Exploration empirique de la richesse lexicale: la perception humaine

Audrey Bonvin et Amelia Lambelet (Fribourg)

Abstract

We investigate how untrained readers perceive the lexical richness of short texts written by 8 to 10-year-olds in French, German and Portuguese. Seven untrained adults rated the lexical richness of a set of argumentative and narrative texts on Likert Scales while performing on a Think Aloud Protocol (TAP) task. Results show that raters consider other criteria than solely vocabulary in their assessments. Coherence, grammar and content seem to have a particular impact. Concerning vocabulary per se, (non-)repetition is mentioned regularly and other characteristics such as register, length of words, number of words appear as well. For this kind of texts, word frequency does not seem to be as salient as expected.

1 La richesse lexicale[1]

Le concept de richesse lexicale (ou de richesse du vocabulaire dans son acceptation scolaire) est une notion courante. On admire une œuvre littéraire dont le vocabulaire est évocateur, ou le discours d'une personne qui contient un vocabulaire riche. Les enseignants utilisent cette notion dans leurs évaluations des productions orales ou écrites des élèves, qu'il s'agisse de la langue de scolarisation ou d'une langue étrangère. En termes de recherche scientifique, la richesse lexicale est tout aussi bien étudiée en lexicographie sur de longs textes tels que des textes littéraires ou des débats politiques (cf. Hubert/Labbé 1988) que sur de plus courts textes, dans le domaine de l'acquisition/apprentissage des langues et par les chercheurs qui s'intéressent à l'attrition L1 dans le contexte de migration (Schmid/Jarvis 2014) ou encore à certains troubles du langage tels que l'aphasie (Fergadiotis et al. 2013). La richesse lexicale est généralement mesurée à l'aide de formules computables variant selon les études. Dans ce travail, cependant, nous interrogeons la perception humaine de la richesse du vocabulaire.

[1] Ce travail constitue une partie du projet «Productions écrites d'enfants issus de la migration» de l'Institut de plurilinguisme (Université de Fribourg et Haute Ecole Pédagogique de Fribourg). Nous remercions Thomas Aeppli pour sa collaboration et les échanges intéressants pendant le processus d'analyse. Merci aussi à Raphael Berthele et Jan Vanhove pour leur soutien durant l'intégralité du projet, ainsi qu'à nos sept informateurs sans lesquels cet article n'aurait pas pu voir le jour.

1.1 Une définition de la richesse lexicale

Si la richesse lexicale est étudiée depuis au moins la première moitié du 20e siècle, elle constitue toujours un concept complexe. En effet, d'une part sa terminologie et ce qu'elle recouvre diffère entre les chercheurs, et, d'autre part, il n'existe pas de mesure automatique de richesse lexicale valide universellement, c'est à dire prenant en compte les différentes composantes que recouvre la notion de richesse lexicale et permettant de comparer toutes sortes de textes. De fait, deux défis majeurs apparaissent dans la littérature quant à l'opérationnalisation du concept de richesse lexicale. Tout d'abord, les algorithmes développés pour la mesurer automatiquement souffrent d'une influence de la longueur des textes, ce qui rend difficile la comparaison de textes de différentes longueurs. Ensuite, les caractéristiques morphosyntaxiques variant d'une langue à l'autre, la comparaison directe de la richesse lexicale de textes rédigés dans des langues différentes est fortement déconseillée.

Dans cet article, nous explorons la richesse lexicale en la considérant comme un terme générique regroupant un certain nombre de mesures variant selon les auteurs (cf. Jarvis 2017; Lindqvist et al. 2013; Read 2000) : la **diversité lexicale** est la mesure à la fois la plus connue et la plus ancienne de richesse lexicale. Elle se base sur le principe de (non-)répétition de vocables (*type* en anglais), l'idée sous-jacente étant qu'une répétition de mêmes mots (autrement dit, la présence de plusieurs occurrences du même vocable) est le signe de compétences lexicales basses (ou inversement, qu'une variation du choix des vocables reflète de bonnes compétences) (cf. Daller 1999: 121). Il s'agit donc d'une mesure attribuant la même valeur à chaque vocable (cf. Daller et al. 2003: 201–203). Une première mesure, très intuitive, de la diversité lexicale est le *type-token ratio* (TTR), soit le nombre de vocables divisé par le nombre d'occurrences (*token* en anglais). S'il est intuitivement utile, le TTR souffre d'une grande sensibilité à la taille des textes, ce qui le rend peu utile lors de comparaisons de textes de différentes longueurs. Plus un texte est long, moins il est probable que de nouveaux mots surviennent et donc, plus les chances de répétitions de mots augmentent (cf. McCarthy/Jarvis 2007 : 460) car le nombre d'occurrences augmente plus rapidement que le nombre de vocables. Pour contrer ce problème, les dernières décennies ont vu le développement de nombreuses autres mesures, moins intuitives, mais plus résistantes aux variations de taille des textes analysés.

Une autre composante de la richesse lexicale est la **sophistication lexicale,** définie comme l'utilisation de mots avancés (traduction de l'anglais *advanced*), difficiles ou rares. En ce qui concerne les mots avancés ou difficiles, on remarque rapidement qu'il n'existe pas de consensus universel sur leur définition. Celle-ci dépend du chercheur ainsi que du groupe cible étudié (p.ex. en fonction l'âge des sujets) (cf. Laufer 1994 : 22). Le plus souvent, c'est la rareté des mots, mesurée selon leur fréquence dans des corpus représentatifs de la langue en question, qui détermine le degré de sophistication des textes. Ainsi, intuitivement, le mot *houspiller* est moins fréquent que *gronder* et est donc plus sophistiqué.

La troisième dimension de la richesse lexicale que nous considérons ici est la **densité lexicale**, c'est à dire la proportion de vocables de contenu (substantifs, verbes, adjectifs et adverbes) dans le texte par rapport aux vocables fonctionnels (prépositions, articles, etc.) (cf. Lindqvist et al. 2013 : 109 ; Ure 1971, cité dans Johansson 2008: 65). Globalement, la densité lexicale nous

informe du degré d'oralité d'un texte, les registres académiques étant en effet entre autres caractérisés par une haute proportion de mots de contenu (cf. Henrichs/Schoonen 2009 pour une discussion).

Enfin, certaines suites de mots ont tendance à apparaitre conjointement dans une langue donnée. Ces suites de mots sont communément appelées **collocations**, soit « des associations conventionnelles de mots, arbitraires et récurrentes, dont les éléments ne sont pas nécessairement contigus et dont la signification est largement transparente » (Nerima et al. 2006 : 96–97). Des exemples-types de collocations en français sont « gros fumeur » (adjectif + nom), « caresser l'espoir » ou encore « exercer une profession » (verbe + nom) (Nerima et al. 2006). Selon divers chercheurs en acquisition des langues secondes (cf. Henriksen 2013; Levitzky-Aviad/Laufer 2013), une bonne maitrise des collocations reflète une compétence lexicale élevée. Finalement, le concept de richesse lexicale inclut également la **proportion d'erreurs lexicales**, en tant que mots utilisés de manière erronée, et l'**originalité lexicale**, soit la proportion de mots utilisés par seulement une personne dans un corpus de textes produits par plusieurs auteurs et considérés comme comparables à un moment donné.

Le développement des statistiques lexicales a permis une avancée remarquable en termes de rapidité et de comparabilité de l'évaluation de la richesse lexicale de textes oraux et écrits. Cependant, les différentes dimensions de la richesse lexicale peuvent chacune être mesurées par divers algorithmes. Comme nous le discuterons dans la suite de cet article, si elles sont de plus en plus élaborées, ces mesures s'avèrent aussi de plus en plus complexes et de moins en moins transparentes quant au concept étudié.

1.2 L'évaluation humaine et non entraînée de la richesse du vocabulaire

À toujours chercher à valider de nouvelles mesures, on oublie parfois de se demander si celles-ci reflètent vraiment ce que tout être humain, ou peut-être plus particulièrement tout enseignant, conçoit par un lexique riche ou varié (cf. Jarvis 2013a). En raison de cette complexité du champ, on peut en effet se retrouver éloignés conceptuellement de l'objet d'étude, utilisant des formules mathématiques plus ou moins adaptées aux textes à évaluer et ne sachant plus précisément ce qui est effectivement évalué. D'une manière plus générale, il est intéressant de se demander si derrière le terme de richesse du vocabulaire se retrouve une perception commune du concept, en d'autres termes si les locuteurs (d'une langue donnée) ont une même vision des éléments permettant à un texte d'être qualifié de riche.

A notre connaissance, peu d'études ont focalisé l'évaluation humaine de la richesse lexicale (en dehors de Vanhove et al. 2019 ; Bonvin/Lambelet 2017; Jarvis 2013a 2013b 2017;). L'étude fondatrice sur cette question est constituée de récoltes de données avec quatre cohortes successives d'étudiants non entrainés auxquels il a été demandé d'évaluer la richesse du vocabulaire de textes en anglais préalablement corrigés (cf. Jarvis 2013a 2013b 2017). Jarvis a ensuite modélisé ces évaluations avec différentes mesures de richesse lexicale dans le but de mettre au jour les dimensions influençant la perception subjective. Jarvis conclut sur la base de ces différentes études que même si la perception de la qualité globale des textes et la perception de la diversité lexicale ont une forte relation, il s'agit néanmoins de deux construits différents.

Dans l'étude présentée ici, nous prenons une perspective plus qualitative pour l'exploration du même phénomène. Pour ce faire, nous avons enregistré les réflexions d'adultes non entrainés à qui nous avons demandé de juger de la richesse du vocabulaire d'une sélection de textes issus du corpus HELASCOT, un corpus composé de plus 3500 textes argumentatifs et narratifs en français, allemand et portugais produits par des enfants de 8 à 10 ans et récoltés de manière longitudinale (cf. Berthele/Lambelet 2017).

Plus précisément, notre but est de répondre aux questions de recherche suivantes :

1. Lorsqu'il leur est demandé d'évaluer la richesse du vocabulaire des textes de notre corpus en verbalisant leurs réflexions à haute voix, quels sont les critères dont les évaluateurs parlent le plus souvent et quel rôle ceux-ci jouent-ils dans l'attribution des scores ?
2. Lesquels d'entre eux sont liés au vocabulaire ?
3. Les critères liés au vocabulaire se rapprochent-ils des composantes de la richesse lexicale apparaissant dans la littérature (cf. 1.1) ?

Nous posons comme hypothèse que les évaluations non entrainées de la richesse du vocabulaire sont liées à la fois à des facteurs endogènes au texte (facteurs lexicaux, mais également concernant la qualité globale du texte), et à des facteurs exogènes tels que le profil de l'évaluateur. Au sein des critères liés au lexique, nous postulons par ailleurs que les évaluateurs se concentrent particulièrement sur les mots lexicalement pleins, car ce sont souvent ces mots qui ont été entrainés à l'école sous forme de recherche de synonymes ou de distinction entre vocabulaire familier et soutenu.

2 Méthode

La méthode choisie pour cette étude est celle du protocole de pensée à haute voix (ou protocole verbal, en anglais *think aloud protocol*), (cf. Ericsson/Simon 1980), suivi d'un court entretien rétrospectif (cf. Kuusela/Paul 2000 : 390).

2.1 Participants

Nous avons choisi un échantillon relativement hétérogène dans le but de refléter différentes visions de la diversité lexicale. La méthode de pensée à haute voix étant difficile à implémenter selon la population, nous avons par ailleurs choisi des participants avec un niveau d'éducation universitaire (et donc plus familiers avec une situation de recherche), pour augmenter les chances d'obtenir assez de segments signifiants pour effectuer les analyses.

Cinq femmes et deux hommes, âgés de 24 à 31 ans, ont participé à l'étude.[2] Quatre d'entre eux sont issus du domaine des langues (TAP2, 4, 5, 6), tandis que les trois autres proviennent des domaines de la psychologie, chimie et génie civile. Deux participants (TAP4 et TAP6) avaient déjà été confrontés à l'évaluation de textes en général, mais sans focus particulier sur la richesse lexicale.

[2] Dans ce document, nous employons uniquement le genre masculin pour des raisons d'anonymat et aussi parce que le genre de l'évaluateur ne s'est pas avéré apporter des informations utiles pour répondre aux questions de recherche.

2.2 Procédure

Les sept participants ont été confrontés à des paires de textes (une lettre argumentative et un récit narratif) issus de la troisième récolte de données du projet HELASCOT et rédigés par des élèves d'origine portugaise. Après la lecture de chaque paire de textes, ils ont procédé à leur évaluation tout en verbalisant leurs réflexions à haute voix (cf. Annexe 1). Les textes à évaluer ont été corrigés au niveau orthographique, mais pas au niveau de la ponctuation ou de la grammaire (en d'autres termes, le mot doit exister ainsi écrit dans la langue en question, mais peut être mal accordé ou mal utilisé). Nous avons varié l'ordre des textes entre les participants pour éviter des effets de proximité. Pour rendre la tâche plus agréable, un environnement favorable et informel a été créé pour la passation et les instructions ont été rédigées en insistant sur le fait qu'il n'y avait pas de réponses justes ou fausses. Selon la langue dominante des évaluateurs, ceux-ci ont évalués la richesse lexicale des textes en allemand, en français ou en portugais. L'évaluateur des textes en portugais a parlé en français afin que les chercheuses non lusophones puissent analyser les enregistrements. La passation a consisté en trois étapes :

1. Pour commencer, chaque évaluateur a estimé de façon holistique la richesse du vocabulaire des paires de textes sur des échelles de Lickert de 1 à 9. Le terme **richesse du vocabulaire** (*Reichtum des Wortschatzes* en allemand) a été choisi, car il s'agit d'un terme communément utilisé, par exemple par les enseignants. Aucune définition n'a été transmise aux participants ; il leur était seulement précisé que nous nous intéressions à leur perception intuitive de la richesse du vocabulaire. Les évaluateurs disposaient en outre des consignes des tâches de production écrite qu'avaient reçues les enfants.
2. Après environ 10 minutes, les participants ont été invités à passer à la deuxième partie, consistant à évaluer les mêmes textes selon cinq critères, avec pour chacun d'entre eux une échelle de Likert de 1 à 9.[3] La sophistication lexicale, la diversité lexicale, l'utilisation appropriée du vocabulaire, la complexité syntaxique présentée comme « la complexité de la structure des phrases telles que des subordonnées » et en dernier, le contenu/la réponse à la tâche.
3. Finalement, deux questions ouvertes ont été posées à chaque participant directement après les deux tâches d'évaluation :
    a. Comment avez-vous procédé pour évaluer les textes lors de la première partie ? Aviez-vous une stratégie ?
    b. Avez-vous des remarques par rapport aux critères utilisés dans la deuxième partie ?

Les instructions complètes, version française, sont disponibles en annexe (Annexe 1). L'étude a duré approximativement 30 minutes par participant. Pendant les deux premières parties, la chercheuse est intervenue uniquement pour répondre aux questions des participants, pour poser des questions lorsqu'elle n'était pas certaine d'avoir compris les réflexions des participants ou encore pour encourager les participants moins loquaces.

[3] Ces critères ont été créés de manières déductives selon la théorie sur la richesse lexicale ainsi que celle concernant l'évaluation de textes d'apprenants L2. Cette tâche a été construite pour tester le matériel potentiel pour l'évaluation à grande échelle (Vanhove et al. 2019 ; Vanhove 2018). Seule l'évaluation globale a par la suite été retenue, en partie en raison du caractère chronophage de la deuxième tâche. D'ailleurs, certaines catégories auraient dû être remaniées d'après les analyses des données verbales.

2.2 Traitement et analyse des données

Toutes les données ont été enregistrées, transcrites selon les règles orthographiques et aménagées pour des questions de lisibilité :

- Les marques d'oralité telles que les hésitations, reformulations, signaux d'écoute ont été effacées respectivement corrigées sauf si l'information était pertinente pour l'analyse (par exemple lorsque cela met en évidence une difficulté à prendre une décision).
- Des informations sur le texte évalué et la tâche effectuée ont été ajoutées pour faciliter l'analyse.

Les conventions de transcription nécessaires à la compréhension des extraits sont les suivantes :

- […] = parties abandonnées (parties non pertinentes, *small talk*, etc.) ;
- (?xxx?) = mot incompréhensible, en cas de doute, la meilleure estimation ;
- () = pause ;
- [xxx] = informations ajoutées par les chercheuses lors de la transcription ;
- _ = interruption de la phrase, du mot ;
- « xxx » = citation d'un extrait d'un texte d'élève.

Une analyse qualitative de contenu inspirée de Mayring (2003) a été effectuée. Le logiciel MaxQDa (VERBI Software, s. d.) a soutenu le travail de catégorisation et codage. Deux codeurs ont d'abord créé indépendamment des catégories selon une approche inductive après plusieurs observations des transcriptions concernant l'évaluation globale de la richesse lexicale (tâche 1) (cf. Mayring 2003: 46). L'une des codeuse (L1= français), première autrice de cet article, a également récolté et transcrit les données. Elle a effectué le processus de construction des catégories et de codage deux fois à 10 mois d'intervalle.[4] Le deuxième codeur (L1 = allemand) est actif professionnellement dans le milieu de l'apprentissage des langues, mais externe à ce projet. Lors de la construction des catégories, il avait uniquement connaissance des deux premières questions de recherche (*Quels sont les critères dont les évaluateurs parlent le plus souvent et quel rôle ces critères jouent-ils dans l'attribution des scores* ? *Lesquels d'entre eux sont liés au vocabulaire* ?), des instructions de la tâche d'évaluation globale de la richesse du vocabulaire et des parties de transcriptions correspondantes afin de ne pas être influencé par les critères d'évaluation choisis dans la deuxième tâche. Il a reçu comme directive de construire des catégories qui permettraient de répondre à ces deux questions de recherche.

Après une mise en commun, cinq catégories (et neuf sous-catégories) présentées dans le tableau ci-dessous ont été retenues. Ces catégories n'étant pas exclusives, certains extraits ont été attribués à plus d'une catégorie. Une liste des (sous-)catégories avec description et nombre de segments est disponible en annexe (cf. Annexe 3).

[4] La première fois, afin de prendre des décisions pour l'étude quantitative (Vanhove et al. 2019). La deuxième fois, dans le but de répondre aux questions de ce chapitre uniquement.

Catégories et sous-catégories		Évaluateurs	Exemples
Vocabulaire	**Répétitions ou variations**	TAP1 ; 2 ; 5 ; 6	TAP1 : *Also hier fängt er an mit den Uhrzeiten im Hinblick. Aber es sind auch „dann“, „dann“, „dann“.* Traduction : Alors ici, il commence en prenant les heures en compte. Mais c'est aussi « après », « après », « après ».
	Autres	Tous	TAP6 : *Il dit « personnellement », c'est pas mal, c'est une jolie formulation.*
Structure	**Longueur des textes**	TAP1 ; 4 ; 5 ; 6	TAP1 : *also das sind zum Beispiel_ das ist auch eine_ sehr wenig, sehr kurz, () und da sind Wiederholungen drin und das sind auch keine ganzen Sätze. Also ich glaube das ist eher schlecht.* Traduction : Donc, ce sont par exemple_ c'est aussi très peu, très court, () et il y a des répétitions dedans et ce ne sont pas non plus des phrases entières. Alors, je crois que c'est plutôt mauvais.
	Contenu	TAP2 ; 4 ; 5 ; 6 ; 7	TAP7 : *Alors le A17 serait le moins bon de tous parce qu'il y a moins d'idées et puis, c'est moins développé. C'est () peut-être même que () c'est le plus jeune qui a fait ça. Je ne sais pas. En tout cas, je mettrais cinq à celui-là.*
	Genre textuel	TAP1 ; 5 ; 6	TAP6 : *J'aime bien le « bisous à tous », c'est comme un sms, je ne sais pas si c'est vraiment l'habitude d'écrire des lettres, on dirait plutôt un sms.*
	Cohérence-cohésion	Tous	TAP3 : *Il manque quand même beaucoup de_ comment dire de_. On dirait qu'on lit le texte, c'est tout le temps en continu, il n'y a pas d'arrêt, il n'y a pas de_ voilà. Même si pendant qu'on lit, on sait ce qu'elle veut dire.*
Justesse et complexité linguistique		TAP1 ; 2 ; 3 ; 4 ; 5 ; 6	TAP6 : [...] *Il a utilisé « voiture » comme sujet de la suite en fait, au lieu de répéter, il s'est dit 'je peux réutiliser le même mot comme complément et sujet'. Donc, on comprend, mais il faut le lire deux fois. La phrase d'après, il reprend le même argument, mais il le tourne autrement, donc du coup, ça ne respecte pas tellement les règles, il n'y a pas trois arguments, il y en a deux : il s'ennuie et ça va plus long. Pas de salutations, pas de date. Pour le deuxième texte, il y a un souci avec les pronoms donc « ont » il écrit « o » « n » « t », alors que c'est le pronom « nous ». On dirait qu'il avait tout dans sa tête, qu'il a écrit ce qui venait et il n'a pas relu. Par enthousiaste, je ne sais pas, il a juste écrit à la suite et du coup il manque des mots, il y a un problème avec les sujets. [...] Là, je mettrais cinq, car il y a très peu de structure, c'est pas respecté les règles et il y a pas mal d'erreurs au niveau de la syntaxe.*
Comparaison	**Différences**	TAP1 ; 4 ; 6	TAP6 : *Du coup, celui-là ça va, c'est surtout la lettre qui est moins bien.*

	Différences entre élèves	TAP1 ; 2 ; 3 ; 4 ; 5 ; 7	TAP4 : *Ja, gut. E12 ist jetzt natürlich ein drastischer Kontrast zum vorherigen.* Traduction : Ok, bon. [le texte] E12, c'est maintenant naturellement un contraste drastique avec le précédent.
	Absence de possibilité de comparer	TAP4 ; 5	TAP4 : *Ich habe auch keine Erfahrung, wie es nachher weiter geht, dementsprechend gebe ich mal sechs, fünf. Eine fünf allerdings unter dem Manko, dass er der erste war, den ich bewerte.* Traduction : Je n'ai non plus pas d'expérience pour savoir comment ça continue, en conséquence, je donne cette fois six, cinq. Un cinq toutefois avec la lacune que c'était le premier que j'ai évalué.
Profil/état de l' élève		TAP1 ; 3 ; 4 ; 5 ; 6 ; 7	TAP7 : *ça, c'est encore moins développé que les trois autres. Celui-là, je mettrais quatre. On voit que c'est vraiment un jeune enfant qui est (), qui dit vraiment ce qu'il ressent sans développer plus loin […]*

Tableau 1 : exemples d'extraits pour chaque catégorie et nombre de participants qui les ont thématisées pendant l'évaluation globale (cf. Tableau 2 en annexe pour les critères mentionnés pendant l'entretien rétrospectif)

3 Les stratégies adoptées par les évaluateurs pour résoudre la tâche

En fin d'enregistrement, les évaluateurs ont été invités à expliquer leurs stratégies ou méthodes pour évaluer la richesse lexicale de manière holistique. Ces entretiens rétrospectifs ont permis de mettre en évidence leurs perceptions des critères importants pour l'évaluation du vocabulaire de manière complémentaire (et parfois contradictoire) aux verbatim récoltés durant la tâche d'évaluation. Les transcriptions des réponses à ces questions, sur lesquelles sont basées les analyses ci-dessous, sont disponibles en annexe (Tableau 2).

Deux participants issus de domaines non linguistiques (TAP1, TAP3) ont clairement indiqué soit ne pas avoir de méthode, soit fonctionner surtout selon leurs (premières) **impressions** (*Eindruck*). L'un d'eux a ajouté qu'il s'agit d'un champ nouveau dans lequel il n'a pas de connaissances pratiques. Malgré cela, cet évaluateur a fait de nombreux commentaires sur la longueur du texte, les répétitions et les liaisons des phrases entre elles, et ce de manière plutôt systématique. L'autre évaluateur, par contre, a considéré la tâche comme un exercice qui se fait **automatiquement**. Il pensait avoir bien réussi. Nous tenons à souligner que cet évaluateur était déjà familiarisé avec ces textes pour avoir travaillé comme aide étudiant dans le projet quelques mois avant cette étude.

Deux évaluateurs ont expliqué travailler en **comparant les textes** des enfants entre eux. L'un d'eux, qui est enseignant, a mentionné travailler à l'*instinct*. Néanmoins, il décrit sa méthode clairement, démontrant qu'il s'agit de processus habituels. Son travail systématique se retrouve d'ailleurs au cours de son évaluation holistique, car la façon d'analyser se ressemble d'un texte à l'autre et aussi parce qu'il verbalise son souhait d'être consistant (1) :

1. *[…] Non si je suis la logique de ce que j'ai fait avant […].* (TAP6)

L'autre évaluateur, TAP2, qui avait demandé pendant l'évaluation holistique s'il pouvait prendre le contenu (argumentation) dans son évaluation, affirme également que la **structure des**

phrases prévaut dans son évaluation. Il mentionne aussi le vocabulaire, mais moins souvent et vaguement. On constate donc un mélange des critères utilisés.

TAP4, qui mentionne aussi le vocabulaire, prend beaucoup en compte la **prise de risque** au niveau grammatical ainsi que le fait de construire un texte intéressant. Il mentionne surtout des critères autres que le vocabulaire, mais conclut en mettant en évidence l'importance presque primordiale du **choix des mots**. Enfin, un évaluateur à profil non linguistique semble avoir été impressionné par les productions des élèves. Il explique avoir évalué plutôt à un niveau **émotionnel** en prenant en compte l'âge des élèves (TAP7). Nous reviendrons sur ces différentes stratégies dans l'analyse de contenu ci-dessous.

4 Analyses détaillées des critères influençant la perception de la richesse du vocabulaire

L'analyse de contenu sera organisée en trois étapes, de manière à illustrer le processus de réflexion des participants :

- les critères lexicaux (richesse du vocabulaire) ;
- les autres critères linguistiques ;
- les critères contextuels.

Pour la thématique « critères lexicaux », nous présenterons d'abord les extraits de verbatim venant de la première tâche (évaluation globale) puis les discuterons en fonction des verbatim issus de la deuxième tâche (évaluation par critères). Pour les deux thématiques suivantes, seuls les verbatim issus de l'évaluation globale seront présentés, car ceux de la deuxième tâche n'apportent pas de nouveau contenu permettant de répondre aux questions de recherche.

4.1 Les critères lexicaux dans la tâche d'évaluation globale

Bien qu'il était demandé aux évaluateurs d'évaluer la richesse du vocabulaire, les commentaires émis pendant l'évaluation holistique ne concernent pas uniquement le vocabulaire, mais aussi la syntaxe ainsi que d'autres critères que nous discuterons ci-dessous (sections 4.3 et 4.4). Lorsqu'il s'agit du vocabulaire, les verbatim montrent plusieurs fois des comportements d'évaluations très globales (de type *bon* ou *mauvais*), voire même des commentaires dans lesquels il n'est pas clair si l'évaluateur parle du vocabulaire ou d'autres aspects :

2. *Il n'a pas trop de richesse dans le texte, c'est vraiment trop, trop bas.* (TAP3)

Des commentaires plus précis sur le vocabulaire apparaissent sous forme de descriptions du vocabulaire au moyen d'attributs ou de citations du texte. Ces commentaires concernent le plus souvent la présence de répétitions ou, au contraire, d'un vocabulaire varié (*Vielfalt* est le mot utilisé en allemand). Ce critère apparait chez quatre participants (dont un, n'ayant pas un profil linguistique, qui mentionne neuf fois la présence de répétitions). Ce qui semble déranger les évaluateurs, ce ne sont pas uniquement les répétitions de mots lexicaux, mais aussi les répétitions de conjonctions ou d'adverbes temporels (*zuerst, dann*) :

3. *Der Schüler E5 () benutzt **nur die Konjunktion „und“,** also hat dort einen () nicht so reichen Wortschatz würde ich sagen, um Sätze zu verknüpfen. Allerdings sind die restlichen () Wörter relativ () vielfach, wenn man das so sagen kann.* (TAP2)

Traduction : L'élève E5 () utilise uniquement la conjonction « et », donc là je dirais qu'il n'a pas tellement un vocabulaire riche pour lier les phrases entre elles. Cependant, les autres mots sont relativement variés, si on peut dire ça comme ça.

Au-delà des mots, les répétitions de bouts de phrases sont également relevées :

4. *Ah ok, also mir fällt auf, dass er **zweimal sagt** „wann es mir so und so" also es wird nicht gut geschrieben, das wirkt (?als wäre?) **wenig Wortschatz.*** (TAP1)
   Traduction : Ah ok, alors je remarque qu'il dit deux fois « et moi quand ça ou ça » alors ce n'est pas bien écrit, ça sonne comme peu de vocabulaire.

La **taille du vocabulaire** se traduit chez plusieurs participants par l'utilisation de plus de mots (exemple ci-dessous) par opposition à peu de vocabulaire (*wenig Wortschatz*, TAP1) ou manque de vocabulaire (*Mangel an Wortschatz*; TAP2).

5. *Er verwendet ich habe das Gefühl **mehr Worte** auch um eine und dieselbe Tatsache vielleicht auch mal unterschiedlich darzustellen oder weiterzuarbeiten.* (TAP5)
   Traduction : J'ai la sensation qu'il utilise plus de mots, aussi pour présenter la même chose différemment ou la retravailler.

De manière similaire, un manque de vocabulaire est relevé par certains évaluateurs. Ainsi, par exemple, TAP6 dit cela à propos de la paire de textes ci-dessous : « par contre, après, on perd le fil, il manque des mots » (cf. extrait 29) :

> Je préfère à voyager en avion parce que je m'ennuie tout le temps dans la voiture. Avec l'avion ça va plus vite qu'en voiture ça passe trop long. Le temps passe plus vite en avion qu'en voiture.
>
> J'ai été dans un musée à LIEU voir des animaux ont devaient deviner le nom de l'animal ont entendus les bruit d'animaux. On a vu un serpent etc. C'est vraiment génial.

Si le contenu de nombreux commentaires se rapproche clairement du concept de **diversité lexicale** (13 extraits codés sur 34 concernant le vocabulaire), des remarques pouvant s'apparenter au concept de la **sophistication lexicale** sont aussi perceptibles, bien que moins clairement. En effet, les participants ne parlent jamais de fréquence ou rareté pendant l'évaluation holistique. Nous discuterons plus bas de potentielles raisons pour cette absence (section 4.2, la tâche d'évaluation par critères). Cependant d'autres critères relevant du concept de la sophistication en tant que qualité du vocabulaire autre que la fréquence sont détectés dans des commentaires épars, à commencer par l'utilisation de mots difficiles (cf. extrait 41). L'attribut *ausgeprägt* (remarquable), utilisé par deux évaluateurs germanophones pour le vocabulaire en est un premier exemple. La **simplicité** du vocabulaire est également relevée :

6. *Und die Worte sind relativ einfach gehalten finde ich.* (TAP5)
   Traduction : Et les mots restent à un niveau relativement simple je trouve.

Proche du concept de sophistication lexicale, on peut citer TAP6 « Il dit *personnellement*, c'est pas mal, c'est une jolie formulation *»* (cf. verbatim dans le tableau 1, catégorie: *vocabulaire-autre*) en évaluant le texte suivant :

> Le 5 avril 2014. Chère tante, personnellement, je préfère l'avion. Tout d'abord, c'est parce que le trajet est moins long que le trajet de la voiture. Ensuite, on peut regarder par la fenêtre pour voir certaines villes. Et aussi, on est bien plus confortable dans l'avion que dans la voiture. Donc, je préfère aller en avion. PRENOM.

Un autre type de remarques concernant le vocabulaire porte sur des aspects pragmatiques :

7. *Celui-là c'est le plus chou de tous [rire] avec le « au revoir et à bientôt ». () C'est difficile, je mettrais aussi un sept.* (TAP7)

Dans cet extrait, l'évaluateur semble attendri par les formules de salutations adressées à la tante. Puisque le but de l'enfant est de convaincre sa tante au travers d'une lettre, la remarque de TAP7 montre que le **choix des mots est approprié** pour atteindre ce but. Dans le même esprit, TAP6 trouve que l'expression *deux fois plus de confort* fait penser à une publicité dans le texte suivant :

> Salut, merci tout d'abord de m'avoir invité à passer tes vacances avec moi. Si tu veux mon opinion moi je préfère l'avion, ça va plus vite, ça te fatigue moins qu'en voiture, et puis dans la voiture pour dormir c'est pas terrible, alors que dans l'avion il y a deux fois plus de confort. Alors pour moi tu choisis l'avion. Bisous à toute.

Ce dernier évaluateur dénote aussi **l'oralité** du vocabulaire utilisé (*C'est aussi du vocabulaire très parlé*), comme dans le mot *shooté* dans le texte suivant :

> C'était mercredi. On est allé au ski. Il y a quelqu'un qui a shooté deux filles et la maîtresse lui a dit tu as shooté les seules filles du groupe. On allait monter en télésiège en haut de la montagne.

Le dilemme de la longueur des textes mentionné dans la partie théorique est ressorti clairement chez un évaluateur avec une formation en linguistique :

8. *Wahrscheinlich müsste ich fast eigentlich die Wörter zählen. Aber dann kann es auch sein, dass ich einen Text von Wortschatz gut bewerte, weil er länger ist, aber vielleicht redundant.* (TAP5)
   Traduction : Je devrais probablement compter les mots en fait. Mais du coup, il se pourrait que j'attribue une bonne évaluation au vocabulaire d'un texte parce qu'il est plus long, mais peut-être redondant.

Finalement, aucun passage dénotant explicitement la **densité lexicale** n'a été détecté. Les passages concernant l'oralité concernent plus le choix des mots lexicaux plutôt que leur proportion. Nous ne nions cependant pas la possibilité que dans certains commentaires, cet aspect soit implicitement caché, comme par exemple dans des remarques sur la structure des textes ou encore le manque de mots.

4.2 Le lexique dans la tâche d'évaluation par critères

Dans la deuxième partie de l'exercice, les participants ont effectué la tâche d'évaluation tout en verbalisant leur cheminement réflexif (mêmes textes, même échelle), mais cette fois-ci en suivant une liste de critères. Cette deuxième approche nous permet une systématisation plus forte des unités d'analyse, mais, comme nous le discuterons, elle reflète aussi des variations dans la compréhension et la saillance des critères d'évaluation.

Le premier critère d'évaluation dans les instructions aux évaluateurs était la **sophistication lexicale** définie comme l'*utilisation de mots rares*. Les verbatim des évaluateurs pendant la tâche montrent que la fréquence n'est pas un élément facile à définir, et que celle-ci dépend d'autres critères. Comme le montrent les extraits ci-dessous, les évaluateurs jugent en effet de la rareté des mots en prenant en compte l'âge des locuteurs, la modalité du texte ou encore la situation de communication :

9. *„überwinden" nicht so häufig für Grundschulkinder.* (TAP1)
   Traduction : « surmonter » pas si fréquent pour des élèves de l'école primaire.

10. *„Party" gleiches Problem wie mit „Klo", weil selten für einen Schreibtext, aber sonst sehr viele unterschiedliche Wörter benutzt. […] An sich werden eher häufigere Wörter benutzt, deswegen auch eine vier.* (TAP2)
    Traduction : « Boum » même problème qu'avec « WC », parce que c'est rarement utilisé pour un texte écrit, mais sinon beaucoup de mots différents utilisés. En soi, il y a plutôt des mots fréquents qui sont utilisés, donc quatre.

11. *Un mot complexe pour moi, c'est un mot qu'on ne dit pas souvent, par exemple quelqu'un qui fait des études élevées en comparaison avec ceux qui sont plus petits. Ça dépend les lieux où il est inséré, ce n'est pas forcément positif.* (TAP3)

Par ailleurs, un évaluateur ne prend pas uniquement en compte le facteur fréquence, mais également la **longueur des mots**, ainsi que les **nuances dans le vocabulaire** en contexte pour évaluer la sophistication lexicale :

12. *Utilisation, donc le lexique. Il n'y a pas tellement de mots rares. Enfin, il y « personnellement » qui est vachement **long**. Il y « trajet » qui revient, () « confortable ». () Après, dans le deuxième texte, il n'y a pas beaucoup de mots plutôt rares. [...] C'est des mots assez fréquents finalement. () « personnellement ». () Le truc, c'est que dans le deuxième texte, il utilise le « y » pour indiquer le lieu, « on y reste ». C'est pas mal ça. () « correctement ». C'est assez drôle qu'il écrive « les plus forts » et « les moins forts », il n'a pas mis les forts et les faibles, il fait un **contraste plus subtile,** plus politiquement correct quelque part. Du coup ça, c'est aussi pas mal. Mais il utilise quand même moins de mots que dans le premier. Ben, je vais mettre_ () Je vais mettre _ J'hésite entre cinq et six. Parce que je me dis, il me semble que c'est quand même un peu supérieur, donc toujours par rapport à l'âge. Mais, en même temps comme ce n'est pas très long le deuxième texte, c'est un peu difficile à voir le potentiel. Bon, je vais mettre six.* (TAP6)

Ces difficultés à travailler avec le concept de fréquence des mots employés pourraient expliquer pourquoi ce critère n'est pas apparu dans les verbatim de la tâche holistique.

Le deuxième critère d'évaluation imposé aux évaluateurs était la diversité lexicale, définie comme la proportion ou le nombre de mots différents. Le critère de la répétition avait déjà été mis en mots par plusieurs évaluateurs lors de l'évaluation holistique de la richesse lexicale. Lors de la tâche par critères, nous constatons que des évaluateurs qui n'ont pas signalé cela clairement auparavant mentionnent les **répétitions** aisément :

13. *Il n'y a pas de mots différents. Certains mots sont dits trop souvent.* (TAP3)
14. *Il y a pas mal de répétitions quand même*. (TAP7)

Il semblerait donc que la notion de diversité telle que décrite dans la section 1.1 soit l'un des facteurs aisément pris en compte par tout un chacun lors de l'évaluation du vocabulaire.

Il est à noter que, durant cette deuxième tâche pour laquelle les évaluateurs évaluent une seconde fois les mêmes textes, il arrive que les évaluateurs mentionnent les mêmes répétitions qu'ils ont déjà observées pendant l'évaluation holistique (extrait 3 pour la comparaison) :

15. *Also **wie schon vorher gesagt**, der Schüler E5 benutzt () fast nur die Konjunktion „und" um die Sätze miteinander zu verbinden. [...] Wohingegen […] unterschiedliche Wörter benutzt, deswegen würde ich bei der lexikalischen Vielfalt auch ein sieben geben. () Abzüge wegen dieser wiederholenden Konjunktion, aber sonst ist es relativ gut.* (TAP2)
    Traduction : Alors, comme déjà dit avant, l'élève E5 utilise () presque uniquement la conjonction « et » pour lier les phrases entre elles et cela dans les deux textes. [...] et utilise des mots différents, c'est pourquoi je donnerais aussi un sept à la diversité lexicale. Des points en moins à cause de la répétition de cette conjonction, mais sinon, c'est relativement bien.

Si les différents temps verbaux sont mentionnés par plusieurs évaluateurs durant l'évaluation holistique (cf. section 4.3), ils sont pris en compte par au moins un évaluateur comme preuve d'une haute diversité lexicale. Peut-on dire que cet évaluateur confond les compétences de grammaire et vocabulaire ? Ou est-ce que l'utilisation correcte de formes fléchies variées fait partie de la richesse lexicale ? Ce même évaluateur (TAP2) cite les mots et le nombre de fois qu'ils sont répétés.

Un défi mentionné par TAP5 et TAP6 est la présence des mots lexicaux contenus dans les instructions à la tâche (par exemple *voiture* et *avion*) qui ont tendance à être répétés plus souvent; ce qui est jugé par conséquent comme normal dans ce contexte *:*

16. *Il faut que pour le lexicale [diversité lexicale], je vais souligner les mots, parce que ça va m'aider à voir combien de fois il les répète. Dans le premier, il y a plus de répétitions que dans le deuxième, mais c'est un peu normal, parce que **le but c'était justement de comparer l'avion à la voiture.** Du coup, **c'est pour ça que ça revient**. […]* (TAP6)
17. *[...] il répète régulièrement « avion » et « voiture », mais comme j'ai dit, **c'est lié au fait qu'il faut les comparer.** [...]* (TAP6)
18. *Puis, l'utilisation répétitive. () Ben, ça va encore en soit. Il y a très peu de mots qui sont vraiment répétés. () Donc, comme il y en a deux dans chaque [texte] qui sont répétés, enfin les mots importants qui sont répétés deux fois. () Ça m'embête vraiment le fait que dans le premier texte, c'est vraiment, **comme on demande de comparer l'avion et la voiture**. Comme ça, bien forcément, ils vont **revenir plusieurs fois**. Du coup, j'ai de la peine à me dire que () que c'est mauvais qu'il a répété plusieurs fois les mots puisqu'il est sensé les comparer. Du coup, j'ai de la peine à () à noter ça parce que finalement, **oui, il y a répétitions, mais c'était demandé**. Ça ce n'est pas évident pour moi. Je me demande si je ne prends pas en compte, je ne prends pas en compte ces mots-là et si je regarde tout le texte, il y a très peu de répétitions. Donc, je vais mettre huit.* (TAP6)

Ces passages concernant successivement différents textes montrent l'ambiguïté de la tâche d'évaluation de la diversité lexicale. Une réflexion similaire est faite à propos de la répétition du substantif *extincteur* par TAP7 (« Il y a deux fois *extincteur*, mais c'est normal. ») lorsqu'il évalue le texte suivant :

> Le lundi 31 mars j'ai été moi et ma classe chez les pompiers ! Un monsieur nous a fait une théorie tellement longue que j'allais m'endormir mais bon. Il nous a appris à utiliser les **extincteurs** ce qu'il faut faire quand il y a le feu ce qu'il faut pas bruler et tout ça. Nous avons été dehors utiliser les **extincteurs** on a été éteindre un feu.

En conclusion, les verbatim des participants laissent apparaitre le fait que la répétition de certains mots est plus acceptable que la répétition d'autres mots, ce qui nous rapproche d'un concept plus complexe de la diversité que la définition procurée dans la littérature (cf. section 1.1). Il est à noter aussi que pour certains évaluateurs, il parait difficile de distinguer entre la diversité et la sophistication lexicale. L'un d'entre eux répond à la question « Avez-vous des remarques par rapport aux critères utilisés dans la deuxième partie ? » :

19. *La différence entre diversité lexicale et sophistication lexicale, () c'est pas la_ () c'est presque pareil pour moi. Bon, sophistication ça veut dire plus que c'est compliqué et puis diversité, c'est varié, variété, oui.* (TAP7)

Un autre évaluateur exprime, au début de l'évaluation par critères, sa difficulté à différencier les critères de diversité et sophistication lexicale. Après une question de la chercheuse visant à préciser ce point, l'évaluateur répond :

20. *Alors le truc, c'est que l'utilisation de vocabulaire () et le nombre de mots différents, le truc c'est que je trouve que c'est vachement lié, parce que s'ils utilisent des mots plutôt rares, ils vont utiliser plusieurs mots différents, je pense, enfin je ne sais pas. Du coup, pour moi, ils sont très liés, c'est assez difficile à distinguer.* (TAP6)

4.3 L'évaluation d'autres aspects linguistiques que le vocabulaire dans la tâche d'évaluation globale

Si les instructions de la tâche d'évaluation holistique ne mentionnent que la « richesse du vocabulaire », il est rapidement apparu lors des analyses que les évaluateurs se basent sur une série de critères syntaxiques, sémantiques, pragmatiques en plus que le vocabulaire lui-même. Nous allons les passer en revue ci-dessous.

4.3.1 La justesse et complexité syntaxique

Une première catégorie d'indices portant sur des éléments externes au vocabulaire regroupe des commentaires concernant la longueur et la structure des phrases (liaisons élégantes ou alignement basique de phrases, structures répétitives, etc.), la justesse ou les fautes (les phrases incomplètes, les erreurs de syntaxe, ainsi que l'utilisation adéquate ou non des temps verbaux et pronoms, etc.).

Sans surprise, il ressort que les participants font des remarques négatives sur les phrases incomplètes, la longueur des textes (cf. tableau 1), ou des utilisations inadéquates des formes fléchies. À l'inverse, la variation et l'utilisation adéquate des temps verbaux ainsi que des mots de liaisons (surtout les conjonctions) sont perçues positivement.

21. *Aussi celle-là, je trouve qu'elle est_, c'est quand même mieux que celle-là d'avant, mais, () c'est quand même mieux, mais il y a* ***par rapport aux verbes*** *et tout, je trouve qu'il manque oui qu'il manque aussi par rapport au féminin au masculin et tout, il y a trop de, ça dépend,*

des fois la personne fait la distinction, des fois elle ne fait pas. Donc je donnerais un trois aussi. (TAP3)

22. *Le deuxième texte,* ***le temps des verbes est un peu bizarre****. Puisque c'est quelque chose qui était censé être au passé et qu'il écrit au présent, bon.* (TAP6)
23. *Bon, c'est écrit au passé. Donc ça, c'est juste.* (TAP6)

Un évaluateur constate que le manque de vocabulaire a des conséquences à un niveau syntaxique voir discursif :

24. *Und alles in allem würde ich sagen, ist der Wortschatz auch nicht so reich und […] denke ich* ***durch den Mangel an Wortschatz [können] keine schönen Sätze*** *konstruiert werden.* (TAP2)
    Traduction : Dans l'ensemble, je dirais que le vocabulaire n'est pas non plus si riche et je pense qu'à cause du manque de vocabulaire, il ne peut pas construire de belles phrases.

Il est à noter qu'il y a parfois des différences entre le poids ou l'importance qu'un évaluateur dit donner à un critère et le nombre de verbatim concernant ce même critère. Par exemple, un participant qui a émis plus de commentaires sur des caractéristiques de justesse et de complexité syntaxique que les autres participants souligne lors de l'évaluation du troisième texte que malgré tout, il ne s'agit pas du critère le plus important pour lui dans l'évaluation :

25. *Ich möchte dabei allerding anmerken, dass mir bei so einer Beurteilung richtige Grammatik nur sekundär ist. Ich allerding finde es wichtig, hat vor allem die Konsistenz, so ein konsistenter Erzählstrang gebildet werden kann und dass die Person sich ausdrücken kann, dass man das versteht.* (TAP4)
    Traduction : J'aimerais tout de même préciser que pour moi, la justesse grammaticale n'est que secondaire dans une évaluation de ce type. Je trouve la consistance surtout importante, qu'un fil narratif puisse être formé et que la personne puis s'exprimer, que l'on comprenne :

4.3.2 La structure du texte

Une deuxième catégorie d'indices non-lexicaux ressortant des protocoles de pensée à haute voix de nos participants concerne la structure du texte. Au sein de celle-ci, nous distinguons des commentaires concernant la taille des textes, l'adéquation des textes à la tâche (le contenu, le genre textuel) ainsi que la **cohésion** et la **cohérence**.

Un évaluateur verbalise ainsi le problème de l'évaluation du vocabulaire sur de courts textes :

26. *Auch dann was die Erzählung anbelangt, () es ist ziemlich kurz eigentlich. () Auch so was wie „ich habe meine Kleider abgezogen" (). Ja finde ich nicht so gut ausgeprägt. Vom Sprachlichen her ist es () schlechter noch als das [von] E5 ich gebe es, aber vom Wortschatz wahrscheinlich...* ***Obwohl Wortschatz hat ja nicht unbedingt etwas mit der Textlänge zu tun****.* (TAP5)
    Traduction : En ce qui concerne le récit également, () c'est assez court en fait. Aussi quelque chose comme « j'ai enlevé mes habits ». () Oui, je ne trouve pas très bien développé. D'un point de vue du langage, c'est () encore plus mauvais que celui de E5 j'avoue, mais d'un point de vue du vocabulaire… Bien que le vocabulaire n'ait pas forcément quelque chose à voir avec la longueur des textes.

La deuxième sous-catégorie structurelle concerne le **contenu**. En particulier, quatre évaluateurs (TAP2, TAP4, TAP5, TAP6) relèvent la (non-)présence des trois arguments demandés dans les instructions aux élèves pour le texte argumentatif ou le nombre d'informations dans le texte narratif (activités, lieu).

27. *Alors, le premier pas de structure de lettre, mais **il y a quand même des arguments** : « c'est plus confortable », « on doit pas s'arrêter pour aller aux toilettes » et « on dort ». Oui, ok, il y a bien trois arguments. Mais bon, le « c'est plus confortable », il est répété deux fois. C'est marrant qu'il dise les toilettes. On dirait vraiment que ça l'a marqué quand il faisait les trajets en voiture.* (TAP6)

L'évaluateur de ce dernier exemple, enseignant de langue à temps partiel, compte même systématiquement les arguments dans les différents textes. Le fait qu'une majorité des évaluateurs fait très attention au contenu alors qu'ils devraient évaluer la richesse du vocabulaire pourrait être dû au fait qu'ils disposent des instructions que les enfants ont reçues lors de la récolte des données. Les évaluateurs pensent donc qu'ils doivent utiliser cette information. D'ailleurs, un évaluateur, apparemment confus, demande explicitement s'il peut prendre la non-présence des trois arguments en compte pour l'attribution des scores :

28. *Gehört es auch mit dazu, wenn ich_ Also er schreibt nicht mit drei Argumenten. Das darf ich auch bewerten oder es ist_?* (TAP2)
    Traduction : Est-ce que ça fait aussi partie [de l'évaluation], si je_ Alors il n'écrit pas avec les trois arguments. Est-ce que je peux évaluer cela aussi ou est-ce_ ?

Parfois, la présence des trois arguments intervient en conjonction avec d'autres critères, jusqu'à prendre une place justificative pour la note attribuée :

29. *Alors, la lettre. De nouveau, il n'y a **pas de structure de lettre (). Mais il y a des arguments** : il y a le restaurant, aller à un stade, je suppose qu'il aime le foot, et il y a un supermarché. **Par contre, après on perd le fil, il manque des mots** et franchement, je ne comprends pas le sens de la fin de la phrase : « on peut trouver chez nous pour voir la maison », je, non, je ne comprends pas. Pour le récit, ce n'est pas trop mal, il dit ce qui s'est passé : ils sont revenus, ensuite ils sont allés là-bas, après ils sont allés camper. Bon, il y a pas mal de fautes mais_ () non, si je suis la logique de ce que j'ai fait avant, il y a un des deux textes qui est plutôt correct. Donc le récit est plutôt correct, même si au niveau de l'orthographe, ce n'est pas super. Mais la lettre, c'est seulement à moitié en fait. Pff, ce n'est pas facile ! () Le truc c'est que si j'enlève un point pour la structure de la lettre_ () Je sais pas quoi faire avec celui-là je dois dire. Parce qu'il a quand même beaucoup écrit, donc, je ne veux quand même pas_ je ne sais pas_ je ne veux pas lui enlever un point parce qu'il a quand même beaucoup écrit, même s'il y a pas mal de faute, on ne comprend pas la dernière phrase de la lettre effectivement. Il n'y a pas de structure de lettre, ça de toute façon, ça enlève quand même un point. Le vocabulaire n'est pas si mal. Ce n'est pas simple. Celui-là, il m'embête un peu. **Bon, je vais mettre six car il y a quand même trois arguments dans la lettre** () et le deuxième texte il est quand même continu et puis, il est plus ou moins correct.*

Donc ça, mais il n'y a pas la structure de la lettre et il y a un problème dans le sens des phrases. Je vais mettre six. (TAP6[5])

Le contenu (idées, arguments), mais aussi le **respect du genre textuel** sont des critères régulièrement utilisés par les évaluateurs. Comme l'illustre l'exemple ci-dessous, ces deux facteurs ne vont pas forcément de pair :

30. *Hier fängt es auch mit einer Anrede an. () Ich glaube, dass das Kind hat ein bisschen missverstanden, weil ich glaube die Tante lädt ein das Kind mitzukommen und hier schreibt das Kind, er lädt die Tante ein. () Und* ***es ist so ein Brief aber es ist keine Argumentation*** *drin. Es sagt nur ja „ich will mit dem Auto" aber nicht warum und_. Aber die Textform „Hallo Tante" mit den Anreden und auch am Ende „bis blad", „ich habe dich lieb" (..) hier wurde wirklich ein Brief geschrieben aber es ist keine Argumentation.* (TAP5)
    Traduction : Ici, ça commence aussi avec une formule d'appel. () Je crois que l'enfant a un peu mal compris, parce que je crois que c'est la tante qui invite l'enfant à venir et ici, l'enfant écrit qu'il invite la tante. C'est bien une lettre, mais il n'y a pas d'argumentation dedans. Il dit seulement « je veux aller en voiture », mais pas pourquoi et_. Mais la forme textuelle « Salut tante » avec la formule d'appel et à la fin « à bientôt », « je t'aime » [...] c'est vraiment une lettre qui a été écrite, mais ce n'est pas une argumentation.

Deux évaluateurs émettent plusieurs fois une remarque quant à la présence ou l'absence d'une structure de lettre et un autre évaluateur souligne positivement pour deux textes narratifs la production d'un *vrai récit*.

Finalement, un critère majeur qui paraît influencer les évaluations est celui de la cohésion et la cohérence. Deux évaluateurs mentionnent plusieurs fois successivement qu'un texte est compréhensible et l'un d'entre eux lie ces remarques directement avec le fait qu'il y a une cohérence *entre les phrases*. Ces commentaires laissent suggérer que tous les textes ne sont pas compréhensibles et cohérents. De plus, TAP4 mentionne expressément qu'il trouve particulièrement important dans ce type d'évaluation qu'on comprenne le texte. Trois évaluateurs notent la présence ou le manque de lien logique entre les phrases ou plus globalement un discours décousu (TAP1, TAP2, TAP5). L'un d'eux se dit même irrité par cela :

31. *Also das sind ein bisschen längere Sätze aber was mich ein sehr irritiert ist, dass die Sätze nicht gut miteinander verbunden sind, also es ist, es kommt (vor) wie Gedankensprünge. Also vielleicht der Wortschatz ist_ () also nur vom Wortschatz her würde ich vielleicht sagen vier.* (TAP1)
    Traduction : Alors, il y a des phrases un peu plus longues, mais ça m'irrite beaucoup que les phrase ne soient pas bien liées les unes aux autres. Alors, peut-être que le vocabulaire est_ () alors uniquement par rapport au vocabulaire, je dirais quatre.

[5] A propos de la paire de textes suivante: Je préfère aller en voiture parce qu'on peut trouver un restaurant. On peut aussi trouver un **stade** on peut même trouver un **supermarché**. On peut regarder dans vitre et on voit des animaux **on peut trouver chez nous pour voir la maison**. / Ma dernière cours d'école on est allé camper, on est revenu à l'école manger un pique-nique, on est allé à l'accrobranche et après on est allé camper.

Par ailleurs, cet extrait indique une tactique consistant à commenter les textes selon des critères structurels, puis à revenir au vocabulaire juste avant de donner un score, mais sans forcément développer.

Un autre évaluateur (TAP6) établit cependant un lien entre la cohérence-cohésion et l'emploi du vocabulaire. Il explique avoir besoin de relire deux fois un passage, car il ne le trouve pas compréhensible : « Il a utilisé voiture comme sujet de la suite en fait, au lieu de répéter, il s'est dit *je peux réutiliser le même mot comme complément et sujet*. Donc, on comprend, mais il faut le lire deux fois. La phrase d'après, il reprend le même argument, mais il le tourne autrement, donc du coup, ça ne respecte pas tellement les règles, il n'y a pas trois arguments, il y en a deux ». Ensuite, cet évaluateur conclut que le manque de mots empêche la compréhension de la lecture du texte : « On dirait qu'il avait tout dans sa tête, qu'il a écrit ce qui venait et il n'a pas relu. Par enthousiaste, je ne sais pas, il a juste écrit à la suite et du coup il manque des mots, il y a un problème avec les sujets. » (cf. Tableau 1 catégories « justesse et complexité linguistique » pour l'extrait complet).

Nous discuterons dans la prochaine section de deux critères majeurs, mais indirectement liés aux textes, ressortant des analyses : le profil de l'élève, et les modalités de l'exercice d'évaluation.

4.4 Quelques critères non textuels

4.4.1 Le profil des élèves

La majorité des évaluateurs contextualisent leurs évaluations en fonction de ce qu'ils considèrent être le stade développemental (ou le niveau cognitif) des élèves. Les évaluateurs parlent de la possibilité que l'élève n'ait pas compris les instructions, du degré de réflexion qu'il amène dans l'exercice (32), de son âge, de son état émotionnel, de son niveau linguistique ou de son statut d'apprenant de langues (33).

32. *Es ist () viel weiter ausgearbeitet. Das Kind hat sich* ***wirklich reingedacht****, was ist in einem Flugzeug, oder was passiert, wenn man mit dem Flugzeug fliegt, warum ist es besser. Hat auch formuliert „ja das ist meine Meinung". Also einfach vom Textaufbau, der natürlich auch länger ist und viel ja reicher auch nicht nur von Argumenten her sondern auch vom Wortschatz her.* (TAP5)
    Traduction : C'est beaucoup plus élaboré. L'enfant a vraiment réfléchi, qu'est-ce qu'il y a dans un avion, qu'est-ce qui se passe quand on vole en avion, pourquoi c'est mieux. Il a aussi formulé « c'est mon avis ». Alors simplement par rapport à la construction du texte, qui est naturellement aussi plus long, et beaucoup plus riche, pas seulement par rapport aux arguments, mais aussi par rapport au vocabulaire.

33. *Der nächste, E8, ist () ganz offensichtlich () schon* ***weiter in seiner Fremdsprache.***[6] *Er benutzt viel mehr Sprach-Features als der erste, also „fände es cool" da zum Beispiel. () Auch generell so der ganze Gedankengang, was man alles machen kann. (...) Wenn man*

[6] Sachant que la chercheuse travaillait à l'Institut de plurilinguisme, TAP4 a probablement conclu qu'il évaluait des textes d'apprenants.

die beiden Texte zusammen betrachtet, es ist ganz offensichtlich, dass das ***Sprachenniveau*** *ein gutes Stück besser ist [...].* (TAP4)
Traduction : Le suivant, E8, est () clairement plus avancé dans sa langue étrangère. Il emploie beaucoup plus de fonctionnalités du langage que le premier, alors « trouverait ça cool » par exemple. () De manière générale, le raisonnement sur tout ce qu'on peut faire. [...] Si on regarde les deux textes, il est clair que le niveau linguistique est bien meilleur [...]

En particulier, l'âge des élèves semble jouer un rôle important pour l'évaluation bien que les évaluateurs ne connaissent que leur âge approximatif. Ainsi par exemple, TAP3 assigne l'emploi de certains mots à une certaine tranche d'âge, tandis que TAP1 fait référence à l'âge en critiquant la façon dont les phrases sont formulées :

34. *Celle-là, je trouve bien. En considérant que* ***c'est peut-être quelqu'un de jeune*** *qui a écrit, parce que relativement au vocabulaire, il y a des mots quand même qu'on dit plutôt* ***quand on est plus petit*** *[rire], et sinon, ça va, je donnerais quand même quatre.* (TAP3)
35. *Sätze sind ein bisschen besser ausformuliert, aber sie machen auch nicht so richtig Sinn. [...] Man merkt,* ***dass es ein Kind geschrieben hat****.* (TAP1)
Traduction : Les phrases sont un peu mieux formulées, mais elles n'ont pas non plus vraiment de sens. [...] On remarque que c'est écrit par un enfant.

L'âge constitue d'ailleurs le critère principal d'un évaluateur au profil non linguistique qui ne mentionne que sporadiquement des points concernant le vocabulaire ou d'autre caractéristiques linguistiques. Cet évaluateur utilise un vocabulaire des affects pour décrire les textes (*cool, émotionnel*); mais aussi un vocabulaire de la cognition pour décrire un élève (*plus évolué*). Il essaye de deviner quels enfants sont plus jeunes en se basant sur les formulations utilisées. De plus, il compare les textes à évaluer avec ceux qu'il pense qu'il aurait lui-même pu produire au même âge (voir aussi tableau 1, contenu) :

36. *Moi, je trouve que* ***pour un enfant de dix ans, c'est pas mal du tout****. Quand même, () moi, je mettrais même un sept. Je ne pense pas qu'****à cet âge*** *j'avais déjà* ***autant d'idées****.* (TAP7)

Le côté émotionnel ressort également chez un enseignant de langue étrangère. Dans ce cas, il ne semble pas que cela influence directement l'évaluation, mais bien plutôt que cela apporte en premier lieu une explication aux problèmes textuels : l'état psychologique de l'enfant (« par enthousiaste », « Je pense qu'il a vraiment aimé et le fait qu'il puisse raconter ça l'a peut-être déstabilisé ») aurait ainsi pu l'amener à écrire tout d'une traite sans liaison entre les phrases (cf. Tableau 1 catégories « justesse et complexité linguistique » pour l'extrait complet).

En conclusion, cette analyse des commentaires sur le profil de l'élève laisse suggérer que les évaluateurs font preuve d'une certaine clémence en fonction de l'âge. Ceci nous amène à émettre l'hypothèse que s'ils savaient qu'ils évaluaient des textes d'enfants issus de la migration, ils seraient aussi (ou encore plus) cléments dans leurs évaluations.

4.4.2 La comparaison entre les élèves

La tâche d'évaluation ne comprenant ni entrainement, ni possibilité de lire une fois l'ensemble des textes avant de commencer, le besoin de comparer le niveau des textes entre les élèves s'est fortement fait ressentir. Six des sept évaluateurs comparent ainsi explicitement les paires de textes avec les une à trois paires précédentes. Ces comparaisons se résument généralement à

dire si les textes sont mieux ou moins bien (37 et 39), mais aussi, plus rarement, peuvent amener à des réflexions sur la difficulté d'attribuer des scores de façon consistante (38). Cette question de la consistance sera reprise dans une perspective plus quantitative dans la discussion :

37. *Das sind viel längere Sätze also nacheinander aber sie sind auch nicht richtig verbunden, finde ich. Also es ist* ***besser*** *halt als glaube ich* ***als davor****. Es sind* ***längere*** *Sätze aber es ist nicht so schön ausformuliert. Ich würde sagen fünf.* (TAP1)
    Traduction : Ce sont donc des phrases beaucoup plus longues, l'une après l'autre, mais elles ne sont pas non plus vraiment liées, je trouve. Bon, c'est mieux que celui d'avant, je crois. Ce sont des phrases plus longues, mais elles ne sont pas si jolies. Je donnerais un cinq.
38. ***Beim Ersten finde ich jetzt es schwierig einzuschätzen****, weil ich noch nicht_ Ich tue mich schwer zu wissen, wie die Kinder_ was sie für eine Performanz haben, in dem Alter von 8 und 10. () [...] Ich mache mal eher eine () vier und hoffe, dass so als Mittelmass ist und hoffe, dass ich es dann nicht zu tief einsetze, dass es dann Kinder gibt, die dann ein bisschen mehr_* (TAP5, évaluation de la première paire de textes)
    Traduction : Pour le premier, je trouve difficile d'évaluer, parce que je ne_ J'ai de la peine à savoir quelle performance ont les enfants entre 8 et 10 ans. Je donne déjà un () quatre et j'espère que c'est une moyenne et que je n'ai pas donné une note trop basse, qu'il y aura des enfants qui auront un peu plus_
39. *Vom Sprachlichen ist es ()* ***schlechter noch als*** *das E5 ich gebe es, [...]. Ja, wenn ich der E5 eine vier gegeben habe, dann muss ich ihm auch eine vier geben, es ist so ungefähr vergleichbar.* (TAP5, évaluation de la deuxième paire de textes).
    Traduction : D'un point de vue linguistique, c'est () encore moins bon que E5 [première paire de textes] j'avoue [...] Si j'ai donné un quatre à E5, alors je dois lui donner aussi un quatre, comme ça c'est à peu près comparable.

Ces comparaisons peuvent aussi avoir lieu à posteriori et amener les évaluateurs à émettre le souhait de modifier leur(s) notation(s) précédente(s).

40. *Jetzt nachdem ich die anderen Texte gemacht habe, hätte der erste wahrscheinlich auch eher eine sechs () bekommen.* (TAP4)
    Traduction : Maintenant que j'ai fait les autres textes, le premier aurait certainement reçu un six.

4.4.3 La comparaison entre les deux productions écrites

Une dernière difficulté rencontrée par les évaluateurs concerne des différences constatées entre le niveau des deux textes (argumentatif et narratif) des élèves (TAP1, TAP4 et TAP6) : parfois le récit est jugé comme meilleur, d'autres fois la lettre argumentative. Un évaluateur semble ainsi particulièrement *irrité* par cette variance, ce qui ressort plutôt de l'exception, les autres s'exprimant de façon plus neutre à ce propos :

41. *Was mich jetzt gerade erstmal sehr irritiert ist, dass ich nicht finde, dass beide Texte gut zusammenpassen, weil* ***der Erste auch sehr unlogisch geschrieben*** *ist, so zu sagen „ich fahre lieber mit dem Auto aber lass uns mit dem Flugzeug fliegen". Und* ***der Zweite ist sehr konkret geschrieben*** *also das finde ich eigentlich sehr gut geschrieben, wie ein richtiger Bericht (wir waren da und da), und eigentlich mit vielen schwierigen Wörtern. Also vom*

Wortschatz denke ich, dass das eigentlich sehr hoch ist. () Ja. Also ja, da finde ich wirklich ganz klar, dass sie nicht so gut zusammenpassen. (TAP1)
Traduction : Tout d'abord, ce qui m'irrite vraiment, c'est que je ne trouve pas que les deux textes aillent bien ensemble. Le premier est écrit de façon très illogique pour ainsi dire « je préfère aller en voiture, mais allons-y en avion ». Le deuxième est écrit de façon très concrète, je le trouve du coup vraiment bien écrit, comme un vrai rapport (on était là et là), et en fait, avec beaucoup de mots difficiles. Pour le vocabulaire, je pense que c'est en fait très haut. () Oui. Donc oui, je trouve clairement qu'ils ne passent pas très bien ensemble.

5 Synthèse et discussion des principaux résultats

La tâche d'évaluation de la richesse du vocabulaire accompagnée des protocoles de pensée à haute voix a été conçue dans le but de répondre à trois questions de recherche. Nous en discutons les principales conclusions dans les prochains paragraphes.

5.1 Quels sont les critères dont les évaluateurs parlent le plus souvent et quel rôle ceux-ci ont-ils dans l'attribution des scores ?

Un point saillant des résultats de notre recherche est que, lorsque l'on demande à des évaluateurs non entrainés d'évaluer la richesse du vocabulaire de textes d'enfants de 8 à 10 ans, ce n'est pas de vocabulaire dont ils parlent le plus souvent. Il semble que la **cohérence et une bonne structure** de texte jouent un rôle très important. On peut émettre l'hypothèse que la cohérence et la structure du texte sont des critères d'analyse à un niveau textuel plus global et qu'ils sont donc logiquement observés en premiers. En effet, une structure de texte cohérente et cohésive est cruciale pour la compréhension du texte et les évaluateurs essaient probablement tout d'abord de comprendre le texte avant de l'évaluer. Cette problématique ressort peut-être ici de façon particulièrement saillante puisque les textes à évaluer ont été rédigés par des enfants du primaire et que plusieurs textes comportent effectivement des problèmes de structure et de cohérence. Ceci pourrait expliquer pourquoi plusieurs évaluateurs commencent fréquemment leurs évaluations par ce type de critères, avant de revenir au vocabulaire juste avant d'attribuer un score au texte. Ces critères perdraient probablement de l'importance si l'évaluation portait sur des textes d'élèves plus avancés dans leur scolarité.

Si les instructions données aux participants de cette étude ont été rédigées de façon à donner un minimum d'informations quant au profil des élèves qui ont écrit les textes, les instructions reçues par ces enfants ont par contre été présentées aux évaluateurs. Ceci semble avoir influencé les évaluations : de nombreux verbatim montrent que les évaluateurs prennent en compte le **contenu des textes** en termes de complétion de la tâche et qu'une connaissance en amont du niveau global des textes est ressentie comme nécessaire par les évaluateurs. En ce qui concerne ce dernier point, les évaluateurs ont exprimé clairement à différentes reprises qu'ils auraient évalué certains textes différemment s'ils avaient vu d'autres textes auparavant, ou su que ces enfants étaient allophones. Cette préoccupation de (non-)consistance dans les évaluations pourrait amener à poser l'hypothèse qu'il est difficile de calibrer la richesse lexicale des textes sur la base d'évaluations humaines non entrainées, les phénomènes de voisinage semblant avoir un impact. Pourtant, une étude quantitative d'évaluations des textes du même corpus, a mis au jour une fiabilité inter-évaluateurs (coefficient de corrélation intra-classe (ICC) moyen entre 0.71 et

0.90 selon la langue) ainsi qu'une régularité intra-évaluateurs (Vanhove 2018 : 82–83). Dans cette étude, des adultes non entrainés (146 francophones, 322 germanophones et 106 lusophones) ont évalué en ligne les productions écrites des enfants en termes de richesse du vocabulaire sur une plateforme internet. Ces productions avaient été corrigées au niveau orthographique et grammatical. Chaque évaluateur a évalué un set d'environ 50 textes répartis aléatoirement et précédés de deux textes d'entrainement. Chaque texte a ainsi été évalué par un nombre d'évaluateurs oscillant entre 3 et 18 selon la langue (Vanhove et al. 2019). De ces résultats, il semble donc que, malgré la perception de non-consistance dans les évaluations ressortant de nos analyses qualitatives, une autre image apparait des analyses quantitatives. Autrement dit, les évaluations humaines n'étaient pas aléatoires.

Le besoin qu'ont éprouvé nos évaluateurs de connaitre à l'avance les textes ou, du moins d'avoir une idée du niveau moyen de la cohorte, confirme en outre les conclusions d'études précédentes (voir par exemple Lumley 2002, pour une étude utilisant aussi un protocole de pensée à haute voix pour une tâche d'évaluation qui a montré que les évaluateurs se réfèrent aux textes qu'ils ont évalués précédemment lorsqu'ils doivent attribuer de nouveaux scores). Ce résultat est pourtant à relativiser dans le cadre de notre étude. Notre but, dans cette recherche qualitative, est en effet de comprendre la perception commune de richesse de vocabulaire et non pas d'obtenir une évaluation équitable.

Finalement, si on compare les critères que les participants mentionnent pendant la tâche de l'évaluation globale avec ceux de la verbalisation rétrospective, on constate qu'ils se basent sur plusieurs critères et non pas seulement sur celui qu'ils considèrent comme leur critère majeur d'évaluation. Seul TAP7, qui, mis à part le vocabulaire global, met l'accent presque uniquement sur l'âge des enfants et le côté émotionnel de leurs textes, est très consistant entre son analyse et la façon de décrire sa méthode.

5.2 Quels critères sont liés au vocabulaire ?

Le vocabulaire est pris en compte, parfois de façon très précise et exemplifiée, mais très souvent de manière plutôt globale. Il nous a été de fait plusieurs fois impossible de savoir comment les participants évaluent le vocabulaire, en particulier dans les cas de figure où l'évaluateur détaille d'autres critères que le vocabulaire, puis, juste avant de donner un score, produit un énoncé commençant avec une mise en relation du style « par rapport au vocabulaire ».

Pourtant deux sous-catégories de critères liés au vocabulaire ont pu être identifiées : la **répétition/variation du vocabulaire**, qui représente la sous-catégorie du vocabulaire la plus saillante, et les **autres commentaires liés au vocabulaire**. Dans cette dernière catégorie, nous avons pu observer des remarques globales (bon/mauvais, riche) sur le vocabulaire, mais aussi des remarques, moins fréquentes, sur sa qualité, sa beauté (*jolie formulation*, *chou*), son oralité, sa simplicité/difficulté, ou la présence de vocabulaire enfantin.

5.3 Les critères liés au vocabulaire se rapprochent-ils des composantes de la richesse lexicale apparaissant dans la littérature ?

Comme mentionné ci-dessus, le critère lexical apparemment le plus utilisé par nos participants est celui de la répétition. Il est question de répétitions de mots fonctionnels, mots lexicaux ainsi

que de parties de phrases. Ces commentaires se rapprochent ainsi tout particulièrement des mesures de **diversité lexicale**. Par contre, l'hypothèse selon laquelle les mots non-lexicaux ne seraient pas pris en compte pour évaluer le vocabulaire est infirmée ici, car certains évaluateurs portent une attention particulière à la répétition de conjonctions (« und », « weil »), soit dans l'évaluation globale soit dans l'évaluation par critères. Les commentaires suggèrent de plus qu'une analyse à n-gram se rapproche également des intuitions humaines d'évaluation, car dans certains extraits, les participants mettent en évidence une répétition de plusieurs mots :

42. *[...] das ist ganz schlecht geschrieben. Und hier hat man auch „ich will mit dem Flugzeug gehen" „ich will mit dem Flugzeug gehen".*[7] (TAP1)
    Traduction : Il y a des répétitions, c'est très mal écrit. Et ici, on a aussi « je veux aller en avion », « je veux aller en avion ».

En ce qui concerne la sophistication lexicale, les participants ne mentionnent que rarement des caractéristiques lexicales liées à la fréquence générale des mots utilisés. Il semble que les évaluateurs de cette étude préfèrent évaluer la qualité du vocabulaire autrement, peut-être de manière plus intuitive. Il en va de même pour les collocations : même si, parfois, les évaluateurs font des réflexions sur des groupes de mots, il ne s'agit pas à proprement parler de collocations. Une explication pour la quasi absence de ces deux critères pourrait résider dans la simplicité des textes analysés et donc, recelant un nombre insuffisant de collocations et de mots considérables comme rares pour pouvoir tirer des conclusions plus précises.

6 Conclusion

Nous concluons tout d'abord que si les protocoles de pensée à haute voix ont permis de mettre en lumière différents critères utilisés (plus ou moins consciemment) avec une certaine régularité par des personnes non entrainées pour évaluer la richesse lexicale de productions écrites, ils nous ont aussi rendus attentifs au fait que chaque évaluateur s'exprime dans un style propre et montre une sensibilité particulière pour certains critères.

En général, les évaluateurs ne se basent pas que sur des critères lexicaux lorsqu'on leur demande d'évaluer la richesse du vocabulaire, et semblent dérangés (ou aidés) par, entre autres, le contenu, des caractéristiques syntaxiques, le respect des règles du genre textuel cible, mais aussi par des critères qui ne sont pas uniquement liés au texte tels que le besoin de comparer les textes entre eux ou de les connaitre à l'avance. Il est aussi intéressant de remarquer que les critères ne peuvent être compris comme des classes absolues, car, comme l'analyse nous le montre, ceux-ci s'entrecroisent régulièrement. Cela laisse supposer un besoin de l'évaluateur de passer par une analyse globale avant de peut-être se recentrer sur l'évaluation du vocabulaire.

Finalement, cette étude met en évidence une limite des méthodes automatiques de richesse lexicale, si appliquées à un niveau individuel : plusieurs verbatim montrent que la subjectivité humaine joue un rôle important, par exemple dans le cas de répétitions que certains évaluateurs considèrent comme normales en vertu de la thématique du texte. Mais ces aspects jouent-ils encore un rôle à grande échelle ? Serait-il possible d'obtenir rapidement, c'est à dire au moyen

[7] Texte évalué : Hallo Tante, ich will mit dem Flugzeug gehen. 1 Weil ich mit dem Auto Angst habe. Und ich will immer mit dem Flugzeug gehen. Und mit dem Flugzeug schneller ist, als das Auto. Herzliche Grüsse VORNAME.

de formules, une évaluation de la répartition de la richesse lexicale des textes d'un corpus tout en respectant, dans une certaine mesure, la perception humaine ?

Une piste de réponse se trouve dans une autre contribution de ce projet, soit l'étude de Vanhove et al. (2019) décrite dans la discussion ci-dessus. Au total, 3060 textes ont ainsi été évalués sur une échelle de Likert. Les moyennes des évaluations par texte ont ensuite été comparées aux valeurs issues de plus de 150 formules dans divers modèles statistiques. Les résultats ont montré que les évaluations humaines peuvent effectivement être prédites dans une certaine mesure et, entre autres, qu'une formule de diversité lexicale facile d'application, l'indice de Guiraud (cf. Guiraud 1954), fonctionne relativement bien sur ce corpus (Vanhove et al. 2019). Notons à nouveau que ces textes avaient été corrigés au niveau grammatical et que les évaluateurs ont eu deux textes d'entrainement, ce qui n'était pas le cas dans l'étude ici présente.

Références

Berthele, Raphael/Lambelet, Amelia (2017): *Heritage and School Language Literacy Development in Migrant Children: Interdependence Or Independence?* Multilingual Matters.

Bonvin, Audrey/Lambelet, Amelia (2017): "Algorithmic and subjective measures of lexical diversity in bilingual written corpora: a discussion". *CORELA HS-21*. doi: 10.4000/corela.4843

Daller, Helmut (1999): *Migration und Mehrsprachigkeit*. Frankfurt/M.: Lang

Daller, Helmut et al. (2003): "Lexical Richness in the Spontaneous Speech of Bilinguals". *Applied Linguistics* 24/2: 197–222.

Ericsson, K. Anders/Simon, Herbert A. (1980): "Verbal reports as data". *Psychological Review* 87/3:215–251. doi: 10.1037/0033-295X.87.3.215

Fergadiotis, Gerasimos et al. (2013): "Measuring lexical diversity in narrative discourse of people with aphasia". *American Journal of Speech-Language Pathology* 22/2:397–408. doi: 10.1044/1058-0360(2013/12-0083

Guiraud, Pierre (1954). Les charactères statistiques du vocabulaire. Essai de méthodologie. Paris: Presses Universitaires de France.

Henrichs, Lotte/Schoonen, Rob (2009): "Lexical features of parental academic language input: the effect on vocabulary growth in monolingual dutch children". In: Richard, Brian et al. (eds): *Vocabulary Studies in First and Second Language Acquisition. The interface between theory and application.* Palgrave Macmillan, London: 1–22.

Henriksen, Birgit (2013): "Research on L2 learners' collocational competence and development – a progress report". In: Bardel, Camilla et al. (eds): *L2 vocabulary acquisition, knowledge and use : New perspectives on assessment and corpus analysis. EUROSLA - the European Second Language Association.* 29–56.

Hubert, Pierre/Labbé, Dominique (1988): « Un modèle de partition du vocabulaire. « In : Dominique Labbé/Philippe Thoiron (eds.) : *Etudes sur la richesse et la structures lexicales*. Slatkine-Champion. ffhal-00758061. 93–114.

Jarvis, Scott (2013a): "Capturing the Diversity in Lexical Diversity". *Language Learning* 63/1: 87–106. doi: 10.1111/j.1467-9922.2012.00739.x

Jarvis, Scott (2013b): "Defining and measuring lexical diversity". In: Jarvis, Scott/Daller,

Michael (eds): *Vocabulary Knowledge: Human ratings and automated measures.* Amsterdam/Philadelphia, Benjamins. 13–44.

Jarvis, Scott (2017): "Grounding lexical diversity in human judgments2. *Language Testing* 34/4: 537–553. doi: 10.1177/0265532217710632

Johansson, Victoria (2008): "Lexical diversity and lexical density in speech and writing: a developmental perspective". *Working paper* 53: 61–79.

Kuusela, Hannu/Paul, Paul (2000): "A comparison of concurrent and retrospective verbal protocol analysis. "*The American journal of psychology*, 113/3: 387–404.

Laufer, Batia (1994): "The Lexical Profile of Second Language Writing: Does It Change Over Time?" *RELC Journal* 25/2: 21–33. doi: 10.1177/003368829402500202

Levitzky-Aviad, Tami/Laufer, Batia (2013): "Lexical properties in the writing of foreign language learners over eight years of study: single words and collocations". In: Bardel, Camilla et al.: *L2 vocabulary acquisition, knowledge and use : New perspectives on assessment and corpus analysis. EUROSLA - the European Second Language Association*: 127–148.

Lindqvist, Christina et al. (2013): "A new approach to measuring lexical sophistication in L2 oral production". In: Bardel, Camilla et al.: *L2 vocabulary acquisition, knowledge and use : New perspectives on assessment and corpus analysis. EUROSLA - the European Second Language Association*: 109–126.

Lumley, Tom (2002): "Assessment criteria in a large-scale writing test: what do they really mean to the raters?" *Language Testing 19/*3: 246–276. doi: 0.1191/0265532202lt230oa

Mayring, Philipp (2003): *Qualitative Inhaltsanalyse: Grundlagen und Techniken.* Weinheim/Basel: Beltz.

Nerima, Luka et al. (2006): « Le problème des collocations en TAL ». *Cahiers de linguistique française* 27: 65–115.

Read, John (2000): *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Schmid, Monika S/Jarvis, Scott (2014): "Lexical access and lexical diversity in first language attrition". *Bilingualism: Language and Cognition*, 17/4: 729–748.

Ure, Jean (1971): "Lexical density and register differentiation". In: George Ernest Perren/John Leslie Melville Trim (eds): *Applications of linguistics. Selected papers.* Cambridge University Press : 443–452.

Vanhove, Jan (2018): *Using text-based indices to predict human ratings of the lexical richness of short French, German, and Portuguese texts written by children*. Technical Report. https://osf.io/6bywg/?show=revision&view_only=c5fd1fe6c76642d4a028ec7be1e482a3

Vanhove, Jan et al. (2019): "Predicting perceptions of the lexical richness of short French, German, and Portuguese texts". *Journal of Writing Research* 10/3: 499–25.

Verbi Software (1989-2019): MAXQDA, Software für qualitative Datenanalyse. Consult. Sozialforschung GmbH, Berlin, Deutschland.

Annexe 1 : instructions pour les évaluateurs (version française, sans le questionnaire biographique)

Instructions pour la pensée à haute voix :

Le but de cette étude est de décrire la façon dont les personnes évaluent des textes écrits selon certains critères. La meilleure manière de découvrir cela est de leur demander de penser à haute voix en même temps qu'ils évaluent ces textes. Quand vous pensez à haute voix, l'objectif est de verbaliser le plus possible ce que vous pensez. À peu près tout ce que vous direz nous sera utile à la compréhension de la façon dont vous procédez. Nous savons que vous ne pourrez pas tout verbaliser mais l'idée est de nous donner le meilleur compte-rendu de votre réflexion ; il faudrait alors parler tout au long de la tâche.

N'essayez pas de planifier ce que vous allez dire. Nous n'avons pas besoin d'avoir un discours structuré, le plus important pour nous d'obtenir une vision précise de vos pensées, même d'une partie de celles-ci. Idéalement, vous devriez essayer de dire directement ce qu'il vous passe par la tête sans trop vous demander comment ça va « sortir ». En effet, le plus important est d'évaluer ces textes au mieux sans trop vous soucier de la façon dont vous verbaliserez ce que vous faites. Ne vous mettez pas de pression dans la résolution de la tâche ; il n'y a pas de bonne ou mauvaise réponse.

Partie 1 : Evaluation globale

Vous allez être confronté à deux très courtes productions écrites par des enfants qui ont entre 8 et 10 ans. L'une est une argumentation, l'autre un récit. Lisez rapidement les deux productions écrites. L'orthographe et la grammaire ont été corrigés. Dès que vous avez fini de les lire, estimez la richesse du vocabulaire de chaque couplet sur une échelle de 1 à 9, où 1 est le score le plus bas et 9 le score le plus haut. 15 textes sont à évaluer.

Salut tata, moi je préfère l'avion parce qu'ils passent donner à boire et à manger, on a pas besoin de s'arrêter pour aller aux toilettes et c'est plus vite. Moi, je veux pas aller en voiture parce que tu dois t'arrêter pour aller aux toilettes, tu as pas besoin de t'arrêter pour manger ou boire et en plus tu prends plus de temps pour arriver. Tchao tata j'espère que tu te décides moi en tout cas je préfère en avion et j'aimerais bien que tu sois d'accord avec moi. Bisou bisou. pense et dis-moi quelque chose à toute.

On a fait une cours le long du lac de Fribourg. On était en shorts et en t-shirts avec une gourde et c'était très cool,

1	2	3	4	5	6	7	8	9

Ne faite qu'une seule croix, s.v.p. !

Votre perception intuitive de la richesse du vocabulaire est ce qui nous intéresse. Essayez donc de ne pas trop réfléchir, mais essayez d'être consistant.

Pour vous aider à calibrer votre jugement, voici les instructions que les élèves ont reçues pour les deux textes :

Exercice 1 : Ecrire une lettre **Préfères-tu voyager en voiture ou en avion ?**	Exercice 2 : Ecrire un article **Ta dernière course d'école**
Ta tante t'a invité pour passer les vacances avec elle. Elle n'a pas encore décidé du moyen de transport et veut savoir ton opinion. **Préfères-tu voyager en voiture ou en avion ?** Ecris une lettre à ta tante où tu lui explique ce que tu préfères. N'oublie pas de lui donner **au moins trois raisons** pour lesquelles tu préfères l'avion ou la voiture. **Essaie de la convaincre !**	Le prochain numéro d'un magazine pour enfants sera consacré aux courses d'école. Tu vas écrire un article pour ce magazine pour y raconter ta dernière course d'école. **Repense à ta dernière course d'école : Où es-tu allé ? Qu'est-ce qui s'est passé ? Qu'as-tu fait pendant cette sortie de classe depuis le début jusqu'à la fin de la journée ?** Raconte aux lecteurs du magazine le plus possible de détails sur cette journée.

Partie 2 : Evaluation selon cinq critères

Maintenant, nous vous prions d'évaluer ce même set de 15 couplets. Cette fois cependant, quatre critères plus précis devront être notés par paire de textes sur l'échelle de 1 à 9.

1) Sophistication lexicale (= utilisation de mots rare)
   (De 1= utilisation de vocabulaire très fréquent à 9= utilisation de vocabulaire très rare) → idem pour les autres critères
2) Diversité lexicale (=la proportion/nombre de mots différents)
3) Utilisation appropriée du vocabulaire
4) Complexité syntaxique (=complexité de la structure des phrases telles que des subordonnées etc.)
5) Contenu/réponse à la tâche

Merci de votre aide !

Annexe 2 : entretien rétrospectif, réponses à la question

Comment avez-vous procédé pour évaluer les textes lors de la première partie ? Aviez-vous une stratégie ?

Sous-catégories	Transcriptions
Intuition	TAP1 : *Ich glaube nicht. Also ich glaube, dass es für mich ein ganz neues Feld ist und ich keine Ahnung und Bildung davon habe, wie man sowas beurteilen sollte. Ich glaube es ist einfach nur ein Eindruck.* Traduction : Je ne crois pas. Alors, je crois que c'est pour moi un tout nouveau champ et que je n'ai aucune idée et formation à ce propos, comment on devrait

	évaluer quelque chose comme ça. Je crois que c'est simplement juste une impression.
	TAP3 : *Pas de technique en fait. () Je ne sais pas trop, mais () je dirais que c'est quand même quelque chose d'automatique*. Chercheuse : *donc première impression ?* TAP3 : *Oui première impression, j'ai bien réussi à voir. L'impression, c'est que voilà.*
Structure, grammaire et vocabulaire	TAP4 : *Eben ich habe ähnliche Kriterien genommen, wie die in der (?ersten?) zweiten Aufgabe vorgelegt wurde. Wie vorher gesagt, wichtig ist mir vor allem bei so was, dass thematisch und logisch dieser Text angegangen wird, dass der oder die SchülerIn sich so ausdrückt, dass man versteht, was sie sagen möchte. [...] Vor allem grammatikalisch ein paar kleine Ungereimtheiten drin sind, finde ich das nicht so wichtig. Was aber sehr positiv auffällt, ist wenn_ Also ich habe es lieber, wenn die Schüler versucht, eine eher ()* ***schwierigere grammatikalische Form aufzubauen*** *und dabei der eine oder der andere Fehler machen, als wenn Unstrukturen [sic.] nacheinandergeheiht werden und dafür fehlerfrei, [...]. So was finde ich zum Beispiel sehr wichtig. Dann, was habe ich noch in Betracht bezogen? Natürlich auch, wie lange ist der Text und wie logisch ist das, was in dem Text geschrieben wird, hat definitiv einen Einfluss gehabt. Und natürlich auch eben Wortwahl.* ***Wortwahl*** *ist auch sehr wichtig. Also wenn ein Wort kommt, dass ich von jemandem in dem Alter oder in dem Sprachniveau nicht erwarten würde, dann bin ich immer sofort positiv überrascht und gehe dann natürlich auch sehr wesentlich weniger streng an den Rest des Textes.* Traduction : Justement, j'ai pris des critères similaires que ceux de la deuxième tâche. Comme je l'ai déjà dit, dans ce type d'exercice, il est important pour moi que le texte soit abordé de façon thématique et logique, que l'élève s'exprime de telle façon qu'on comprenne ce qu'elle/il veut dire. [...] Surtout s'il contient quelques imprécisions grammaticales, je ne le trouve pas important. Ce qui par contre se fait remarquer très positivement, c'est quand_ Alors, je préfère quand l'élève produit des tournures grammaticales plutôt difficiles et fait du coup l'une ou l'autre faute, que [des tournures] sans structure et sans faute, [...]. Je trouve quelque chose comme ça très important, par exemple. Ensuite, à quoi ai-je encore fait attention ? Naturellement, la longueur et la logique du texte, ce qui est écrit dans le texte, ont définitivement aussi une influence. Et naturellement, le choix des mots aussi. Le choix des mots est aussi très important. Donc, quelqu'un emploie un mot, alors qu'à son âge ou niveau langagier, on aurait pas attendu cela, alors je suis toujours directement très positivement surpris et je suis essentiellement moins sévère avec le reste du texte.
	TAP5 : *Aber ich würde ja () fast eine drei geben. Obwohl es noch ein bisschen gefährlich ist, finde ich, weil ich habe fast das Gefühl, dass ich langsam die Rhetorik oder auch generell von dem Ausbau und nicht unbedingt nur den Wortschatz. Wahrscheinlich müsste ich fast eigentlich die Wörter zählen. Aber*

	dann kann es auch sein dass ich einen Text von Wortschatz gut bewerte, weil er länger ist, aber vielleicht redundant. […] Ich tue mich sehr schwer zu bewerten [...]. (Remarque: la question générale n'a pas été posée à TAP5, car il mène déjà cette réflexion profonde lors de l'évaluation holistique et cela aurait paru étrange dans le flux de l'entretien). Traduction : Mais je donnerais presque un trois. Bien que cela soit un peux dangereux, j'ai presque le sentiment que je [fais attention] de plus en plus à **la rhétorique ou plus généralement à la construction et pas forcément au vocabulaire.** Je devrais probablement compter les mots en fait. Mais du coup, je pourrais bien noter le vocabulaire d'un texte parce qu'il est plus long, mais il sera peut-être plus redondant. (..) C'est difficile pour moi d'évaluer [...].
Comparaison entre les textes	TAP2 : *[...] Also habe ich mir die **Satzstruktur** angeguckt, und das ist, was mir eingefallen ist **und () den Wortschatz. Und dann habe ich verglichen**. […] Insofern verglichen, dass ich geguckt, **was habe ich bei anderen kritisiert und habe das dann mit dem Text, den ich gerade bewertet habe, versucht da auch anzuwenden und da auch zu vergleichen**. Und sonst habe ich mir **viel die Satzstruktur** angeguckt.* Traduction : [...] Alors j'ai beaucoup regardé la structure des phrases et c'est ce qui m'est venu à l'esprit et () le vocabulaire. Et ensuite j'ai comparé. (..). Par comparer, je veux dire que j'ai regardé ce que j'ai critiqué chez les autres et ai ensuite essayé d'employer cela aussi avec le texte que j'étais en train d'évaluer. Et sinon, j'ai beaucoup regardé la structure des phrases.
	TAP6 : *Ce que je fais d'habitude, c'est que le premier, je fais un peu à l'instinct en fait. Je me dis, bon **qu'est-ce qui serait parfait** ? et puis **par rapport au parfait**, à combien il est en-dessous ? Et par rapport à celui-là, je **compare les autres** et puis, je pense que ça m'arriverais de revenir en arrière. Si j'arrive à un qui est plus proche du but, **ça m'arriverais de revenir en arrière et de refaire carrément pour recomparer**.*
Profil de l'élève	TAP7 : *D'abord j'ai pensé () **émotionnel**. Je me dis que **c'est un enfant de 8 à 10 ans** et () je trouve qu'ils ont quand même pas mal développé () les réponses. () Bon, peut-être que je devrais_, j'aurais dû lire une fois tous les textes de tout le monde pour bien évaluer les compétences de chacun. Après, chaque personne () m'étonne toujours une petit peu, ce qu'ils donnent comme réponse.*

Tableau 2 : extraits des réponses à la question « Comment avez-vous procédé pour évaluer les textes lors de la première partie ? Aviez-vous une stratégie ? » lors de l'entretien rétrospectif. Pour chaque participant, une catégorie majeure a été créée afin de donner ici une vue d'ensemble des critères conscientisé. Evidemment, la plupart des extraits se chevaucheraient entre plusieurs catégories dans un codage plus fin.

Annexe 3 : (sous-)catégories avec description ainsi que le nombre de segments codés

MAXQDA 02.06.2017

Code System [164]
- **Question-_méthode [0]**
  - **Méthode_intuition [2]**
  - **Méthode_comparaison [2]**
  - **Méthode-vocabulaire [1]**
  - **Méthode_profil [1]**
  - **Méthode_grammaire [1]**
  - **Méthode-structure [1]**
- **Structure [0]**

Aspects plus sémantiques et pragmatiques

- **Longueur du texte [8]**
- **Contenu [17]**

Réponse à la tâche (p.ex. 3 arguments)
Nombre d'idées

- **Genre textuel [10]**
- **Cohérence-cohésion [17]**

Manque de logique dans l'alignement des évènements
Texte compréhensible
Sens du texte
Ponctuation

- **Justesse et complexité linguistique [26]**

Longueur des phrases
Structure des phrases
Fautes de syntaxe
Elégance vs. inélégance
Justesse
Temps verbaux

- **Comparaison_textes [0]**
  - **Différences_deux_textes [7]**

Comparaison entre les deux textes du même élève (argumentation et narration)

- **Différences_entre_élèves [18]**
  - **Négatif [6]**

Absence de comparaison possible

- **Profil locuteur [14]**

Aspects cognitifs (pensée (typique) de l'enfant, humeur, affects, compréhension de la tâche)
Aspects linguistique (Language utilisé par les enfants)

- **Vocabulaire [0]**
  - **Voc_répétitions/variation [13]**

Diversité
Quantité de vocabulaire
Répétition d'expressions

- **Voc_autre [20]**

Qualité (p.ex oralité)

Difficulté
Autres qualificatifs du vocabulaire (riche, etc.) - mais seulement si cet adjectif n'est pas directement expliqué par la répétition
Utilisation de mots "inacceptables" ou au contraire adaptés au contexte
Manque de vocabulaire
Vocabulaire en général
Arguements imprécis (p.ex.: bien/pas bien)