappel, noté BERTScore-R, semble particulièrement adapté à notre tâche, puisqu'il permet de valoriser la restitution d'informations sans pénaliser les enfants utilisant leur propre vocabulaire.

4 Expériences et résultats

4.1 Protocole expérimental

Afin d'évaluer la pertinence de ces différentes mesures, nous avons calculé leur coefficient de corrélation de Spearman avec la note globale de compréhension. Fondé sur le rang des observations, cet indicateur mesure l'intensité d'une relation monotone entre deux variables. Ce choix est ici privilégié car la relation entre les scores automatiques et les évaluations manuelles n'est pas strictement linéaire.

L'analyse qualitative des écarts entre scores automatiques et notes humaines s'appuie également sur une comparaison de leurs rangs respectifs. Pour identifier les résumés dont le classement diverge significativement de l'évaluation humaine, nous appliquons la méthode de Tukey (1977), selon laquelle une valeur est considérée comme aberrante si elle se situe hors de l'intervalle :

$$[\,Q_1 - 1,5 \times IQR\,;\; Q_3 + 1,5 \times IQR\,]$$

où $IQR = Q_3 - Q_1$ représente l'écart interquartile, Q_1 et Q_3 étant respectivement les premier et troisième quartiles. Nous qualifions ainsi de *résumés mal évalués* ceux dont l'écart de rang dépasse la borne supérieure de cet intervalle, calculée spécifiquement pour chaque mesure et chaque modèle.

Enfin, pour comparer l'efficacité concrète de ces mesures, nous projetons les rangs obtenus sur une échelle de notation sur 10 (voir Annexe A). Cette méthode tend vers une correspondance avec l'évaluation humaine en s'appuyant sur l'ordonnancement des données, bien qu'elle reste tributaire de leur distribution. À partir de cette projection, nous mesurons l'exactitude, stricte et à un point près, des classements automatiques par rapport aux notes de référence.

4.2 Test des mesures

Référence	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-R CamemBERT	BERTScore-R DeBERTa
Story	0,82	0,81	0,82	0,84	0,80
Detailed	0,82	0,79	**0,84**	**0,86**	0,85
Short	0,76	0,67	0,71	0,84	0,84
Child	0,81	0,71	0,79	0,85	0,84

TABLE 3 – Corrélations de Spearman entre le score de compréhension et différentes mesures d'évaluation.

ROUGE Les scores obtenus en combinant les différentes variantes de ROUGE avec le texte source ou les résumés de référence sont globalement très faibles. À titre d'exemple, la moyenne pour ROUGE-2 n'atteint que 0,02 avec le texte d'origine et 0,05 avec le résumé détaillé. La performance maximale observée s'élève à 0,25 (ROUGE-1) en utilisant le résumé court comme référence. Cette faiblesse des scores s'explique par la nature des productions : il s'agit de reformulations d'enfants dont la majorité a obtenu une note de compréhension inférieure ou égale à 3 (Figure 1).

Ces scores ROUGE sont néanmoins très intéressants car ils corrèlent fortement avec les notes de compréhension données par les annotatrices (Table 3). Bien que les scores ROUGE soient moins élevés lorsque les références considérées sont plus longues (Story ou Detailed), leur corrélation avec les notes des annotatrices est plus importante, atteignant un coefficient de Spearman de 0,84 pour le couple ROUGE-L/Detailed. Cette convergence se confirme par l'analyse des rangs : l'exactitude stricte s'élève à 41,6%, indiquant que, pour près de la moitié des observations, le score automatique se situe dans l'intervalle de rangs défini par l'évaluation humaine. En autorisant une marge d'erreur d'un point, ce taux de fiabilité atteint 83,2%. Enfin, l'analyse des écarts extrêmes de rang révèle un taux de résumés mal évalués marginal, limité à 3,6%.

BERTScore Nous avons évalué plusieurs modèles BERT, tant multilingues que monolingues. La plupart de ces modèles imposent une limite de tokens pour la construction des plongements, entraînant une troncature du texte au-delà de ce seuil. Parmi les modèles multilingues compatibles avec BERTScore, le modèle `deberta-v3-base` [4] (He *et al.*, 2021b,a) figure parmi les plus performants au niveau de la corrélation sur le jeu de données WMT16 To-English [5]. Ce modèle a initialement été choisi car le document fourni par BERTScore indiquait une absence de limite de tokens, ce qui permettait donc que les textes servant de référence ne soient pas tronquer. Après vérification, DeBERTa a en réalité une limite d'entrée fixée à 512 tokens, comme la majorité des modèles compatibles avec BERTScore.

Contrairement à `deberta-v3-base` pour lequel BERTScore définit par défaut la couche offrant la meilleure corrélation sur WMT16, aucune configuration préétablie n'existe pour les modèles monolingues français. Nous avons donc testé les couches 7 à 11 de plusieurs modèles présélectionnés. Le modèle `camembert-base` [6](Martin *et al.*, 2020) a montré une corrélation constante sur nos données. Par souci de cohérence et de comparabilité, nous avons retenu ce modèle en exploitant sa 9^{e} couche, correspondant à la couche par défaut de `deberta-v3-base`, les deux architectures possédant un nombre de couches identique, soit 12 couches. CamemBERT a une limite de tokens en entrée similaire à DeBERTa, soit 514 tokens, mais un nombre de paramètres plus important (110 millions de paramètres pour CamemBERT contre 86 millions pour DeBERTa-v3).

Comme pour ROUGE, les scores moyens observés sont relativement bas (entre 0,40 et 0,55) et présentent une forte corrélation avec les notes de compréhension (Table 3). Les meilleurs résultats sont obtenus avec les résumés détaillés, bien que l'influence de la référence soit ici moins marquée. Les deux modèles affichent des performances comparables : CamemBERT obtient une exactitude stricte de 44,9% contre 44,6% pour DeBERTa. Lorsque l'on considère l'exactitude à un point près, ces scores s'élèvent respectivement à 86,9% et 85,9%. Quant à la proportion de résumés mal-évalués, elle demeure marginale avec 2,6% pour CamemBERT et 3,1% pour DeBERTa.

4. https://huggingface.co/microsoft/deberta-v3-base
5. Classement disponible via le dépôt GitHub de BERTScore : https://docs.google.com/spreadsheets/d/1RKOVpselB98Nnh_EOC4A2BYn8_201tmPODpNWu4w7xI/edit?gid=0#gid=0
6. https://huggingface.co/almanach/camembert-base

Réflexion sur les résultats Si ces mesures n'atteignent pas les standards habituels de l'état de l'art [7], elles reflètent néanmoins la qualité des résumés des enfants, comme le montrent les fortes corrélations de Spearman. Les mesures ROUGE obtiennent de meilleures corrélations lorsque la référence est longue (Story ou Detailed), tandis que les BERTScore sont moins sensibles au changement de référence. On observe toutefois que DeBERTa, modèle multilingue, est légèrement impacté lorsque le texte original est utilisé comme référence.

Ces résultats nous amènent à réfléchir à la relation entre la longueur, la richesse lexicale et la qualité du résumé. Le calcul du coefficient de corrélation de Pearson met en évidence un lien fort de la note humaine avec la taille du lexique (0,81), mais également avec la longueur brute du résumé (0,77). Toutefois, cette relation ne signifie pas qu'un résumé plus long est nécessairement meilleur : il ne s'agit pas simplement d'aligner un grand nombre de mots, mais de mobiliser un vocabulaire riche et pertinent au regard de l'histoire originale. Sans cette adéquation sémantique, les métriques ROUGE ou fondées sur BERT ne pourraient converger avec le jugement humain. Afin de mieux comprendre ce phénomène, nous analysons dans la sous-section suivante plusieurs résumés.

4.3 Analyse des erreurs

Pour la suite de notre étude, nous avons restreint l'analyse aux trois mesures suivantes : ROUGE-L, CamemBERT et DeBERTa, avec le résumé détaillé comme référence. Le croisement de leur résultats permet d'identifier les productions considérées comme mal-évaluées par l'ensemble de ces métriques (voir Figure 2).

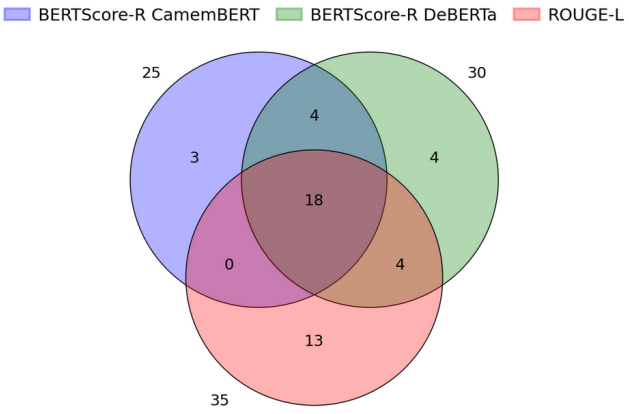

FIGURE 2 – Diagramme de Venn : distribution des résumés mal évalués selon trois mesures.

On y observe un noyau dur de 18 résumés systématiquement mal évalués, suggérant des difficultés structurelles ou sémantiques qui échappent tant aux approches lexicales qu'aux modèles contextuels. Si ROUGE-L présente le plus grand nombre de cas de mauvaise évaluation (35 résumés), dont 13 lui sont exclusifs, les BERTScores manifestent une plus grande robustesse avec respectivement 25 erreurs pour CamemBERT et 30 pour DeBERTa, dont la majorité est commune aux deux modèles.

Afin de cerner les limites des mesures automatiques classiques face aux exigences de la tâche, nous présentons trois extraits de résumés mal-évalués issus de ce consensus d'erreurs (les résumés complets et leurs rangs sont disponibles en Annexe B) :

Résumé 1 (taille > 100 mots — note = 2 — BERTScore et ROUGE-L ≈ 5)

C'est l'histoire de Léonard. Il écoute pas ses voisins. Il alla se promener au bord de la rivière. Il rentre su le . [...] il est clair ensuite il va à l'apothicaire et le drapeau d'or "prend un bain de soleil ça va te réchauffer" et ensuite il tombera vite malade et il n'avait plus de provisions ni plus de bois.

7. Un score ROUGE-L est généralement jugé satisfaisant au-delà de 0,3 et un BERTScore au-delà de 0,8.

Dans ce premier cas, les mesures automatiques surestiment la qualité du résumé. Ce décalage s'explique par la présence de nombreux éléments lexicaux communs avec la référence, tels que les personnages (Léonard, le meunier, le bûcheron) et certaines unités informationnelles (la promenade au bord de la rivière, l'absence de provisions). Ces correspondances ponctuelles génèrent de bons scores de similarité ; toutefois, elles occultent la confusion structurelle du résumé. L'évaluation humaine sanctionne ici un manque de clarté que les mesures de similarité ne parviennent pas à détecter.

Résumé 2 (taille < 20 mots — note = 3 — BERTScore ≈ 1,5, Rouge-L ≈ 2)

c'est ça parle d'un homme, qui a pas voulu aider ses voisins, donc il en est tombé malade. Voilà.

À l'inverse, les scores automatiques sous-estiment ici la production. Si l'enfant démontre une compréhension globale de l'intrigue, son récit reste extrêmement concis : les noms des personnages et le détail des actions sont largement omis. Cette absence de termes pivots réduit mécaniquement le recouvrement lexical et sémantique avec la référence. Cependant, l'annotatrice valorise la cohérence d'ensemble du propos, là où les modèles automatiques, pénalisés par la pauvreté lexicale, échouent à reconnaître la validité de la compréhension.

Résumé 3 (taille ≈ 100 mots – note = 0 – BERTScore ≈ 3, Rouge-L ≈ 3.5)

Ça parle de la jeune fille à la, la jeune fille. La jeune fille il voulait se marier, elle voulait se marier je voulais dire. […] Son père a dit "maintenant, ram ramène le le pierre le pierre le pierre qui peut se transformer l'eau, le sable je voulais dire, le sable en n en eau.

Le troisième résumé présente une autre forme de surestimation. Bien que la production soit inaboutie, elle intègre de nombreux éléments factuels présents dans la référence, ce qui gonfle les scores de similarité. Cependant, l'enfant commet un contresens majeur en ne saisissant pas le refus de la jeune fille de se marier. Ce point est crucial car il constitue l'élément déclencheur de toute la suite de l'intrigue ; son absence ou sa mauvaise interprétation semble donc jugée rédhibitoire, signifiant que l'essence même de l'histoire n'a pas été comprise. Ici, la limite des mesures automatiques réside dans leur incapacité à hiérarchiser l'information : elles valorisent la présence d'idées secondaires là où l'humain sanctionne l'échec de la compréhension du nœud dramatique.

5 Conclusion et perspectives

Dans cet article, nous avons proposé un protocole expérimental pour évaluer la compréhension de la lecture chez l'élève de Cycle 3 à travers la production de résumés oraux. Nous avons testé différentes mesures automatiques et analysé les caractéristiques des résumés. Nos travaux ont permis de confronter des métriques de similarité textuelle standard (ROUGE et BERTScore) aux jugements subjectifs d'annotatrices expertes. Un premier résultat majeur réside dans la forte corrélation observée entre les scores automatiques et les évaluations humaines. Malgré la spécificité du public (enfants de 9-11 ans) et de la modalité (parole spontanée), ces mesures parviennent à capturer une part significative de la qualité de la compréhension.

Toutefois, deux limites principales apparaissent : la distribution fortement asymétrique des résumés, avec une prédominance de productions jugées manuellement de faible qualité, et l'usage uniquement de transcriptions corrigées, sans tenir compte des erreurs de reconnaissance automatique. L'intégration

future de transcriptions brutes, issues de modèles comme Whisper, constituera une étape cruciale pour valider la robustesse de ces mesures en conditions réelles sur la plateforme pédagogique.

Au-delà de l'aspect technique, cette étude initiale met en avant les caractéristiques du résumé oral chez l'enfant. Nos analyses qualitatives révèlent que si les métriques actuelles sont sensibles à la richesse lexicale, elles peinent encore à hiérarchiser l'information et à détecter des contresens narratifs majeurs. Ce constat souligne la nécessité de dépasser la simple mesure de similarité. Les différentes analyses menées guideront l'affinement de la grille d'évaluation et l'exploration d'approches de type *LLM-as-a-judge*, qui pourraient offrir des retours plus fins et pédagogiquement plus riches aux élèves et enseignants.

Références

BROSSETTE B., LEFÈVRE E., GRAINGER J. & LÉTÉ B. (2025). On the relation between single word and multiple word processing during learning to read. *Journal of Experimental Child Psychology*. DOI : https ://doi.org/10.1016/j.jecp.2025.106223.

BROWN A. & DAY J. (1983). Macrorules for summarizing texts : the development of expertise. *Journal of Verbal Learning and Verbal Behavior*. DOI : 10.1016/S0022-5371(83)80002-4.

CHALL J. S. (1983). *Stages of Reading Development*. McGraw-Hill.

ELLEMAN A. & OSLUND E. (2019). Reading comprehension research : Implications for practice and policy. *Policy Insights from the Behavioral and Brain Sciences*. DOI : 10.1177/2372732218816339.

FABBRI A. R., KRYŚCIŃSKI W., MCCANN B., XIONG C., SOCHER R. & RADEV D. (2021). SummEval : Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, **9**, 391–409.

HE P., GAO J. & CHEN W. (2021a). DeBERTaV3 : Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

HE P., LIU X., GAO J. & CHEN W. (2021b). DeBERTa : Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

KRYŚCIŃSKI W., MCCANN B., XIONG C. & SOCHER R. (2020). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, p. 9332–9346.

LABBÉ E., BROSSETTE B., CAMELIN N., CAUDRELIER T., CAVALLI E., FERRANÉ I., LUTZ B., MORICEAU V., PELLEGRINI T., PINQUIER J., PRAT C. & GELIN L. (2025). Annotation de résumés oraux d'élèves de primaire pour l'analyse automatique des capacités de compréhension de la lecture. In *CORIA-TALN-RJCRI-RECITAL*.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

NIPPOLD M., HESKETH L., DUTHIE J. & MANSFIELD T. (2005). Conversational versus expository discourse. *Journal of Speech, Language, and Hearing Research*. DOI : 10.1044/1092-4388(2005/073).

O'DONNELL R., GRIFFIN W. & NORRIS R. (1967). A transformational analysis of oral and written grammatical structures in the language of children in grades three, five, and seven. *The Journal of Educational Research*.

OPHUIS-COX F., ROZENDAL L., CATRYSSE L., JOOSTEN-TEN BRINKE D. & CAMP G. (2024). The effects of summarization and factual retrieval practice on text comprehension and text retention in elementary education. *Journal of Experimental Psychology : Applied*. DOI : 10.1037/xap0000507.

RICE M. & HOFFMAN L. (2015). Predicting vocabulary growth in children with and without specific language impairment : A longitudinal study from 2 ;6 to 21 years of age. *Journal of Speech, Language, and Hearing Research*. DOI : 10.1044/2015_JSLHR-L-14-0150.

SCIALOM T., DRAY P.-A., LAMPRIER S., PIWOWARSKI B., STAIANO J., WANG A. & GALLINARI P. (2021). QuestEval : Summarization asks for fact-based evaluation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, p. 6594–6604.

STEVENS E., PARK S. & VAUGHN S. (2019). A review of summarizing and main idea interventions for struggling readers in grades 3 through 12 : 1978–2016. *Remedial and Special Education*. DOI : 10.1177/0741932517749940.

TUKEY J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

YUAN W., NEUBIG G. & LIU P. (2021). BARTScore : Evaluating Generated Text as Text Generation. In M. RANZATO, A. BEYGELZIMER, Y. DAUPHIN, P. S. LIANG & J. W. VAUGHAN, Éds., *Advances in Neural Information Processing Systems*, volume 34, p. 27263–27277 : Curran Associates, Inc.

ZHANG H., YU P. S. & ZHANG J. (2025). A systematic survey of text summarization : From statistical methods to large language models. *ACM Computing Surveys*, **57**(11), 1–41.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating text generation with BERT. In *International Conference on Learning Representations*.

ZHAO W., PEYRARD M., LIU F., GAO Y., MEYER C. M. & EGER S. (2019). MoverScore : Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, p. 563–578.

A Correspondance rang-note

Pour comparer les scores automatiques (valeurs continues) aux notes humaines (valeurs discrètes), nous avons établi une table de correspondance basée sur l'ordonnancement des données. Les résumés partageant une même note humaine se voient attribuer un rang humain identique, calculé comme le rang médian de leur groupe. Ces rangs médians servent ensuite de référence pour définir des intervalles de notation automatique (voir Table 4). Tout résumé dont le score automatique se situe dans l'un de ces intervalles reçoit la note correspondante (ex : un rang automatique entre 1 et 40 équivaut à un 0).

Note / 10	Rang médian (référence humaine)	Intervalle de rang automatique
0	20,5	[1 ; 40]
1	168	[41 ; 295]
2	406	[296 ; 516]
3	608	[517 ; 699]
4	748	[700 ; 796]
5	832,5	[797 ; 868]
6	890,5	[869 ; 912]
7	930,5	[913 ; 948]
8	956	[949 ; 963]
9	969,5	[964 ; 975]

TABLE 4 – Grille de correspondance entre les rangs humains partagés et les intervalles de notation automatique

B Exemples de résumés jugés mal-évalués

B.1 Résumé 1

C'est l'histoire de Léonard. Il écoute pas ses voisins. Il alla se promener au bord de la rivière. Il rentre su le . Il il voit des flocons de neige tomber. Il demanda il demanda la meunier. Elle lui dit, "va prendre un bain de soleil". Elle lui claque les appors ensuite le le boucheron lui dit "va prendre un bain de soleil ça va te réchauffer" ensuite il est clair ensuite il va à l'apothicaire et le drapeau d'or "prend un bain de soleil ça va te réchauffer" et ensuite il tombera vite malade et il n'avait plus de provisions ni plus de bois.

Mesure	**Note**	**Rang**
Référence	2	$\approx$ 406
BERTScore (CamemBERT)	$\approx$ 5,5	853
BERTScore (DeBERTa)	$\approx$ 5	821
ROUGE-L	$\approx$ 5	813

TABLE 5 – Évaluation du résumé 1

B.2 Résumé 2

C'est ça parle d'un homme, qui a pas voulu aider ses voisins, donc il en est tombé malade. Voilà.

Mesure	Note	Rang
Référence	3	≈ 608
BERTScore (CamemBERT)	$\approx 1{,}5$	226
BERTScore (DeBERTa)	$\approx 1{,}5$	221
ROUGE-L	≈ 2	448

TABLE 6 – Évaluation du résumé 2

B.3 Résumé 3

Ça parle de la jeune fille à la, la jeune fille. La jeune fille il voulait se marier, elle voulait se marier je voulais dire. Et son père a dit que "si tu veux te marier, va me chercher un fleur rare". Et les elle est y allée et il a vu elle a vu un fleur le fleur rare. Il l'a ramené à son père. Son père a dit "maintenant, ram ramène le le pierre le pierre le pierre qui peut se transformer l'eau, le sable je voulais dire, le sable en n en eau.

Mesure	Note	Rang
Référence	0	$\approx 20{,}5$
BERTScore (CamemBERT)	≈ 3	544
BERTScore (DeBERTa)	≈ 3	564
ROUGE-L	$\approx 3{,}5$	651

TABLE 7 – Évaluation du résumé 3