*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Annotation de résumés oraux d'élèves de primaire pour l'analyse automatique des capacités de compréhension de la lecture

Etienne Labbé[2] Brice Brossette[3] Nathalie Camelin[4] Tiphaine Caudrelier[4] Eddy Cavalli[4] Isabelle Ferrané[2] Barbara Lutz[1] Véronique Moriceau[2] Thomas Pellegrini[2] Julien Pinquier[2] Cantin Prat[4] Lucile Gelin[1, 2]
(1) Lalilo by Renaissance Learning, France
(2) IRIT, Université Paul Sabatier, CNRS, Toulouse, France
(3) Laboratoire d'Etude des Mécanismes Cognitifs (EMC), Université Lumière Lyon 2, Lyon, France
(4) Laboratoire d'Informatique d'Avignon (LIA), Avignon, France
etienne.labbe@irit.fr, lucile.gelin@renaissance.com

## RÉSUMÉ

Le projet CHICA-AI vise à construire une activité assistée par ordinateur pour l'entraînement des compétences de compréhension de la lecture des élèves de primaire. Cette activité consiste à demander à l'élève de résumer à l'oral un texte narratif, afin d'identifier ses difficultés de compréhension et fournir un retour personnalisé à l'élève et à son enseignant. Pour cela, nous mettrons en place un système automatique d'analyse fine des résumés oraux, capable d'extraire les informations pertinentes et de les combiner pour remplir une grille de critères pédagogiques et psycho-cognitifs. Nous présentons ici les défis du projet, ainsi que les premiers travaux réalisés : création de l'activité dans la plateforme Lalilo et du contenu pédagogique, collecte d'enregistrements audios, construction du protocole d'annotation. Nous présentons enfin les analyses préliminaires faites sur les premières annotations, qui serviront à l'entraînement et l'évaluation de notre système automatique.

## ABSTRACT

**Annotation of oral summaries from primary school students for automatic analysis of reading comprehension skills**

The CHICA-AI project aims to build a computer-assisted learning activity for training the reading comprehension skills of primary school pupils. This activity involves asking students to orally summarize a narrative text, in order to identify their comprehension difficulties and provide personalized feedback to the pupil and their teacher. To do this, we'll be implementing an automatic system for fine-grained analysis of oral summaries, capable of extracting relevant information and combining it to fill out a grid of pedagogical and psycho-cognitive criteria. We present here the challenges of the project, as well as the first tasks carried out : creation of the activity in the Lalilo platform and of the pedagogical content, audio recordings collection, construction of the annotation protocol. Finally, we present the preliminary analyses executed on the first annotations, which will be used to train and evaluate our automatic system.

MOTS-CLÉS : Traitement automatique de la parole et du language, apprentissage de la lecture, IA pour l'éducation.

KEYWORDS: Natural language and speech processing, reading learning, AI for education.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 10–16.

# 1 Introduction

La maîtrise de la lecture est essentielle à l'autonomie de l'enfant, mais son apprentissage est un véritable défi. En France, les compétences en lecture et en compréhension des élèves de CM1 ont baissé depuis le début des années 2000. L'étude PIRLS de 2016 révèle que 40,5% des jeunes de 15 ans ne maîtrisent pas la lecture et que 21,5% d'entre eux rencontrent de sérieuses difficultés (Colmant & Le Cam, 2017). L'enquête PISA de 2015 a mis en évidence des écarts de performance importants entre les élèves les plus performants et les moins performants(OCDE, 2016). L'enquête PISA de 2018 a montré que les élèves socio-économiquement favorisés dépassent leurs pairs défavorisés d'environ quatre années de scolarité(OCDE, 2019). Enfin, un rapport de l'Observatoire national de la lecture de 2007 [1] a établi un lien entre les problèmes de lecture et les échecs dans diverses matières. Une solution à ces problèmes pourrait résider dans un apprentissage assisté par ordinateur (*Computer-Assisted Learning*, CAL) efficace pour aider les enseignants et les élèves.

L'entreprise Lalilo développe un tel système CAL : un assistant pédagogique basé sur l'intelligence artificielle (IA) pour différencier l'apprentissage de la lecture. Ce système offre une grande variété de tâches associatives pour couvrir les deux aspects clés de la lecture : la reconnaissance des mots et la compréhension du langage. Ce dernier aspect est toutefois plus difficile à traiter par les systèmes CAL en raison de la complexité des processus impliqués dans la compréhension. Les exercices à réponse ouverte, par opposition aux QCM traditionnels, peuvent aider les étudiants à s'exercer à des aspects plus complexes de la compréhension. Dans ce contexte, Lalilo souhaite proposer une nouvelle intervention CAL avec une activité de résumé oral d'un texte, des retours personnalisés pour les élèves et des métriques informatives pour les enseignants.

L'objectif du projet CHICA-AI est de développer le système automatique CAL qui analysera et évaluera les résumés oraux des élèves. Nous avons choisi d'évaluer les résumés non seulement à un niveau global, mais surtout à un niveau plus fin, comme le font les psychologues et les enseignants en classe. Dans cet objectif, nous avons conçu une grille d'évaluation basée sur les sciences psycho-cognitives et la pédagogie. Le système CAL visera à extraire les informations pertinentes du résumé oral de l'élève et à donner une note pour chaque critère de la grille. Les technologies suivantes seront utilisées pour extraire les informations : reconnaissance automatique de la parole (RAP), compréhension du langage parlé (CLP), traitement automatique du langage naturel (TALN).

Pour mener à bien cet objectif, nous devrons surmonter deux défis majeurs.

**Défi n°1 : Adaptation de technologies à un cas d'utilisation complexe de la vie réelle**

- Parole d'enfant : la performance des modèles de RAP sur la parole d'enfant est loin de la performance sur les adultes, et il n'y a pas de systèmes de CLP existants qui sont adaptés à la parole d'enfant ;
- Parole spontanée avec une charge mentale élevée : la tâche de résumé oral nécessite des mécanismes cognitifs complexes, ce qui introduit à des événements de parole spontanée dans le discours de l'enfant, tels que les disfluences, répétitions, auto-corrections, hésitations, etc. La présence de ces événements augmente la difficulté des tâches automatiques ;
- Utilisation en classe : la plateforme Lalilo est conçue pour une utilisation à l'école. Cela implique que les enregistrements contiennent une grande variété de bruits de classe, qui peuvent nuire à la performance des systèmes de RAP.

1. «La lecture au début du collège», téléchargeable : http://onl.inrp.fr/ONL/publications/publi2007/

**Défi n°2 : Construction d'une activité complexe d'entraînement de la compréhension**

Cette intervention CAL est inédite et doit être créée de toute pièce. Les aspects suivants devront être pris en compte pour garantir l'efficacité de l'intervention sur les progrès en lecture de l'élève :

1. Expérience psycho-cognitive et pédagogique
   (a) Evaluation du résumé : Comment évaluer un résumé oral ? Quels sont les critères importants et ceux qui le sont moins ? Comment déterminer si un critère a été satisfait ?
   (b) Remédiation à la suite de l'activité : Comment détecter les difficultés des élèves ? Lorsque ces difficultés sont détectées, de quoi l'élève a-t-il besoin pour les surmonter ?
2. Défis techniques
   (a) Annotation des données : Comment annoter les données pour entraîner les modèles à remplir la grille d'évaluation ? L'évaluation est très subjective (même chez les enseignants expérimentés), comment assurer la qualité et la représentativité des annotations ?
   (b) Remplir automatiquement la grille d'évaluation : Comment combiner les informations extraites automatiquement pour répondre à chaque critère de la grille d'évaluation ? Comment mesurer la performance du système ?

Nous présentons dans cet article le travail effectué depuis le début du projet CHICA-AI. La première étape (section 2), a été de créer l'activité de résumé oral dans la plateforme Lalilo, ce qui nous a permis de récolter des enregistrements audios de résumés d'élèves. La seconde étape (section 3) a été de mettre en place une tâche d'annotation et un protocole rigoureux pour assurer la bonne qualité et la pertinence des annotations pour l'entraînement et l'évaluation des systèmes automatiques. Nous présentons en section 4 les analyses préliminaires de notre processus d'annotation.

# 2 L'activité de résumé oral

## 2.1 Déroulement de l'activité

L'activité de résumé oral est déployée dans la plateforme Lalilo depuis mai 2024. Les élèves suivent une progression pédagogique définie par l'équipe pédagogique de Lalilo, et atteignent l'activité de résumé oral au niveau CM1-CM2.

L'activité de résumé oral est précédée par plusieurs leçons pour apprendre aux élèves à extraire les informations pertinentes d'un texte et à construire un résumé à partir de ces informations. Une fois ces leçons terminées, les élèves ont accès à l'activité, qu'ils peuvent effectuer (pour l'instant) 5 fois, avec 5 textes différents. Les textes sont des textes narratifs créés par l'équipe pédagogique de Lalilo avec une difficulté croissante. L'activité est composée de plusieurs phases :

1. Lecture du texte par l'élève ;
2. Réponse à des questions de compréhension sur le texte pour l'aider à extraire les informations pertinentes. Plus l'élève est avancé dans sa progression, moins il y a de questions ;
3. Enregistrement du résumé oral ;
4. Validation par l'enfant de son résumé.

Au terme du projet, l'activité comportera une cinquième étape, durant laquelle l'enfant recevra un retour personnalisé à l'aide du système d'analyse automatique. L'objectif est de l'aider à mettre en place des stratégies pour progresser en compréhension de la lecture.

## 2.2 Grille d'évaluation des résumés

Nous avons construit une grille à partir de critères pédagogiques et psycho-cognitifs pour évaluer de façon fine les résumés des élèves et trouver les meilleures stratégies de remédiation. Nous nous sommes basés sur la grille de (Casazza, 1993), que nous avons adaptée pour qu'elle corresponde à des textes narratifs niveau primaire et à du travail de compréhension de la lecture (Cèbe *et al.*, 2004).

Les critères sont divisés en deux catégories : stratégie de compréhension et mise en forme du résumé. La première catégorie contient des critères sur le contenu du résumé : les personnages principaux et secondaires, les lieux de l'histoire, les idées principales à retrouver et leur chronologie dans le récit de l'élève. La seconde catégorie permet d'évaluer la qualité de la forme du résumé : reformulation des idées, présence de connecteurs logiques et temporels, mise en oeuvre des compétences syntaxiques et sémantiques, etc.

# 3 Construction des annotations

## 3.1 Tâche d'annotation

La tâche d'annotation est réalisée sur une application développée par nos soins et fondée sur la bibliothèque python Streamlit[2]. Cette application permet de facilement traiter les enregistrements les uns après les autres et enregistre les annotations sous format json. Le processus d'annotation est composé de plusieurs phases.

La première phase consiste à écarter les enregistrements non annotables, qui ne seront pas utilisés dans le projet. En effet, les enfants étant en autonomie sur la plateforme Lalilo, il arrive souvent que les enregistrements reçus ne présentent pas de contenu analysable. Nous rejettons les enregistrements pour les raisons suivantes :

- NO_VOICE : L'élève ne parle pas
- ADULT : On entend seulement un adulte qui parle (souvent l'enseignant)
- OUT_OF_EXERCISE : L'élève parle mais ne fait pas l'exercice
- TOO_NOISY : L'enfant parle mais il y a trop de bruit pour comprendre ce qui est dit
- NOT_INTELLIGIBLE : L'enfant parle de façon non intelligible, on ne le comprend pas

La seconde phase, si l'enregistrement n'est pas rejeté, consiste à transcrire le résumé le plus fidèlement possible. Cette transcription sera utilisée pour entraîner les futurs systèmes de RAP et CLP, et servir de base textuelle pour les systèmes de TALN. Nous utilisons le système Whisper-large-v3[3] pour établir une transcription automatique à corriger. Des symboles sont ajoutés manuellement pour transcrire les pauses, hésitations, sons non verbaux et mots inintelligibles.

La troisième phase comporte les critères de la catégorie "stratégie de compréhension" de la grille. Les personnages, lieux et idées font l'objet de deux annotations : un score d'identification (1-4) et un repérage d'entités nommées (consistant à surligner des mots dans la transcription du résumé). La chronologie est évaluée avec un système de classement des idées principales identifiées par l'élève. L'annotation correspond à la chronologie telle qu'elle est dans l'esprit de l'élève, et non telle que présentée dans le résumé, c'est à dire en prenant compte de l'utilisation de connecteurs temporels (par exemple "avant ça", "plus tôt dans l'histoire"...). Cette phase se termine par un score global, entre

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

0 et 10, du contenu du résumé de l'élève, représentant sa compréhension de l'histoire.

La quatrième phase correspond aux critères de la catégorie "mise en forme du résumé" qui ne sont pas annotables automatiquement. Nous cherchons à détecter la présence d'avis personnels de l'enfant sur l'histoire (à proscrire dans un résumé) ou de parole hors de l'exercice (signe d'une déconcentration de l'élève). Nous demandons également à l'annotatrice de donner, entre 1 et 5, un score global de mise en forme du résumé et une score global d'expression orale.

## 3.2 Protocole d'annotation

En premier lieu, nous avons mené en 2024 une étude pilote à plus petite échelle sur les données collectées au cours des premiers mois sur la plateforme Lalilo. Dans cette étude préliminaire, une première version du guide d'annotation et de l'interface d'annotation a été conçue pour répondre à la tâche décrite ci-dessus. L'étude a été réalisée avec des étudiantes orthophonistes et a permis d'obtenir divers retours pour améliorer et corriger certains points de l'interface d'annotation et du guide. Après amélioration du protocole, une première réelle salve d'annotation a été réalisée en 2025 avec six autres étudiantes orthophonistes. Pour les former à cette tâche d'annotation très complexe, nous avons divisé le processus d'annotation en différentes étapes :

1. Formation initiale : présentation du protocole d'annotation (guide, interface) ;
2. Lot de formation (n°1) : Annotation d'un lot de 15 enregistrements choisis pour couvrir une diversité de cas (un exemple simple, moyen et complexe pour chaque histoire), chaque enregistrement étant annoté par toutes les annotatrices ;
3. Mesure de la qualité des annotations sur le lot 1 : Calcul de différentes mesures inter-annotateurs pour chaque domaine. L'objectif est de s'assurer que chaque annotatrice a bien compris les lignes directrices et fournit des annotations cohérentes et correctes.
4. Formation complémentaire et correction collaborative du lot 1
5. Lot de validation (n°2) : Annotation d'un lot de 45 audios répartis par paires d'annotatrices ;
6. Mesure de la qualité des annotations sur le lot 2 : Si les accords inter-annotateurs sont en dessous de nos seuils prédéfinis pour la validation, retour à l'étape 4. Dans le cas contraire, nous procédons à l'annotation du lot final.
7. Lot final : Le reste des données est réparti entre les annotatrices, en conservant 20% d'enregistrements en commun pour calculer des accords inter-annotatrices.

Pour estimer la qualité des annotations, nous avons utilisé plusieurs métriques différente en fonction de la nature du champ annoté. Les scores globaux sont évalués par le coefficient de corrélation intraclasse (*Intraclass Correlation Coefficient*, ICC) (Koch, 2006), et plus précisément l'ICC3 (modèle à deux facteur mixte). Les séquences (transcriptions et chronologies) sont comparées par le ratio de la distance de Levenshtein (au niveau du mot pour les transcriptions et au niveau de l'index de l'idée pour les chronologies). Les autres critères catégoriques sont évalués par le Kappa de Cohen.

# 4 Résultats

Le jeu de données obtenu contient au total 2085 annotations, dont 1065 ont été rejetés par les annotatrices. La majorité des rejets concernent soit des fichiers audio ne contenant aucune parole (481 annotations), soit des fichiers audio ne contenant que des paroles hors de l'exercice (364 annotations). Il en résulte 1020 annotations de 875 fichiers audio différents.

Les résultats des accords inter-annotateurs sont donnés dans la table 1. Les accords sur la transcription sont restés supérieurs au seuil de 0,85 pour tous les lots. Dans le lot 1, 2 et final, les accords sur les scores de compréhension (compréhension globale, personnages principaux, etc.), étaient au-dessus du seuil attendus pour l'ICC3 (0,6), et ont même dépassé 0,9, indiquant que les annotatrices ont bien compris ces critères. Pour la chronologie, les ratios sont plus bas, dû au fait qu'une idée non-reconnue par une annotatrice mais reconnue par une autre pour un même fichier fait légèrement baisser le score. Si on ne considère que les idées en commun, l'accords est bien plus élevé (>0,97). Les scores de mise en forme et d'expression globaux sont plus difficiles à annoter et plus variables. Après le lot 2, ils étaient au dessus du seuil acceptable (0,6), mais ont chuté à 0,50 et 0,55 pour le lot final. En dépit de cela, les accords peuvent être considérés comme globalement satisfaisants en raison de la complexité de l'annotation mise en œuvre, et ils suggèrent que le protocole pourra être appliqué à nouveau pour annoter les résumés oraux futurs de la plateforme.

TABLE 1 – Résultats des accords inter-annotateurs sur les lots d'entrainement 1 et 2 ainsi que sur les données communes du lot final.

| **Champ d'annotation** | **Métrique** | **Lot 1** | **Lot 2** | **Lot final** |
|---|---|---|---|---|
| Transcription | Levenshtein | 0,9102 | 0,8559 | 0,8781 |
| Compréhension globale | ICC3 | 0,6510 | 0,9089 | 0,8141 |
| Personnages principaux | ICC3 | 0,6775 | 0,9282 | 0,9306 |
| Idées principales | ICC3 | 0,7303 | 0,8092 | 0,8859 |
| Lieux | ICC3 | 0,9666 | 0,8922 | 0,8836 |
| Personnages secondaires | ICC3 | 0,7816 | 0,9666 | 0,8633 |
| Chronologie des idées | Levenshtein | 0,7776 | 0,8135 | 0,8551 |
| Mise en forme globale | ICC3 | 0,3173 | 0,7517 | 0,5092 |
| Expression globale | ICC3 | 0,6945 | 0,6708 | 0,5594 |

# 5 Conclusion

Nous avons présenté dans ce travail les objectifs et premiers résultats du projet CHICA-AI, visant à fournir un entraînement à la compréhension de la lecture aux élèves de primaire à travers une activité de résumé oral avec analyse automatique et retour personnalisé. Les premiers travaux effectués ont permis de récolter des enregistrements audios qui seront utilisés pour entraîner le système d'analyse automatique des résumés, ainsi que de les annoter pour les différentes tâches envisagées (RAP, CLP, TALN). Nos analyses montrent que notre protocole d'annotation, très sophistiqué en raison de la complexité de la tâche, permet d'obtenir des annotations de qualité satisfaisante pour la plupart des critères d'évaluation d'un résumé. Les chercheurs et chercheuses impliquées dans le projet pourront ainsi utiliser ces annotations pour entraîner et évaluer leurs systèmes, et ainsi construire une activité bénéfique dont l'usage entraîne une amélioration dans les capacités de compréhension des élèves.

# Remerciements

Les auteurs remercient l'Agence Nationale de la Recherche (ANR) pour le soutien financier apporté au projet CHICA-AI dans le cadre de l'Appel à Projet Générique 2023, ainsi que Renaissance Learning, qui finance le reste du projet et permet l'existence de la plateforme Lalilo.

# Références


CASAZZA M. E. (1993). Using a model of direct instruction to teach summary writing in a college reading class. *Journal of Reading*, **37**(3), 202–208.

CÈBE S., GOIGOUX R. & THOMAZET S. (2004). Enseigner la compréhension. Principes didactiques, exemples de tâches et d'activités. In *Lire écrire, un plaisir retrouvé*. MEN-DESCO. HAL : hal-00922482.

COLMANT M. & LE CAM M. (2017). PIRLS 2016. DOI : 10.48464/ni-17-24, HAL : halshs-03846903.

KOCH G. (2006). *Intraclass Correlation Coefficient*. DOI : 10.1002/0471667196.ess1275.pub2.

OCDE (2016). *Résultats du PISA 2015 (Volume I) : L'excellence et l'équité dans l'éducation*. PISA, Éditions OCDE. DOI : https ://doi.org/10.1787/9789264267534-fr.

OCDE (2019). *Résultats du PISA 2018 (Volume I) : Savoirs et savoir-faire des élèves*. PISA, Éditions OCDE. DOI : https ://doi.org/10.1787/ec30bc50-fr.