L’évaluation de la qualité des écrits d’enfants en TAL

BOUHOUCHE
Sami
(Numéro d’étudiant – 10468326)

DAVID
Nicolas Leewys
(Numéro d’étudiant – 11719192)

UFR LLASIC
Département des Sciences du Langage et Didactique des Langues

Master 2 Mention Sciences du Langage Parcours Industries de la Langue
UE : TAL et apprentissage des langues

Enseignant responsable : Monsieur Mathieu LOISEAU

Année universitaire
2018 – 2019

Introduction

S'inscrivant dans le domaine du Traitement Automatique des Langues (TAL) et de la linguistique de corpus, appliquée à l'étude d'écrits scolaires au sein du cycle de l'enseignement primaire en France, ce projet a pour ambition de mesurer la « qualité » des productions écrites du corpus LONGIT au moyen de différents indicateurs : lexique, diversité lexicale, structures syntaxiques, cohérence et cohésion du texte, orthographe, usage de la ponctuation… Il s'agirait par la suite de calculer certains de ces indicateurs, de permettre un calcul pour chaque texte et d'offrir une visualisation dynamique des résultats. Tel est l'objectif fixé par les enseignants chercheurs des laboratoires LaRAC et LIDILEM de l'Université Grenoble Alpes.

Nous commencerons donc par dresser un état de l'art pour implanter les principes et méthodes, tantôt traditionnels tantôt automatiques, qui régissent l'analyse et l'évaluation textuelles. Puis, nous présenterons le corpus LONGIT et les enjeux qu'il représente. Finalement, nous proposerons un système pour tenter de satisfaire à l'objectif fixé et de répondre à la problématique suivante : dans quelle mesure l'évaluation de la qualité des écrits d'enfants en TAL est-elle possible ?

1. Analyse et évaluation textuelles conventionnelles

Dans le domaine des *Arts, Lettres et Langues* ainsi que des *Sciences Humaines et Sociales*, de nombreuses disciplines se sont intéressées à l'analyse et l'évaluation textuelles. Il convient alors de dresser un état de l'art afin de déceler ces différents axes d'analyse et d'évaluation mais, ce travail ne peut guère s'avérer exhaustif. De ce fait, en nous appuyant sur la typologie établie par Calas (2011), nous dressons dans cette partie cet état de l'art selon les six axes suivants : paratexte, organisation textuelle, énonciation, lexique, grammaire et rhétorique.

1.1 Paratexte

Le paratexte constitue l'ensemble des éléments textuels et non textuels qui entourent un texte. Ces éléments fournissent toute une série d'informations et sont les principales auxiliaires nécessaires lors de l'entreprise d'une première lecture d'un texte, facilitant par la suite sa compréhension. Le paratexte est composé de deux sous-ensembles d'éléments : le péritexte et l'épitexte. Le péritexte inclut notamment les titres, les intertitres, les dédicaces, les épigraphes, les préfaces et les notes tandis que l'épitexte (moins usité que le péritexte) peut inclure des publicités et des présentations d'autres œuvres et ouvrages. Parmi tous ces éléments, les titres demeurent ceux étant les plus utiles car ils donnent des indications sur le thème du texte (titres thématiques) ou sur ce que renferme le texte (titres rhématiques) – quoique, par moments, le titre de certains textes est à la fois thématique et rhématique.

1.2 Organisation textuelle

Divers procédés sont utilisés pour construire et structurer un texte. Cette organisation textuelle prend forme à partir d'un bas niveau, celui de la typographie, pour s'étendre à un niveau de construction et d'harmonisation textuelle plus complexe.

1.2.1 Typographie

Considérée comme l'art d'utiliser les caractères pour agencer un texte, la typographie recèle plusieurs informations de base qui orientent vers la compréhension tout en permettant de déceler et de saisir l'enjeu d'un texte. La variation dans l'usage des majuscules et des minuscules, l'alternance entre caractères romains et italiques et l'emploi de caractères gras constituent des jeux typographiques de base, essentiels à l'organisation textuelle.

1.2.2 Ponctuation

La ponctuation est définie par Catach (1996, p.7) comme un :

> « Système de renfort de l'écriture, formé de signes syntaxiques, chargés d'organiser les rapports et la proportion des parties du discours et des pauses orales et écrites. Ces signes participent ainsi à toutes les fonctions de la syntaxe, grammaticales, intonatives et sémantiques. »

Ces « signes » fournissent alors des informations linguistiques qui véhiculent des :

- **Traits prosodiques :** tout comme l'intonation à l'oral, la ponctuation établit le rythme et la segmentation d'un texte

- **Traits syntaxiques :** la ponctuation contribue à la segmentation des phrases d'un texte, au balisage des constituants syntaxiques ainsi qu'à l'établissement de la modalité phrastique

- **Traits sémantico-énonciatifs :** les signes de ponctuation concourent au signalement des différents types de discours ainsi qu'aux changements thématiques et/ou énonciatifs

1.2.3 Volumétrie

L'appropriation de l'espace « blanc » d'une page est un indicateur volumétrique de la répartition et de la propagation textuelle. La volumétrie demeure une mesure exclusivement quantitative mais comme l'écriture est une forme d'expression du savoir humain, plus le volume textuel s'avère conséquent, plus l'expression de ce savoir s'extériorise, se concrétise et s'organise textuellement.

1.2.4 Cohésion, répétition et progression

Selon Adam (2011), un texte ne peut être réduit à une simple juxtaposition de phrases car toute unité textuelle minimale est régie par les trois grands principes suivants :

i. **Principe de cohésion :** un texte doit être un ensemble d'énoncés élémentaires liés

ii. **Principe de répétition :** un texte doit assurer la récurrence d'éléments énonciatifs

iii. **Principe de progression :** un texte doit être une suite progressive d'énoncés élémentaires

Dans le champ de la linguistique textuelle et de l'analyse du discours, la progression est un processus clé et dynamique dans l'organisation textuelle et Adam (*ibid.*, p.75-77) distingue les trois types de progression suivants :

i. **Progression thématique linéaire :** reprise systématique du rhème[1] d'une phrase ou d'un paragraphe dans la phrase ou dans le paragraphe qui suit

ii. **Progression thématique à thème constant :** le même thème[2] apparaît dans plusieurs phrases ou dans plusieurs paragraphes de manière successive

iii. **Progression thématique combinée / à thèmes dérivés :** mélange et alternance entre progression thématique linéaire et progression thématique à thème constant

Deux unités textuelles permettent aux principes de cohésion, de répétition et de progression d'être actualisés : le paragraphe et la phrase. Les paragraphes adoptent notamment un ensemble de modalités d'enchaînement tantôt anaphorique (reprise d'un référent déjà annoncé précédemment dans le texte) tantôt cataphorique (annonce d'un référent à venir dans le texte) pour assurer la progression thématique. Quant aux phrases, c'est par le biais élémentaire des connecteurs (logiques, inter-propositionnels et intra-propositionnels) qu'elles assurent les relations sémantico-logiques du texte – permettant par conséquent l'enchaînement de la trame énonciative et la progression thématique.

1.3 Énonciation

Selon Perret (2010, p.9), l'énonciation est « l'acte de parler, dans chacune de ses réalisations particulières, c'est-à-dire qu'est acte d'énonciation chaque acte de production d'un certain énoncé » et l'énoncé est différent de la phrase dans la mesure où « un énoncé doit avoir été dit ou écrit pour communiquer, alors qu'une phrase peut n'être qu'un exemple de grammaire, parfaitement abstrait et hors situation ». Pour Jakobson (1963, p.178-179), l'énonciation est une situation de communication « qui comprend les lieux, le temps de production des échanges entre les protagonistes ». Et grâce aux trois paramètres énonciatifs établis par Benveniste (1966, 1974) : le « moi » (*ego*), l' « ici » (*hic*) et le « maintenant » (*nunc*), l'énonciation accomplit selon Jakobson (*ibid.*) « une scène triangulaire au sein de laquelle la référence s'établit de manière spécifique à partir de ces trois bornes ».

1.3.1 Fonctions du langage

Tout texte modélise une situation de communication, d'énonciation, et doit ainsi exercer les six fonctions du langage suivantes (Jakobson, 1963) :

[1] Du grec *Rhèma* – « ce qui est dit » du thème : propos, commentaire, focus… sur le thème
[2] Du grec *Thèma* – « ce qui est posé » par le discours : topic, canevas, composition… du discours

i. **Fonction expressive :** le locuteur exprime ses sentiments, ses émotions, sa subjectivité et son investissement dans le discours

ii. **Fonction conative :** interpellation du récepteur / allocutaire pour attirer son attention et interagir avec lui

iii. **Fonction référentielle :** établissement du lien entre le message envoyé et le contexte (l'univers extralinguistique – le réel)

iv. **Fonction phatique :** maintien du contact entre le locuteur et l'allocutaire pour assurer la bonne circulation du message

v. **Fonction métalinguistique :** interrogation sur le code (le langage) utilisé et son fonctionnement

vi. **Fonction poétique :** mis en œuvre de divers procédés linguistiques pour accentuer l'essence du message

1.3.2 Discours et récit

Benveniste (1974) postule qu'un texte instaure soit un système de discours, soit un système de récit mais peut également alterner entre les deux systèmes. D'un côté, le système de discours actualise un instant d'énonciation pour mettre en exergue le contenu d'un texte : personnages, actions, évènements… De l'autre, le système de récit opère un certain effacement de la situation d'énonciation et procède à une mise en forme du texte au moyen de la temporalité des événements. Le tableau (*Tableau 1*) qui suit, repris de Calas (*ibid.*, p.38) présente cette dichotomie entre discours et récit.

Tableau 1 : Dichotomie entre discours et récit (Calas, 2011)

	Discours	**Récit**
Caractéristiques	Première personne Situation de communication visible Temps de base : le présent	Troisième personne Situation de communication masquée Temps de base : le passé simple
Événements	Insertion du récit et de ses caractéristiques	Type narratif Type descriptif
Paroles (pensées)	La parole directe domine, si c'est celle du locuteur ou du narrateur	Les paroles rapportées (Discours direct, indirect, indirect libre, narrativisé)

1.3.3 Paroles rapportées

Genette (1982) identifie quatre mécanismes pour traiter l'expression de la parole dans un texte :

i. **Discours direct :** par l'emploi de déictiques, marques typologiques (guillemets et tirets) et incises, et l'ancrage au sein de la situation d'énonciation, le mode du discours direct possède la spécificité d'être mimétique qui projette et isole les paroles sous forme de dialogues et/ou de conversations

ii. **Discours indirect :** le mode du discours indirect rapporte les paroles sous une forme synthétique et reformulée, et introduites par des verbes tels que *dire, raconter, déclarer, demander*…

iii. **Discours indirect libre :** le mode du discours indirect libre n'emploie guère d'indices grammaticaux et lexicaux pour mettre en relief les paroles et les intègre pleinement au texte, ce qui rend parfois leur identification complexe

iv. **Discours narrativisé :** le discours narrativisé opère un effacement quasi intégral des paroles et les réduisent à de simples verbes d'expression et de communication

Quand les paroles sont entièrement intériorisées et exprimées dans la tête d'un personnage, le texte prend la forme d'un psycho-récit et quand l'effacement du narrateur laisse place à la liberté d'expression d'un personnage, le texte prend la forme d'un monologue intérieur.

1.3.4 Polyphonie

Le dialogisme (la manifestation d'énoncés et de situations d'énonciation) est une particularité inhérente à tout texte et c'est cela qui le rend polyphonique selon les travaux de Bakhtine (1977). Un texte polyphonique comporte une pluralité (*poly*) de voix (*phonie*) instaurée par trois types d'acteurs distincts (Ducrot, 1984) : le sujet parlant, le locuteur et l'énonciateur.

1.3.5 Personnalisation

La personnalisation est une technique qui vise à circonscrire les différents participants / actants d'un texte au moyen de l'analyse d'éléments grammaticaux qui nourrissent l'énonciation, étant donné leur nature déictique : noms propres, appellatifs, déterminants et pronoms personnels.

1.3.6 Proximité et distance

Dans un texte, il est possible d'identifier une stratégie de proximité en observant l'emploi d' « une catégorie de déictiques indiquant les objets proches (voici, ceci), par opposition aux déictiques indiquant des objets éloignés (voilà, cela). » (Dubois *et al.*, 2012, p.389), ou une stratégie de distance que Dubois *et al.* (*ibid.*, p.154) définit ainsi :

> « Par certains mots, consciemment ou non, un locuteur peut laisser voir qu'il n'appartient pas, ou ne veut pas appartenir, ou n'a rien de commun avec le groupe ou les personnes dont il parle. Ces mots sont des marques de distance […]. On peut aussi parler de distance à propos du rapport que le locuteur veut établir non entre lui et autrui, mais entre lui et son discours. Plus la distance est grande, plus le discours est didactique. L'individu en tant que tel n'intervient pas dans les énoncés (disparition de tout ce que se réfère à lui personnellement, comme le pronom je, par exemple). »

1.4 Lexique

Dans un texte, les unités lexicales contribuent à fabriquer le sens sous 3 formes particulières :

i. **La dénotation :** unité lexicale qui donne un sens dénotatif à un élément en le désignant

ii. **La connotation :** unité lexicale qui en complément d'attribuer un sens dénotatif à un élément permet de suggérer des informations supplémentaires

iii. **La polysémie :** unité lexicale qui est soumise à une pluralité d'acceptions et qui véhicule divers sens

Ces unités lexicales entretiennent alors des relations sémantiques. Ce sont des relations d'inclusion qui établissent une hiérarchie entre les unités lexicales (relation d'hyponymie et d'hyperonymie), des relations d'équivalence qui les placent sur un même piédestal (relation de synonymie) ou encore, des relations d'opposition qui mettent au jour leur sens opposé (relation d'antonymie).

Ces relations sémantiques construisent des réseaux sémantiques, grâce à des champs lexicaux et des champs sémantiques (l'association de formes lexicales et de sens), pour au final établir l'isotopie d'un texte étant selon Eco (1985, p.131) « la constance d'un parcours de sens qu'un texte exhibe quand on le soumet à des règles de cohérence interprétative ».

Le lexique façonne également les registres de la langue employés dans un texte au sens des trois registres/styles de la rhétorique classique :

i. **Registre simple :** les unités lexicales employées sont d'ordre familier pour rendre l'expression du sens efficace et faciliter la compréhension

ii. **Registre moyen :** les unités lexicales les plus couramment employées et comprises par une communauté linguistique

iii. **Registre sublime :** les unités lexicales employées sont d'ordre soutenu pour afficher une certaine élégance au niveau de l'expression

Ces registres de langue font aussi ressortir différentes variations (socio)linguistiques auxquelles sont soumises les unités lexicales :

i. **Variation diatopique :** unités lexicales employées au sein d'un contexte géographique particulier

ii. **Variation diachronique :** évolution dans l'emploi et l'usage des unités lexicales au fil du temps

iii. **Variation diastratique :** unités lexicales employées en fonction d'un statut social et/ou de l'appartenance à un groupe social

1.5 Grammaire

Outre l'expression du bon usage et le respect des règles de la langue, les procédés grammaticaux sont essentiels dans toute construction textuelle pour assurer les fonctions suivantes :

i. **L'actualisation du rôle des déterminants :** la présence des déterminants est fondamentale afin de garantir l'expression des procédés lexicaux à l'exemple de la dénotation

ii. **L'actualisation du rôle des adverbes :** les adverbes agissent en tant que connecteurs et sont indispensables pour exprimer l'intensité

iii. **L'actualisation du rôle des verbes :** noyaux des phrases et des propositions, les verbes organisent et coordonnent le sens

iv. **La caractérisation lexicale :** procédé de mise en contexte et de désambiguïsation des unités lexicales

v. **La structuration phrastique :** les structures de phrase adoptées modulent les effets de rythme ainsi que les modalités d'énonciation

vi. **La concordance des temps verbaux :** la mise en œuvre d'un récit et/ou d'un discours et l'oscillation entre les deux reposent sur la concordance des temps verbaux

1.6 Rhétorique

Pour Calas (*ibid.*, p.69), la rhétorique, qui puise ses origines de la grecque antique au temps des grands orateurs se manifestant en public, est « un ensemble de règles permettant à l'orateur d'être efficace dans le choix de mots destinés à emporter l'adhésion de son interlocuteur ». Les procédés rhétoriques qui sont alors appliqués successivement pour construire un sujet et lui conférer une certaine validité (sujet ayant pour but ultime d'informer et de convaincre) sont les suivants :

i. **L'invention (*inventio*) :** il s'agit de trouver quoi dire

ii. **La disposition (*dispositio*) :** une fois l'invention trouvée, il est nécessaire de l'ordonner

iii. **L'élocution (*élocutio*) :** usage d'un registre sublime pour embellir les unités lexicales

iv. **L'action (*actio*) :** endosser le rôle d'un acteur pour jouer le discours et ainsi vivifier le sujet

v. **La mémoire (*memoria*) :** avoir recours à des procédés stéréotypés et efficaces usités dans le passé

2. Analyse et évaluation textuelles en TAL

Vu que l'un des principaux objectifs du TAL est de modéliser la langue écrite et parlée afin de procéder à son traitement automatique, divers dispositifs sont créés pour satisfaire à cet objectif. Or, au sein de cette élaboration de dispositifs, les phases d'analyse et d'évaluation figurent parmi les plus importantes et indispensables avant toute mise à disposition auprès d'un grand public. *Tokéniseur*, étiqueteur morphosyntaxique, analyseur syntaxique, synthétiseur vocal… n'y échappent pas. Le plus souvent, les données d'analyse et d'évaluation sont d'ordre textuel. Nous présentons ainsi dans cette partie un ensemble de métriques et de procédés non-exhaustifs qui contribuent à mener à bien cette analyse et cette évaluation.

2.1 Matrice de confusion

Bien souvent, l'accomplissement des tâches de classification à base de statistiques nécessite une matrice de confusion. C'est un tableau de contingence, à l'exemple du tableau suivant (*Tableau 2*), au sein duquel les colonnes représentent le nombre d'occurrences d'une classe estimée/prédite (la référence) et les lignes, le nombre d'occurrences d'une classe réelle/actuelle (la réalisation).

Tableau 2 : Exemple d'une matrice de confusion

		Prédiction	
		Positif	Négatif
Réalisation	Positif	Vrai positif (VP)	Faux positif (FP)
	Négatif	Faux négatif (FN)	Vrai négatif (VN)

2.2 *Accuracy*

Une fois la matrice de confusion réalisée, *l'accuracy* est la mesure qui calcule la performance globale d'un système en appliquant la formule qui suit :

$$Accuracy = ((VP + VN) / (VP + FP + FN + VN)) * 100$$

2.3 Précision

La précision est la mesure qui exprime le nombre d'éléments pertinents identifiés par un système par rapport au nombre total d'éléments identifiés et s'obtient ainsi :

$$Précision = (VP / (VP + FP)) * 100$$

2.4 Rappel

Le rappel mesure le nombre d'éléments pertinents identifiés par un système par rapport au nombre total d'éléments pertinents potentiels en procédant au calcul suivant :

$$Rappel = (VP / (VP + FN)) * 100$$

2.5 F-mesure

La F-mesure, aussi appelé F-score ou encore F_1, est la moyenne harmonique de la précision et du rappel. La F-mesure assure une pondération de la précision et du rappel et offre par conséquent une vision plus réaliste de ces deux mesures. Elle est calculée de la manière suivante :

$$F\text{-}mesure = 2 * ((Précision * Rappel) / (Précision + Rappel))$$

2.6 *Word Error Rate* (WER)

Dans le domaine de la reconnaissance automatique de la parole, le *Word Error Rate* (WER – Taux Erreur Mot en français), est une mesure utilisée pour évaluer une transcription automatique de parole par rapport à une transcription de référence. Le résultat est compris entre 0 et 100 (en termes de pourcentage) et se calcule au moyen de la formule suivante :

$$WER = ((I + D + S) / M) * 100$$

Où *I, D* et *S* sont le nombre d'insertions, de délétions et de substitutions observées dans la transcription automatique de parole et *M* étant le nombre total de mots contenus dans la transcription de référence.

2.7 Score BLEU

Le score BLEU (*bilingual evaluation understudy*) est une mesure utilisée dans le domaine de la traduction automatique pour évaluer la qualité d'un texte qui a été traduit automatiquement d'une langue naturelle A vers une langue naturelle B. L'algorithme de cette mesure se base sur un traitement à base de n-grammes et peut ainsi être utilisé pour évaluer un texte candidat à sa référence dans une même langue. Le score BLEU se situe entre 0 et 1 (plus le score est élevé, plus le texte candidat est proche du texte de référence) et s'obtient par l'application de la formule de base suivante :

$$Score\ BLEU = m / w_t$$

Où *m* est le nombre de mots similaires qu'on retrouve à la fois dans le candidat et la référence et w_t le nombre total de mots contenus dans le candidat.

2.8 *Latent Semantic Analysis* (LSA)

Avec l'immensité de données textuelles disponibles de nos jours, les techniques automatiques offrent de nombreux moyens de les classifier et de les modéliser en vue de leur exploitation. Le *Latent Semantic Analysis* (LSA – Analyse Sémantique Latente en français) figure parmi ces techniques qui facilitent cette classification et cette modélisation. Le LSA utilise un modèle de *Bag of Words* (BoW – sac de mots en français) qui produit une matrice terme-document (*document-term matrix*) révélant le nombre d'occurrences de termes dans un document. L'usage du LSA permet dans un premier temps de classifier un texte en lui attribuant une classe, un domaine : le procédé du *text classification*. Dans un deuxième temps, le LSA permet de découvrir automatiquement les thèmes cachés au sein d'un texte. Le modèle identifie les mots les plus saillants et fréquents du texte et leur attribue un score de

cohérence (*coherence score*) qui est compris entre 0 et 1. Ce procédé modélise les thèmes et les rhèmes d'un texte ainsi que le sujet qu'il véhicule : le procédé du *topic modelling*.

2.9 Fréquences lexicales

La linguistique de corpus a suscité un engouement considérable au niveau de la récolte massive de données pour constituer des corpus, notamment à l'écrit. Un certain nombre de choix préalables sont nécessaires pour traiter ces données et ces corpus, et l'ensemble de ces choix constitue selon Muller (1992) la norme de dépouillement. Cette norme se base sur une approche statistique et favorise l'observation de la répartition des fréquences d'occurrence d'unités lexicales au sein d'un corpus, essentiellement écrit. Cette approche d'étudier les fréquences lexicales révèle donc diverses informations utiles pour établir les propriétés lexico-métriques et lexico-syntaxiques d'un corpus écrit. Vocabulaire spécifique, expressions récurrentes, données texto-métriques… figurent parmi les propriétés les plus pertinentes à observer pour ainsi déterminer la richesse lexicale d'un texte : la taille du vocabulaire rapportée à la longueur, exprimant la diversité du vocabulaire employé.

2.10 Analyse automatique d'erreurs lexicales

Dans le domaine du TAL et de l'apprentissage des langues, les enseignants-chercheurs s'intéressent de plus en plus à l'analyse automatique d'erreurs lexicales qui se manifestent à l'écrit. Tout récemment, Wolfarth (2015) s'est intéressée aux apports du TAL dans l'élaboration et l'exploitation d'un grand corpus écrit, constitué à partir de productions écrites dans un cadre scolaire. Elle s'est appesantie sur la transcription de ces productions et de leur annotation automatique en terme d'erreurs. Elle a érigé une typologie des erreurs lexicales (au niveau des mots) pour aboutir à un schéma d'annotation automatique des erreurs. Cela démontre la possibilité de traiter automatiquement les erreurs lexicales à l'écrit et dans le cadre de son travail, les erreurs suivantes ont été traitées :

- **Séquences de segments**
  - Exemple : *Listoir / *ton bèt
- **Segments**
  - Exemple : chat / *fé / *i
- **Séquences de graphèmes**
  - Exemple : *apér / *fesa
- **Graphèmes**
  - Exemple : ch / a / t / s
- **Lettres**
  - Exemple : a, b, c
- **Signes graphiques**
  - Exemple : . / , / ‘

3. Présentation et traitement du corpus LONGIT

S'insérant dans l'élan du projet national *Lire et écrire au CP*[3], le corpus LONGIT, sous l'égide du Laboratoire de Recherche sur les Apprentissages en Contexte (LaRAC)[4] de l'Université Grenoble Alpes, se constitue encore actuellement dans le cadre du projet

[3] *cf.* http://ife.ens-lyon.fr/ife/recherche/lire-ecrire
[44] LaRAC : https://larac.univ-grenoble-alpes.fr/

CORPUSCOL[5]. Le projet CORPUSCOL est soutenu par le pôle *Initiatives de Recherche Stratégiques* de l'*Initiative d'Excellence* (IRS – IDEX) de la Communauté Université Grenoble Alpes et arbore l'ambitieux objectif de :

> « […] décrire les performances réelles des élèves de l'école élémentaire dans le domaine de la production d'écrit et leur progressive construction, et à proposer un outil évolutif pour aider les enseignants à évaluer les productions écrites. Il ambitionne donc à terme de soutenir la formation des enseignants, pour un accompagnement exigeant et explicite de la production d'écrit et de son évaluation (Bautier, 2016). Le moyen en est la publication d'un vaste corpus de référence d'écrits d'apprenants avec mise à disposition d'un outil d'interrogation de ce corpus et la mise au point d'une grille évolutive d'évaluation des textes narratifs. »

Le corpus est pour l'instant constitué de productions d'écrits d'enfants du CE1 et du CE2. La méthodologie adoptée par l'équipe du LaRAC pour le constituer, dans un contexte scolaire, est la suivante :

- L'enseignant(e) lit une histoire aux écoliers en début de matinée sans y apporter de commentaires
- Après la récréation du matin, il leur est demandé de reproduire l'histoire à l'écrit
- L'enseignant(e) ne doit pas aider les élèves mais les rassurer si nécessaire, en leur précisant bien que ce qui est recherché est « comment ils peuvent raconter une histoire qu'ils ont entendu une seule fois »
- L'enseignant(e) précise aux écoliers que cet exercice de production écrite ne sera pas évalué, jugé ou noté
- L'enseignant(e) est invité(e) à encourager les élèves à écrire ce dont ils se souviennent et comme ils s'en souviennent
- Cet exercice de production écrite dure 20 minutes

Après cette phase de collecte, les phases qui suivent sont la numérisation, la transcription et la normalisation des productions écrites. Dans le cadre de notre travail, nous avons eu accès à un échantillon brut de 186 productions du corpus LONGIT. Chaque production est composée de sa transcription et de sa normalisation en complément de la classe et de l'identifiant de l'élève-scripteur qui font office de métadonnées. Pour mener à bien notre travail, nous avons écarté de notre traitement analytique les transcriptions pour exploiter pleinement les normalisations.

4. Développement et spécificités du PCPA

Après le nettoyage de l'échantillon brut, les 186 productions ont été isolées dans un répertoire à partir duquel nous avons appliqué la chaîne de traitement automatique suivante :

[5] Projet CORPUSCOL : http://scoledit.org/corpuscol/

i. Étiquetage morphosyntaxique des productions

ii. Calcul lexico-métrique, à partir de l'étiquetage morphosyntaxique, des indices suivants :

- le nombre d'occurrences de formes/tokens (mots pleins et ponctuations)
- le nombre d'occurrences de mots
- le nombre d'occurrences de hapax (mots à occurrence unique)
- le nombre de caractères
- le nombre de caractères par mot (longueur moyenne des mots)
- le nombre de phrases
- le nombre de mots par phrase (longueur moyenne des phrases)
- le nombre de types de mots et leur fréquence
- le nombre d'occurrences de chaque catégorie grammaticale
- la liste des lemmes catégories triées par fréquence décroissante

iii. Calcul des métriques d'évaluation suivantes :

- Richesse lexicale (*Type-Token Ratio*)
- Score BLEU
- Précision
- Rappel
- F-mesure

iv. *Topic modelling* basé sur l'intégralité des 186 productions

Cette chaîne de traitement s'effectue au moyen d'un système (programme) que nous avons baptisé le PCPA : *Python Corpus Processor and Analyzer*. Le langage de programmation Python 3 a été utilisé pour l'écriture du code du PCPA. Le module *treetaggerwrapper* a permis l'usage de l'étiqueteur morphosyntaxique *TreeTagger* pour l'étiquetage morphosyntaxique des productions. Nous avons entièrement écrit la partie du code concernant le calcul lexico-métrique ainsi que la mesure de la richesse lexicale. Le score BLEU, la précision, le rappel, et la F-mesure ont été calculés et obtenus grâce à la librairie *Natural Language Toolkit* (NLTK) pour Python. Le texte lu en classe a été pris comme référence et chaque production écrite comme réalisation. Pour le *topic modelling*, nous avons utilisé le modèle LSA inclus dans le module *Gensim* pour Python en exploitant et réadaptant un script Python créé par Avinash Navlani et disponible sur *DataCamp*[6].

5. Exemplification des résultats

Après l'exécution du PCPA, un fichier individuel est généré pour chaque production pour illustrer les résultats de l'analyse comme l'exemplifie le tableau suivant (*Tableau 3*) :

Tableau 3 : Illustration des données issues de l'analyse d'une production

Les données suivantes sont issues de l'analyse de la production : prod_CE1_92

[6] *Latent Semantic Analysis using Python :* https://www.datacamp.com/community/tutorials/discovering-hidden-topics-python

Cette production contient :

127 forme(s)/token(s)

116 mot(s)

36 hapax

524 caractère(s)

Une moyenne de 5 caractère(s) par mot

2 phrase(s)

Une moyenne de 58 mot(s) par phrase

60 type(s) de mots dont la distribution fréquentielle est la suivante :

Le : 4
lundi : 1
Léon : 1
doit : 1
faire : 1
une : 2
rédaction : 6
il : 11
cherche : 4
mange : 2
le : 8
soir : 4
maman : 1
se : 1
coupe : 1
doigt : 1
avec : 2
un : 4
couteau : 2
lendemain : 3
raconte : 1
sa : 4
poussa : 1
soeur : 1
dans : 1
les : 1
escaliers : 2
elle : 1
va : 1
à : 1
l' : 1
hôpital : 2
raconta : 2
a : 1
met : 2
vaisselle : 1
du : 1
vinaigre : 2
cela : 1
marche : 1
très : 1
bien : 2

posa : 1
pot : 3
matin : 1
tombe : 1
sur : 1
passant : 2
y : 1
avait : 1
remplaçante : 3
la : 2
dit : 1
aux : 1
enfants : 2
maîtresse : 1
s' : 1
est : 1
ouvert : 1
crâne : 1

Liste (triée par fréquence décroissante) du nombre d'occurrences de chaque catégorie :

('NOM', 40)
('DET:ART', 22)
('VER:pres', 16)
('PRO:PER', 15)
('PRP', 5)
('DET:POS', 4)
('VER:simp', 4)
('NUM', 3)
('PUN', 3)
('VER:ppre', 3)
('ADV', 3)
('PRP:det', 2)
('SENT', 2)
('NAM', 1)
('VER:infi', 1)
('PRO:DEM', 1)
('VER:impf', 1)
('VER:pper', 1)

Liste (triée par fréquence décroissante) des lemmes et de leurs catégories :

(('DET:ART', 'le'), 16)
(('PRO:PER', 'il'), 11)
(('DET:ART', 'un'), 6)
(('NOM', 'rédaction'), 6)
(('VER:pres', 'chercher'), 4)
(('NOM', 'soir'), 4)
(('DET:POS', 'son'), 4)
(('NOM', 'lendemain'), 3)
(('NUM', '@card@'), 3)
(('PUN', '/'), 3)
(('VER:ppre', '@ord@'), 3)
(('NOM', 'pot'), 3)
(('NOM', 'remplaçant'), 3)
(('VER:pres', 'manger'), 2)
(('PRO:PER', 'se'), 2)
(('PRP', 'avec'), 2)
(('NOM', 'couteau'), 2)
(('NOM', 'escalier'), 2)

```
                (('NOM', 'hôpital'), 2)
                (('VER:simp', 'raconter'), 2)
                (('VER:pres', 'mettre'), 2)
                (('NOM', 'vinaigre'), 2)
                (('ADV', 'bien'), 2)
                (('NOM', 'passant'), 2)
                (('NOM', 'enfant'), 2)
                (('SENT', '.'), 2)
                (('NOM', 'lundi'), 1)
                (('NAM', 'Léon'), 1)
                (('VER:pres', 'devoir'), 1)
                (('VER:infi', 'faire'), 1)
                (('NOM', 'maman'), 1)
                (('VER:pres', 'couper'), 1)
                (('NOM', 'doigt'), 1)
                (('VER:pres', 'raconter'), 1)
                (('VER:simp', 'pousser'), 1)
                (('NOM', 'sœur'), 1)
                (('PRP', 'dans'), 1)
                (('PRO:PER', 'elle'), 1)
                (('VER:pres', 'aller'), 1)
                (('PRP', 'à'), 1)
                (('VER:pres', 'avoir'), 1)
                (('NOM', 'vaisselle'), 1)
                (('PRP:det', 'du'), 1)
                (('PRO:DEM', 'cela'), 1)
                (('NOM', 'marche'), 1)
                (('ADV', 'très'), 1)
                (('VER:simp', 'poser'), 1)
                (('NOM', 'matin'), 1)
                (('VER:pres', 'tomber'), 1)
                (('PRP', 'sur'), 1)
                (('PRO:PER', 'y'), 1)
                (('VER:impf', 'avoir'), 1)
                (('VER:pres', 'dire'), 1)
                (('PRP:det', 'au'), 1)
                (('NOM', 'maître|maîtresse'), 1)
                (('VER:pres', 'être'), 1)
                (('VER:pper', 'ouvrir'), 1)
                (('NOM', 'crâne'), 1)

        Richesse lexicale (Type-Token Ratio) : 0.5172413793103449

        Score BLEU : 0.057669557867685864

        Précision : 0.5076923076923077

        Rappel : 0.171875

        F-mesure : 0.25680933852140075
```

Un fichier pour le *topic modelling* est également généré et contient les 100 mots les plus saillants de l'ensemble des 186 productions. Chaque mot possède un score de cohérence et une observation minutieuse conduit à inférer le système de personnages construit autour d'un personnage principal (0.260*"léon") et de personnages secondaires (0.111*"soeur" / 0.085*"maîtresse" / 0.054*"maman" / 0.037*"garçon" / 0.028*"parent" / 0.047*"dame" / 0.021*"père"), le cadre spatio-temporel (0.363*"dimanche" / 0.180*"lundi" / 0.071*"lendemain" / 0.060*"hôpital" / 0.050*"école" / 0.027*"matin" / 0.022*"toilette") et

l'intrigue global : l'histoire de Léon qui a des rédactions mais qui ne fait rien (0.557*"a" / 0.260*"léon" / 0.225*"rédaction" / 0.177*"fait" / 0.176*"rien").

6. Discussion et perspectives d'amélioration

Au terme de notre travail, il serait erroné d'inférer que nous soyons parvenus à satisfaire pleinement l'objectif fixé en préambule. Nous avons certes pu fournir une analyse lexicale (lexico-métrique) et sémantique (LSA – *topic modelling*) pour chacune des 186 productions de l'échantillon du corpus LONGIT. Elle a été agrémentée de diverses métriques pour faire ressortir la diversité (richesse lexicale – TTR) et la fidélité de restitution de chaque production par rapport au texte de référence (score BLEU, précision, rappel, et F-mesure).

Or cette analyse demeure purement centrée sur des statistiques. Dans une étape ultérieure, il serait bien de l'orienter davantage vers l'étude des structures syntaxiques, de l'orthographe et de la ponctuation. Cela permettrait de tendre plus vers une approche d'analyse automatique d'erreurs lexicales et de l'apprentissage de l'écrit, à l'exemple du travail mené par Wolfarth. Au niveau de la sémantique, nous sommes plutôt satisfaits des résultats issus du *topic modelling* qui offrent de multiples informations sur la cohérence et cohésion globales du corpus par rapport au texte de référence. Toutefois, cette méthode s'avère une fois de plus régie par des statistiques. Il conviendrait donc de l'associer aux méthodes d'analyse et d'évaluation textuelles conventionnelles afin d'offrir une représentation plus complète et aboutie de la cohérence et de la cohésion de chaque production.

Au final et en dépit des divers atouts qui contribuent à la robustesse du PCPA, force est de constater que certains aspects, tantôt intrinsèques tantôt extrinsèques, lui imposent certaines limites analytiques et fonctionnelles. Les éventuelles améliorations du système pourraient ainsi être menées dans le sens à résoudre partiellement voire complètement les contraintes suivantes :

- Offrir une visualisation dynamique des résultats
- Lier cette évaluation de la qualité à l'étude de l'évolution des écrits – quand l'élève passe d'une classe à une autre
- Exploiter les balises contenues dans les normalisations pour rendre compte de la structure et du paratexte de chaque production
- Exploiter les transcriptions en vue du traitement et de l'analyse automatique des erreurs lexicales qu'elles contiennent afin de tendre vers une analyse plus pointue des structures syntaxiques, de l'orthographe et de la ponctuation
- Contrer les limites imposées par *TreeTagger* afin de mieux gérer l'identification des phrases (limitée pour l'instant aux nombres d'occurrence de l'étiquette « SENT »), les expressions poly-lexicales et référentielles, et la désambiguïsation
- Attribuer une portée plus pédagogique et didactique au PCPA pour le rendre plus pertinent auprès d'un corps enseignant et faire de ce système « un outil évolutif pour aider les enseignants à évaluer les productions écrites »

Conclusions

Ayant inscrit notre travail d'évaluation de la qualité des écrits d'enfants dans le domaine du TAL et de la linguistique de corpus, tout en exploitant les apports du domaine des *Arts, Lettres et Langues* ainsi que des *Sciences Humaines et Sociales*, nous sommes parvenus à adopter une démarche cohérente et progressive afin de mener à bien l'élaboration d'un système d'analyse et d'évaluation. Nous avons commencé par dresser un état de l'art des principes et méthodes, traditionnels et automatiques, d'analyse et d'évaluation textuelles. Puis, nous avons présenté le corpus LONGIT dont les enjeux ont façonné le développement de notre système : le PCPA. Finalement, et après avoir exemplifié les résultats produits par le PCPA, nous avons mis en exergue ses atouts et avons proposé quelques améliorations qui pourraient y être apportées. Certes, notre système n'est qu'à ses premiers balbutiements et mérite d'être amélioré et « robustifié ». Néanmoins, il demeure fonctionnel et pourrait servir d'esquisse à une version plus avancée et dotée d'une plus grande complétude.

Références bibliographiques

Adam, J.-M. (2011). *La linguistique textuelle.* Paris : Armand Colin.

Bakhtine, M. (1977). *Le marxisme et la philosophie du langage.* Paris : Les Éditions de Minuit.

Benveniste, É. (1966). *Problèmes de linguistique générale Tome I.* Paris : Gallimard.

Benveniste, É. (1974). *Problèmes de linguistique générale Tome II.* Paris : Gallimard.

Calas, F. (2011). *Leçons de stylistique.* Paris : Armand Colin.

Catach, N. (1996). *La ponctuation.* Paris : Presses Universitaires de France (Que sais-je ?).

Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., Marcellesi, J.-B. et Mével, J.-P. (2012). *Le dictionnaire de linguistique et des sciences du langage.* Paris : Larousse.

Ducrot, O. (1984). *Le dire et le dit.* Paris : Les Éditions de Minuit.

Eco, U. (1985). *Lector in fabula.* Paris : Grasset.

Genette, G. (1982). *Palimpsestes.* Paris : Le Seuil.

Jakobson, R. (1963). *Essais de linguistique générale.* Paris : Les Éditions de Minuit.

Perret, M. (2010). *L'Énonciation en grammaire de texte.* Paris : Armand Colin.

Poibeau, T. (2017). *Machine Translation.* Massachusetts : The MIT Press.

Wolfarth, C. (2015). *Apport du TAL à la constitution et l'exploitation d'un corpus scolaire de cours préparatoire. Linguistique.* <dumas-01167286>