Transformations syntaxiques pour une aide à l'apprentissage de la lecture : typologie, adéquation et corpus adaptés

Núria Gala[1,*], *Amalia* Todirascu[2], *Delphine* Bernhard[2], *Rodrigo* Wilkens[2], et *Jean-Paul* Meyer[2]

[1]Aix Marseille Univ., Laboratoire Langage et Parole (LPL-CNRS), 13090 Aix-en-Provence, France
[2]Université de Strasbourg, Linguistique, Langues, Parole (LiLPa), 67084 Strasbourg, France

Résumé. Dans cet article, nous présentons une typologie de transformations syntaxiques permettant une adaptation des contenus textuels à destination d'enfants faibles lecteurs et dyslexiques. Pour arriver à cette proposition, nous avons analysé des textes parallèles originaux et adaptés. Nous avons aussi appliqué des transformations lexicales, morpho-syntaxiques et discursives à des corpus habituellement lus entre CE1 et CM1 que nous avons soumis à des enfants dans ces classes, tous profils confondus. Sur la base de ces deux études, nous avons défini une typologie de transformations syntaxiques, avec des informations supprimées, conservées ou ajoutées, qui pourra servir de guide pour adapter des textes et faciliter l'apprentissage de la lecture dans des cas d'enfants en difficulté.

Abstract. Syntactic transformations to help learning to read : typology, adequacy and adapted corpora. In this paper, we present a typology of syntactic transformations targeted at adapting textual contents addressed to poor-readers and dyslexic children. To make this proposition, we have analyzed a set of parallel texts (original and adapted). We have also applied lexical, morpho-syntactic and discursive transformations to corpora usually read at primary school (second to fourth grades). The different versions have been read by different reader profiles at school. Based on both studies, we have defined a typology of syntactic transformations, with deleted, kept or added information, that could be used as guidelines to adapt texts and facilitate reading to children facing difficulties.

1 Introduction

Cet article propose une caractérisation de phénomènes syntaxiques présents dans des corpus dits 'simplifiés' ou 'adaptés'. Ces corpus, issus de publications existantes ou transformés à partir de textes originaux, se destinent à des publics ayant des difficultés lors de

* Auteur pour correspondance : nuria.gala@univ-amu.fr

l'apprentissage de la lecture, principalement des enfants faibles lecteurs ou dyslexiques, mais aussi, éventuellement, des adultes illettrés. À notre connaissance, il n'existe pas dans la littérature un état de l'art exhaustif visant une typologie de corpus et de phénomènes linguistiques destinés à faciliter la lecture par ce public.

Notre proposition de typologie est issue d'un travail empirique réalisé dans le cadre du projet ANR ALECTOR[i]. D'une part, des analyses de corpus édités ont été menées. D'autre part, nous avons manuellement adapté 79 textes que nous avons fait lire à des enfants scolarisés (1.057 enfants de CE1 à CM1, tous profils confondus) et à des enfants bénéficiant d'une remédiation orthophonique (30 enfants faibles lecteurs ou dyslexiques d'entre 9 et 12 ans, ayant principalement des difficultés au niveau du décodage). L'objectif de cette expérience était d'identifier les phénomènes entraînant des difficultés lors de la lecture et la compréhension des textes. À l'issue de ces études, nous avons élaboré une typologie de phénomènes spécifiques ayant un impact dans l'amélioration de l'activité de lecture, principalement au moment de son apprentissage par des jeunes enfants (cf. projet ALECTOR). En particulier, nous étudions ici les transformations identifiées lors de l'étude de corpus opérées au niveau syntaxique. Nous présentons la typologie réalisée à la suite de cette étude et la méthodologie mise en place.

Dans un premier temps (section 2) nous exposons le contexte de notre travail. Dans la section 3, nous présentons la méthodologie, les corpus utilisés et l'analyse des transformations syntaxiques identifiées dans les corpus. La section 4 propose la typologie de transformations que nous avons définie d'un point de vue des opérations. Enfin, la section 5 présente une typologie linguistique des transformations syntaxiques proposées.

2 Contexte

Les initiatives pour proposer des guides permettant d'adapter les contenus didactiques à un public en difficulté (enfants dyslexiques, avec handicaps cognitifs, avec autisme, ...) se multiplient. Cependant, la plupart de ces initiatives agissent uniquement sur la forme des textes. Il s'agit, par exemple, d'extraits d'œuvres originales (avec des suppressions de passages considérés moins informatifs), comme dans des éditions destinées à des publics jeunes, ou bien des aménagements visuels (taille et couleur des caractères), comme dans les éditions destinées à des publics en difficulté de lecture.

Dans cette section, nous décrivons plusieurs initiatives (2.1) et nous présentons les objectifs du projet ANR ALECTOR (2.2) où l'on vise à transformer non seulement le forme mais aussi le contenu (lexique et structure) des phrases, dans le but de proposer des textes adaptés à des apprentis lecteurs en difficulté.

2.1 Aménagements et outils pour des publics en difficulté de lecture : quelques initiatives existantes

Au niveau de la Commission Européenne, il existe des projets destinés à accompagner les personnes ayant un handicap intellectuel dans le processus d'apprentissage de la lecture[ii]. Dans le cadre de ces initiatives, des guides de rédaction en Français Facile à Lire et à Comprendre (FALC) indiquent des recommandations concernant la présentation et le contenu des productions écrites ou publiées sur Internet. Ainsi, il est recommandé de « privilégier les mots simples », et « les mots complexes doivent être systématiquement expliqués », sans faire appel aux métaphores (encore faut-il identifier qu'est-ce qu'un mot 'simple' ou un mot 'complexe'). Les phrases doivent être courtes. Il faut également privilégier les phrases actives à la place des phrases passives, ainsi que les phrases positives à la place des phrases négatives. Il faut, enfin, répéter les informations plus importantes et ajouter des explications aux mots difficiles à comprendre (par exemple, les mots rares).

Les associations proposant des exercices et des textes à destination des enseignants et des parents d'enfants dyslexiques décrivent aussi les transformations appliquées sur les textes pour le public dyslexique (par exemple, la *British Dyslexia Association*[iii]). Dans la grande majorité, ils proposent des aides au niveau de la typographie et de la forme du texte (police, espacement de caractères, espacement entre lignes) et conseillent des couleurs différentes pour les syllabes et/ou lignes. Dans des cas spécifiques comme dans la collection *Colibri* (éditions Belin), les textes sont proposés selon des niveaux de difficulté en fonction de la fréquence des correspondances graphème-phonème en français et leur complexité croissante (par exemple, le graphème '-eau' correspond au premier niveau, 'ss' ou deuxième, 'rr' ou troisième et 'gr' au quatrième).

Globalement, ces initiatives ne transforment pas le contenu d'un texte. Lorsqu'il y a des modifications par rapport à un texte original, il s'agit de proposer des transformations lexicales telles que donner des explications pour les mots et les expressions idiomatiques, ou remplacer les mots complexes (abstraits, ou conceptuellement difficiles) par des mots plus simples. D'autres transformations s'appliquent au niveau syntaxique ou discursif, comme découper les phrases complexes (longues) en phrases plus simples, transformer les phrases passives en phrases actives, etc. Cependant, ces recommandations s'appliquent dans des cas très particuliers, et peuvent varier d'une association à l'autre.

Par ailleurs, il existe des systèmes et des outils de simplification automatique de textes qui reposent sur des typologies des transformations réalisées sur les corpus. Les aménagements proposés interviennent surtout sur le lexique ou la syntaxe (Saggion 2017). Par exemple, Brunato et ses collaborateurs (2014) proposent des macro-catégories d'opération de simplification automatique : découpage (au niveau des conjonctions ou des relatives), regroupement dans une seule phrase, changement d'ordre (syntaxique), insertion (sujet, verbe, autres constituants), suppressions (modifieurs, sujet, verbe), transformations (lexicales, morpho-syntaxiques et syntaxiques). Les méthodes de simplification utilisant les approches à base de traduction automatique considèrent que le langage simplifié est le résultat d'une traduction du langage général vers le langage simplifié (Alva-Manchego et al., 2017). On retrouve les mêmes catégories de transformation identifiées parmi les alignements : le découpage (alignement 1-N) et le regroupement (alignements N-1). De plus, dans les alignements 1-1 de phrases différentes, ils détectent la suppression (si le mot est présent dans le texte original et non aligné dans le texte simplifié), la réécriture (si un mot est remplacé par un autre mot isolé), l'insertion (si le mot est non aligné et présent uniquement dans le texte simplifié), ou le remplacement (si on remplace un groupe de mots par un autre). Quant à Koptient et ses collaborateurs (2019), ils proposent une typologie de simplifications adaptée aux textes de spécialité (techniques, et plus particulièrement les textes médicaux). Parmi les transformations appliquées, ils identifient la substitution lexicale (avec ou sans glissement sémantique), l'ajout d'explications, la substitution syntaxique (transformer la phrase passive vers la phrase active), la pronominalisation et la suppression d'informations.

Ces typologies, pensées pour être mises en œuvre de façon (semi-)automatique, ont comme transformations communes : (1) les substitutions lexicales et syntaxiques avec ou sans glissement sémantique, (2) l'ajout d'information et (3) la suppression de contenu.

2.2 Aide à la lecture

Notre objectif de départ était l'identification de phénomènes linguistiques qui ont un impact dans la lecture et la compréhension des textes par les enfants faibles lecteurs. Dans cette optique, nous avons créé des ressources linguistiques (un corpus de textes parallèles, voir section 3.2) que nous avons soumis à des enfants. Le résultat de ce travail nous a permis de poser les bases pour la définition d'un outil d'aide à la lecture. *In fine*, l'idée est de développer un système qui, s'appuyant sur ces ressources et sur le guide de transformations

que nous présentons dans cet article, adapte (semi-)automatiquement les textes de façon à produire des versions adaptées (simplifiées). Ces versions ont pour but d'être utilisées dans le cadre d'un **entraînement à la lecture**. Le but n'est pas d'appauvrir les textes *per se* mais de proposer des aménagements qui permettent à un enfant en difficulté de s'entraîner pour mieux lire, c'est-à-dire, d'acquérir les mécanismes nécessaires pour **comprendre ce qu'il lit**.

Dans le cas concret des enfants dyslexiques, les difficultés pour l'acquisition de la lecture ont été largement décrites dans la littérature (par exemple, Rello, 2014; Ziegler et al., 2014; Gala et Ziegler, 2016). Du fait d'un décodage lent et laborieux, plus une unité est longue, plus elle est difficilement lue. Par exemple les mots longs (plus de 7 caractères), les syllabes complexes (avec des irrégularités phonème-graphème) et les phrases longues (avec plusieurs propositions) sont des écueils récurrents. Ainsi, pour ce public, il est préconisé de mettre en place des transformations de type (1) substitution et (3) suppression.

Dans ce qui suit, nous présentons une typologie de transformations syntaxiques réalisée sur la base d'analyses de corpus originaux et transformés et visant à alléger les textes sans pour autant diminuer leur contenu sémantique.

3 Méthodologie et corpus

Afin de proposer un ensemble de transformations pour obtenir des textes simplifiés à partir de textes originaux, nous avons étudié un ensemble de textes édités à destination d'enfants dyslexiques et normolecteurs (3.1). Nous avons aussi étudié, d'une part, les recommandations existantes dans la littérature (guides d'écriture provenant d'associations sur la dyslexie, publications scientifiques comme Rello 2014 ou Ziegler et al., 2014). D'autre part, nous avons analysé les écrits littéraires destinés à des publics adultes illettrés (concrètement, ceux de la collection *La Traversée*[iv]). En tenant compte les caractéristiques de ces deux ensembles d'ouvrages, nous avons nous-mêmes simplifié 79 textes originaux pour enfants (3.2). Notre objectif était de créer un corpus parallèle testé dans des écoles et cabinets d'orthophonie et pouvant servir de référence pour des futurs développements en simplification automatique de textes.

Pour identifier la typologie des transformations morpho-syntaxiques et syntaxiques, nous avons, ainsi, comparé des textes originaux et des textes transformés. D'une part, cette comparaison a été réalisée au niveau des opérations de transformation (substitution et suppression, principalement). D'autre part, nous avons étudié les différences de catégories lexicales dans les corpus que nous avons rassemblés. L'analyse des corpus alignés a permis de mettre en évidence des transformations lexicales, morpho-syntaxiques et discursives.

Dans la suite de cet article, nous présentons en détail les transformations syntaxiques identifiées dans les corpus et appliquées pour valider nos hypothèses. À partir de l'analyse de corpus et des expériences réalisées avec les tests de lecture, nous avons établi la typologie de transformations syntaxiques que nous présentons dans la section 4.

3.1 Analyse du corpus CONTES : textes édités à destination d'enfants dyslexiques ou normolecteurs

Le corpus que nous appelons 'CONTES' est constitué de neuf paires de textes originaux et simplifiés (contes pour enfants, niveau CE1-CM2, dont, par exemple, deux contes de Charles Perrault niveau CM2). Le Tableau 1 récapitule le nombre de tokens et de phrases dans l'ensemble de ce corpus.

Tableau 1. Nombre de tokens et de phrases dans les corpus CONTES.

Corpus CONTES	Tokens	Phrases	Tokens/phrase
Textes originaux	12.223	770	15,87
Textes simplifiés	9.792	684	14,32

Il est possible de constater que les textes simplifiés sont généralement plus courts pour ce type de public, les suppressions et les substitutions par des unités plus courtes sont privilégiées (cela peut ne pas être le cas pour d'autres types de public où des ajouts peuvent être considérés comme utiles pour améliorer la lecture et la compréhension de textes).

Dans un deuxième temps, nous avons comparé la distribution des catégories de discours (verbe, nom, adjectif, adverbe, pronom, conjonction de coordination et de subordination) dans les textes originaux et simplifiés. Il est possible de faire état de plusieurs différences : le pourcentage de certains types de pronoms (et en particulier celui de pronoms relatifs) dans les textes simplifiés est inférieur à celui des textes originaux (1,77% de pronoms relatifs dans les textes originaux et 1,58% dans les textes simplifiés), le nombre d'adverbes est aussi réduit (de 7,36 % dans les textes originaux à 6,69 % dans les textes simplifiés), et c'est également le cas pour les conjonctions de coordination (3,25 % dans les textes originaux contre 2,85 % dans les textes simplifiés). La fréquence relative des noms communs (15,44 % dans les textes simplifiés contre 13,39 % dans les textes originaux) et des articles définis (6,54 % dans les textes simplifiés contre 5,82 % contre dans les textes originaux) est plus importante pour les textes adaptés. Ces différences s'expliquent par les transformations réalisées dans le texte : découpage des phrases longues, suppression de subordonnées relatives, transformation du discours rapporté en discours direct, remplacement des pronoms par des groupes nominaux.

L'analyse de la distribution des catégories lexicales montre qu'il y a bien un lien entre type de texte (original / simplifié) et distribution des catégories (test exact de Fisher ou du χ^2, $p<0,05$). Cela étant, les catégories qui contribuent le plus au χ^2 global sont très peu nombreuses : on constate une association négative entre la catégorie Nom et les texte originaux et, à l'inverse, une association positive entre la catégorie Nom et les textes simplifiés (fréquence relative de 13,39 % *vs*. 15,44 %). À l'inverse, la présence de noms propres est associée de manière positive aux textes originaux et de manière négative aux textes simplifiés (fréquence relative de 3,10 % *vs*. 1,72 %). La présence plus importante de noms communs est donc une des caractéristiques des textes simplifiés, tandis que celle des noms propres est une caractéristique des textes originaux. Il ne faut toutefois pas oublier que ces analyses sont réalisées à partir d'une annotation automatique des corpus en catégories lexicales, annotation qui n'est pas toujours à cent pour cent exacte. Nous avons donc procédé à des analyses supplémentaires : la première est essentiellement manuelle et a été réalisée sur une sous-partie du corpus, et la seconde est entièrement automatique et a donc pu être effectuée pour tout le corpus.

Nous avons utilisé l'outil MEDITE[v] (Fenoglio et Ganascia, 2007) pour repérer les transformations (suppression de séquence de mots, remplacement de mots, changement de l'ordre des mots) effectuées dans le texte simplifié par rapport au texte original. Sur la base de ces observations, nous avons aligné manuellement 98 paires de phrases (provenant de 7

paires de textes) et nous avons repéré les opérations effectuées au niveau lexical, syntaxique ou discursif. Le choix des phrases a été fait compte tenu des informations qui ont été conservées. Parmi ces alignements, il y a **46,31 % de transformations syntaxiques**. La plupart du temps nous avons observé des transformations complexes qui impliquent plusieurs catégories de transformations (suppression d'information, reformulation, ajout d'information). Les remplacements représentent 51,28 % des transformations syntaxiques, suivies par 25,64 % des suppressions d'informations secondaires (propositions relatives, adverbes, constructions avec participes, modifieurs, etc.). Les subordonnées relatives font souvent l'objet de transformations, soit par réécriture, soit par suppression (environ 25 % des modifications). Les découpages des phrases complexes (quand il y a coordination ou signe de ponctuation) représentent 17 % des transformations.

Nous avons également développé une méthode d'analyse automatique des transformations entre corpus original et corpus simplifié. Dans un premier temps, les corpus sont analysés avec StanfordNLP (Qi et al., 2018) selon la chaîne de traitement suivante : tokenisation, étiquetage morphosyntaxique, lemmatisation et analyse en dépendances. La lemmatisation produite par StanfordNLP est améliorée par l'utilisation de Flemm (Namer, 2000). Les textes originaux et simplifiés sont ensuite alignés automatiquement avec CollateX (Dekker et Middell, 2011)[vi], qui est un outil similaire à MEDITE. Enfin, les informations issues de l'annotation automatique sont utilisées pour classer automatiquement les changements repérés par CollateX dans diverses catégories. La figure 1 ci-dessous montre les opérations de transformation les plus fréquentes observées sur le corpus. La plus fréquente est split, qui correspond à un découpage de phrases.

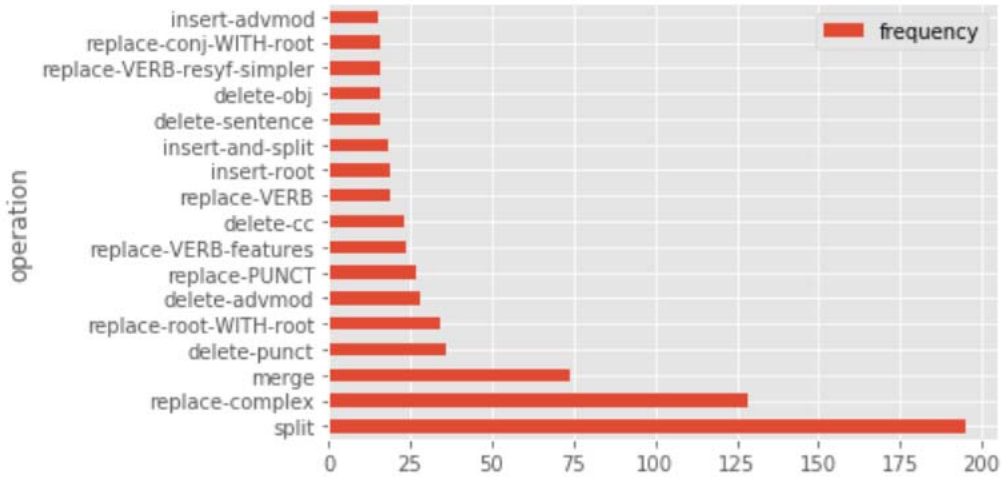

Fig. 1. Fréquence des opérations de transformation syntaxique dans les corpus CONTES.

Enfin, l'exemple proposé par la Figure 2 montre une table d'alignement produite par CollateX. Deux opérations sont identifiées par notre méthode : *split* (découpage de phrases) et *delete-cc* (suppression d'une conjonction de coordination).

W1	Ce pauvre enfant était le souffre-douleur de la maison	, et	on lui donnait toujours tort.
W2	Ce pauvre enfant était le souffre-douleur de la maison	.	On lui donnait toujours tort.

Fig. 2. Alignement phrase originale *vs*. phrase simplifiée avec l'outil CollateX.

3.2 Création et analyse du corpus ALECTOR : corpus parallèle pour enfants de sept à neuf ans

Le corpus ALECTOR est constitué de 183 textes différents (52.704 tokens au total), dont 79 textes originaux et leurs équivalents simplifiés (Gala et al. 2020). Il s'agit d'un ensemble de textes proposés dans les classes de CE1 à CM1, aussi bien littéraires (contes, fables, histoires) que documentaires (sciences de la vie et de la terre, majoritairement). L'ensemble de ces textes a été manuellement simplifié et proposé comme tests de lecture lors d'une étude longitudinale de 3 ans visant à évaluer les bénéfices de la simplification de textes dans l'apprentissage de la lecture[vii].

Tous les corpus originaux ont subi des transformations 'à tous les niveaux' (lexical, morpho-syntaxique, discursif). Par ailleurs, 25 corpus de CE1 ont subi différentes opérations de simplification (uniquement lexicale et/ou uniquement syntaxique). Afin d'établir une comparaison avec le corpus CONTES présenté à la section 3.1, le tableau 2 montre les données concernant les 79 corpus originaux comparés à leurs 79 versions simplifiées 'à tous les niveaux' :

Tableau 2. Nombre de tokens et de phrases dans le corpus ALECTOR : 158 textes au total (79 textes originaux avec leurs versions simplifiées au niveau lexical, morpho-syntaxique et discursif).

Corpus ALECTOR	Tokens	Phrases	Tokens/phrase
Textes originaux	24.146	1.557	15,5
Textes simplifiés	22.097	1.767	12,5

Si on tient compte de la totalité du corpus, c'est-à-dire 79 corpus originaux et 104 versions simplifiées (79 'à tous les niveaux', 15 simplifiés uniquement au niveau du lexique, 10 corpus simplifiés uniquement au niveau des structures syntaxiques), on obtient 104 paires de textes constitués tel que le montre le tableau 3 (au total, 25 textes originaux ont différentes versions de simplification):

Tableau 3. Nombre de tokens et de phrases dans le corpus ALECTOR complet : 208 textes au total (104 paires de textes originaux et simplifiés).

Corpus ALECTOR	Tokens	Phrases	Tokens/phrase
Textes originaux	36.071	2.374	15,19
Textes simplifiés	32.721	2.492	13,13

Dans les deux corpus, CONTES et ALECTOR, on observe une baisse importante du nombre de tokens par phrase entre les deux versions. Cependant, à la différence du corpus CONTES, le nombre de phrases dans les textes simplifiés du corpus ALECTOR est supérieur à celui des textes originaux. En effet, comme on peut le voir à la suite des mêmes traitements (alignements de corpus, classements automatiques des changements repérés par CollateX) l'augmentation du nombre de phrases s'explique par une tendance significative au découpage (*split* demeure l'opération de transformation la plus fréquente, comme dans les textes CONTES). Néanmoins, alors que le corpus CONTES fait état de l'opération

delete sentence, cette transformation n'a pas été mise en place dans le corpus ALECTOR (toutes les phrases ont été conservées).

Les figures 3 et 4 présentent les différentes opérations de transformation dans les corpus littéraires et scientifiques, en tenant compte du niveau du texte :

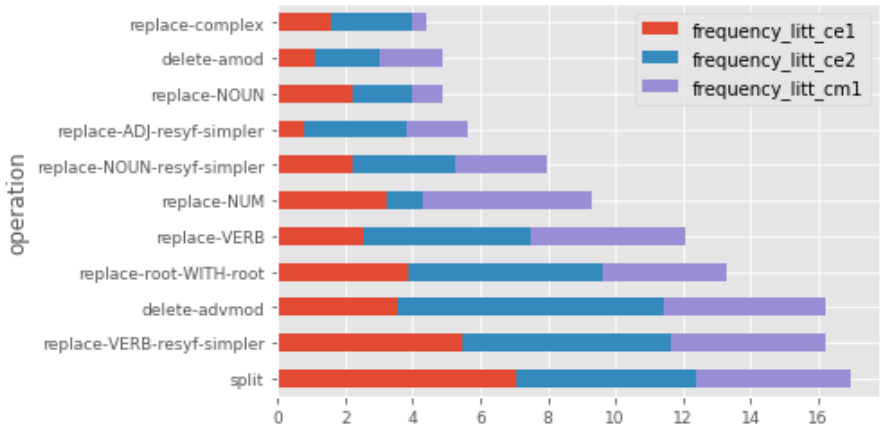

Fig. 3. Fréquence des opérations de transformation syntaxique dans le corpus ALECTOR (textes littéraires).

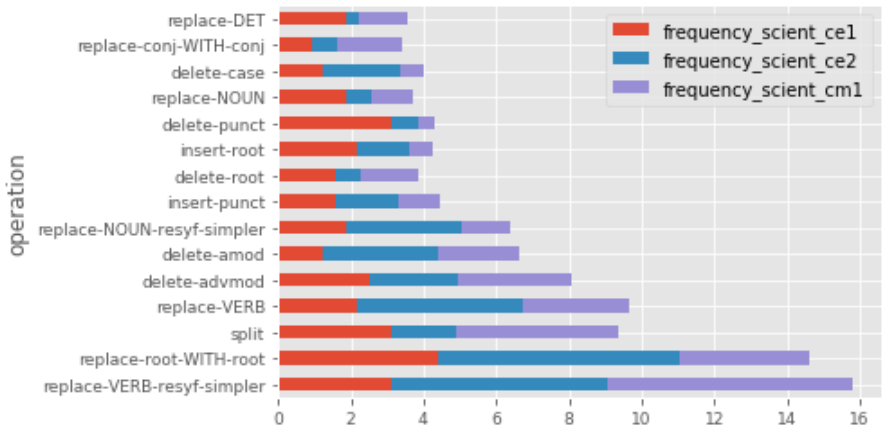

Fig. 4. Fréquence des opérations de transformation syntaxique dans le corpus TEXTES (textes documentaires scientifiques).

Comme on peut le voir, l'opération *split* est parmi les trois opérations les plus fréquentes dans les deux types de corpus. Par ailleurs, certaines opérations de transformation sont lexicales : elles incluent la substitution ou la suppression d'une catégorie lexicale (verbe, nom, adjectif). La ressource de synonymes gradués ReSyf[viii] (Billami et al., 2018) a été utilisée comme référence dans ces cas-là, par exemple pour les opérations *replace-Noun-resyf-simpler*, *replace-verb-resyf-simpler*.

Pour ce qui est de l'étude des catégories du discours, nous avons aussi comparé leur distribution dans les deux versions, originale et simplifiée. On observe, tout comme pour le corpus CONTES, une augmentation des noms communs dans les corpus simplifiés (due, notamment, à la substitution de pronoms par leurs référents). Le nombre de verbes (formes personnelles) augmente aussi au détriment des verbes à l'infinitif (formes non personnelles). En revanche, le nombre de pronoms, de conjonctions, d'adjectifs et

d'adverbes diminue, le premier à cause de remplacements par noms, les trois derniers par des suppressions.

De même, comme pour le corpus CONTES, l'analyse de la distribution des catégories dans le corpus ALECTOR montre qu'il y a bien un lien entre type de texte (original / simplifié) et distribution des catégories (test du χ^2, $p<0,05$). Les catégories qui contribuent le plus au χ^2 global sont toutefois différentes. Les pronoms relatifs (version originale 1,15 % *vs* version simplifiée 0.86 %) sont associés de manière positive aux texte originaux et de manière négative aux textes simplifiés. À l'inverse, les signes de ponctuation (version originale 13,58 % *vs* version simplifiée 14,40 %) sont associés de manière positive aux textes simplifiés et de manière négative aux textes originaux.

Enfin, la figure 5 montre deux versions d'un même texte dans le corpus ALECTOR (les couleurs sont celles de l'outil MEDITE, rouge pour les suppressions, vert pour les insertions et bleu pour les remplacements ; elles sont formelles et ne tiennent pas compte de notre typologie linguistique) :

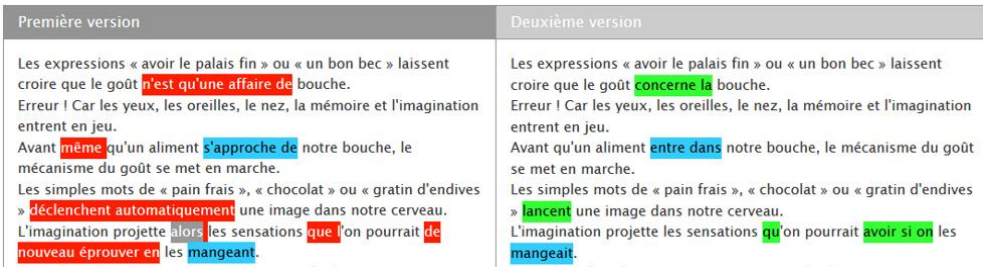

Fig. 5. Exemple d'une version originale *vs*. simplifiée en parallèle, visualisée avec l'outil MEDITE (Corpus ALECTOR 85_SCI CE2 Goût).

4 Transformations syntaxiques : typologie

Comme nous l'avons mentionné plus haut, les opérations de simplification présentés sur les corpus CONTES et ALECTOR s'appliquent à plusieurs niveaux : lexical, morpho-syntaxique et discursif. Dans cet article, nous présentons uniquement une caractérisation au niveau des variations syntaxiques (des propositions concernant des transformations lexicales ont déjà été abordées (Gala et Ziegler, 2016)) ; quant aux simplifications des chaînes de référence, elles ont été traitées dans (Wilkens et Todirascu, 2020) entre autres). Notre travail de caractérisation des variations syntaxiques en vue d'une aide à la lecture a pris en compte des stratégies de transformation automatisables mais aussi des stratégies de reformulation manuelles qui ont été appliquées sur le corpus ALECTOR.

Sur la base de la typologie annoncée plus haut (2.1) et des analyses réalisées sur les corpus, nous proposons une typologie spécifique à la syntaxe avec (4.1) des **substitutions**, (4.2) des **suppressions** et (4.3) des **découpages**. Quelle que soit la transformation, l'objectif est de réduire la complexité de lecture pour faciliter la compréhension du texte, par le biais du maintien, autant que possible, de la structure sujet-verbe-objet (SVO).

4.1 Substitutions morpho-syntaxiques

Au niveau des substitutions, les textes adaptés privilégient les formes actives aux formes actives et les propositions positives aux négatives. Ces deux transformations sont fréquentes dans les textes et recommandations étudiées :

Original. *Ces fèves sont entassées* au soleil, sur des feuilles de bananier.

Simplifié. *On met ces fèves* au soleil, sur des feuilles de bananier.
Exemple 1. Corpus ALECTOR 158_SCI CM1 Chocolat.

Original. Son mari *n'a jamais osé lui dire* que sa soupe avait une odeur qui le gênait.
Simplifié. Monsieur Dupond *ose, pour la première fois, dire* à sa femme qu'il n'aime pas l'odeur de sa soupe.
Exemple 2. Corpus CONTES La soupe de la discorde.

Les constructions modales ou les locutions verbales peuvent être substituées par des formes plus simples. Nous considérons que la perte sémantique est minimale, en revanche, l'effort de décodage est moindre car il y a moins de tokens (unités) à lire :

Original. La température est très basse (…), *elle peut devenir extrêmement* élevée.
Simplifié. La température est très basse (…). Elle devient très élevée.
Exemple 3. Corpus ALECTOR 29_SCI CE1 Mesure.

Original. Un éléphant *vint à passer.*
Simplifié. Un éléphant *passa*.
Exemple 4. Corpus ALECTOR 178_LIT CM1 L'éléphant et l'oiseau.

Les relatives peuvent être supprimées mais aussi transformées en une conjonction entre deux syntagmes :

Original. Il était une fois une pauvre veuve *qui vivait avec* son fils Jacques.
Simplifié. Il était une fois une pauvre veuve *et* son fils Jacques.
Exemple 5. Corpus CONTES Jack et le haricot magique.

Enfin, les attributives peuvent être remplacées par des groupes nominaux.

Original. Ils avaient chacun leur particularité : *un était plutôt naïf, l'autre plutôt peureux et le dernier plutôt bavard.*
Simplifié. Les trois garçons étaient toujours ensemble: *le naïf, le peureux et le bavard*.
Exemple 6. Corpus CONTES Les trois sots.

4.2 Suppressions morpho-syntaxiques

Au niveau des suppressions, les textes adaptés évitent les ajouts ou les modifieurs en début de phrase (notamment des participes passés et présents) :

Original. *Pris de panique*, chacun des trois sots se cache rapidement.
Simplifié. Chacun des trois se cache rapidement.
Exemple 7. Corpus CONTES Les trois sots.

Original. *Se sentant rejetés*, ils étaient toujours ensemble.
Simplifié. Ils étaient toujours ensemble.
Exemple 8. Corpus CONTES Les trois sots.

Les constructions clivées et pseudo-clivées, fréquentes dans les textes documentaires, permettent de porter un focus sur un élément de la phrase. Comme elles impliquent une complexification avec le rajout de la proposition subordonnée, les phrases sont modifiées en une proposition simple avec l'ordre SVO.

Original. *C'est le vent qui* apporte la pluie.

Simplifié. Le vent apporte la pluie.
Exemple 9. Corpus ALECTOR 25_SCI CE1 Vent.

Original. *Ce sont les océans qui* sont les plus sensibles à cette attraction.
Simplifié. Les océans sont les plus sensibles à cette attraction.
Exemple 10. Corpus ALECTOR 89_SCI CE2 Marées.

Par ailleurs, les adjectifs, les adverbes et certaines subordonnées relatives sont supprimés si ils ne sont pas essentiels pour la compréhension du texte (ils apportent des nuances sémantiques mais alourdissent le processus de décodage) :

Original. (…) une fille (…) aux *longs* cheveux *tout* bouclés.
Simplifié. (…) une fille (…) aux cheveux bouclés.
Exemple 11. Corpus ALECTOR 102_ LIT CE1 Émilie et le crayon magique.

Original: *En chemin*, ils aperçoivent, au loin, des bandits *qui se dirigent vers eux*.
Simplifié : Ils aperçoivent au loin des bandits.
Exemple 12. Corpus CONTES Les trois sots.

Enfin, afin d'alléger la lecture et faciliter la compréhension, les textes adaptés évitent les incises et les informations entre parenthèses (c'est le cas pour les textes à destination des enfants sur lesquels nous avons travaillé). Ces structures sont possibles dans d'autres types de documents, comme dans les encyclopédies destinées à des enfants normo-lecteurs : dans ce cas les notions complexes sont explicitées (Cardon, 2018).

4.3 Découpages de phrases

Les découpages représentent une part importante de transformations présentes dans les textes simplifiés. Ainsi, ces découpages interviennent quand il y a des conjonctions, des points virgules ou des énumérations. Les transformations proposées permettent de réduire les phrases à des propositions uniques (une seule forme verbale personnelle) :

Original. Moi, dit la cadette, je n'aurai que ma jupe ordinaire *; m*ais en récompense, je mettrai mon manteau à fleurs d'or.
Simplifié. Moi, dit la cadette, je n'aurai que ma jupe ordinaire*. M*ais en récompense, je mettrai mon manteau à fleurs d'or.
Exemple 13. Corpus CONTES Cendrillon.

Dans les cas de listes et énumérations dans des propositions, les textes adaptés contiennent des phrases courtes, d'où un découpage des phrases dans les versions originales lorsqu'il y a plus d'une proposition. Une conjonction de coordinations finale est généralement ajoutée dans les cas des coordinations ou énumérations.

Original. *De tout temps,* l'homme a utilisé son corps pour mesurer *: la main, l'œil* donnent des informations.
Simplifié. L'homme a toujours utilisé son corps pour mesurer*. La main et l'œil* donnent des informations.
Exemple 14. Corpus ALECTOR 28_SCI Mesure.

Dans cet exemple, tout comme dans l'exemple 3, entre autres, plusieurs transformations interviennent dans une même phrase.

5 Typologie linguistique des variations

Pour conclure, la figure 6 schématise la typologie des transformations syntaxiques d'un point de vue linguistique. Les rectangles arrondis indiquent les modifications à effectuer dans les textes originaux, tandis que les rectangles carrés indiquent les objectifs de simplification. Le classement a été fait en tenant compte des recommandations de simplification pour faciliter la lecture et la compréhension de textes auprès d'enfants en difficulté, issues de nos études de corpus.

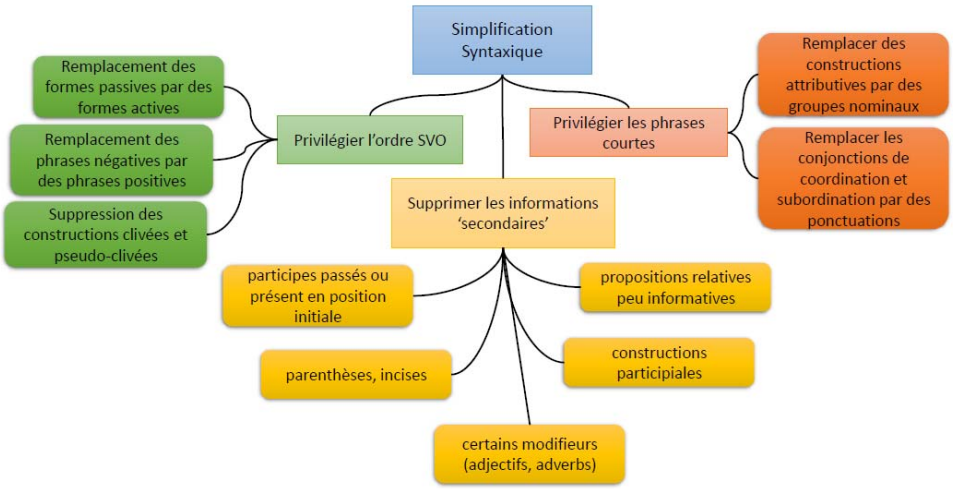

Fig. 6. Typologie des variations syntaxiques en vue d'une simplification de textes pour l'aide à la lecture destinée aux enfants apprentis lecteurs.

En l'état actuel de la technologie, des traitements automatiques permettent la prise en compte des opérations formelles comme les remplacements et les suppressions, afin de garantir l'ordre SVO et raccourcir les phrases. Néanmoins, la suppression d'informations 'secondaires' (hormis les parenthèses et incises, qui -par ailleurs- sont inexistantes dans les corpus adressés à des enfants) demande des traitements plus complexes notamment au niveau sémantique et discursif. Nous poursuivons actuellement nos travaux dans cette direction.

6 Conclusions

Dans cet article, nous avons présenté une typologie de transformations syntaxiques appliquées à des textes originaux (littéraires et scientifiques) destinés à des enfants normo-lecteurs. L'objectif était d'étudier des textes adaptés et de proposer du contenu textuel pour les enfants faibles lecteurs et dyslexiques. Les versions adaptées se caractérisent par le maintien du contenu tout en facilitant l'accès lexical et syntaxique. Cette étude fait partie d'un travail de recherche pluridisciplinaire au sein du projet ANR ALECTOR. Une des issues du projet est la proposition d'un guide de transformations (rendu accessible sur les pages web du projet) afin que des enseignants et autres professionnels de la lecture puissent se l'approprier pour adapter des textes et venir en aide aux enfants en difficulté (proposition de matériel pédagogique ciblé).

Les travaux présentés, *modulo* quelques ajustements destinés à chaque type de public cible, peuvent également être proposés à des apprenants de français langue étrangère ou à des lecteurs francophones avec un handicap cognitif.

Le projet ALECTOR (ANR-16-CE28-0005) est financé par l'Agence Nationale de la Recherche. Nous remercions Ludivine Javourey-Drevet, Aurore Brunel, Mathilde Combes, Marie Nandiegou, Stella Reboul, Stéphane Dufau et Johannes Ziegler pour leurs contributions dans les transformations manuelles des textes, dans les tests de lecture dans cinq écoles du département du Var et dans les premières analyses des effets de la simplification dans l'apprentissage de la lecture (publications à venir).

Références bibliographiques

Alva-Manchego, Bingel, J., Paetzold, G., Scarton, C. et Specia, C. (2017) Learning How to Simplify From Explicit Labeling of Complex-Simplified Text Pairs, *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), 295-305.

Billami, M. B., François, T. et Gala, N. (2018) ReSyf: A French Lexicon with Ranked Synonyms. *Proceedings of the 27th International Conference on Computational Linguistics (COLING-2018*), Santa Fe, New Mexico, USA, 2570-2581

Brouwers, L., Bernhard, D., Ligozat, A. L. et François, T. (2014) Syntactic Sentence Simplification for French. Proceedings of the *3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) at EACL 2014*, Gothenburg, Suède, 47-56.

Brunato D., Dell'Orletta F., Venturi G. et Montemagni S. (2014) Defining an annotation scheme with a view to automatic text simplification. Proceedings of the *First Italian Conference on Computational Linguistics* (CLiC-it 2014), ISBN 978-8-86741-472-7, Pisa, Basili R., Lenci A., and Magnini B. (eds.), published by Pisa University Press srl, Pisa (Italia), 87-92.

Cardon, R. (2018) Approche lexicale de la simplification automatique de textes médicaux. Dans *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, 175-189.

Dekker, R. H. et Middell, G. (2011). Computer-supported collation with CollateX : Managing textual variance in an environment with varying requirements. *Supporting Digital Humanities*, 17–18.

Fenoglio I. et Ganascia J-G. (2007) MEDITE: un logiciel pour l'approche comparative de documents de genèse. *Revue Genesis*, pp. 166-168.

Gala, N., Tack, A., Javourey-Drevet, L., François, T. et Ziegler, J.-C. (2020) Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. *Proceedings of the 12th International Conference on Language Resources and Evaluation* (LREC 2020), poster session. Marseille, France.

Gala, N., François, T., Javourey-Drevet, L. et Ziegler, J.-C. (2018) Vers une simplification automatique de textes pour une meilleure compréhension. Dans *Langue Française « L'apprentissage de la lecture en français langue maternelle et seconde »,* Armand Colin, 123-131.

Gala, N. et Ziegler, J.-C. (2016) Reducing lexical complexity as a tool to increase text accessibility for children with dyslexia. *Proceedings of the workshop Computational Linguistics for Linguistic Complexity (*CL4LC*) at the 26th International Conference on Computational Linguistics (COLING-2016).* Osaka, Japon.

Koptient, A., Cardon, R. et Grabar, N. (2019) Simplification-induced transformations: typology and some characteristics. In Proceedings of the *18th BioNLP Workshop and Shared Task, Association for Computational Linguistics*, Florence, Italy, 309-318.

Namer, F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues*, 41(2), 523–547.

Qi, P., Dozat, T., Zhang, Y. et Manning, C. D. (2018). Universal dependency parsing from scratch. In Proceedings of the *CoNLL 2018 shared task: multilingual parsing from raw text to universal dependencies*. Brussels, Belgium : Association for Computational Linguistics, 160-170.

Rello, L. (2014). DysWebxia: a text accessibility model for people with dyslexia. *Ph.D. thesis*, Universitat Pompeu Fabra, Barcelona.

Saggion, H. (2017) *Automatic Text Simplification (Synthesis Lectures on Human Language Technologies)*. 1 ed. Morgan & Claypool Publishers.

Saggion, H., Gomez-Martin, E., Anula, A., Bourg, L. et Etayo, E. (2011) Text simplification in SImplext : Making texts more accessible. *Procesamiento del Lenguaje Natural (SEPLN)* n° 47, 341-342.

Wilkens, R. et Todirascu, A. (2020) Simplifying Coreference Chains for Dyslexic Children. *Proceedings of the 12th International Conference on Language Resources and Evaluation* (LREC 2020), poster session. Marseille, France.

Zorzi, M., Barbiero, C., Facoetti, A., Lonciari, I., Carrozzi, M., Montico, M. et Ziegler, J. C. (2012). Extra-large letter spacing improves reading in dyslexia. *Proceedings of the National Academy of Sciences*, 109(28), 11455-11459.

Ziegler, J.-C., Perry, C. et Zorzi, M. (2014). Modeling reading development through phonological decoding and self-teaching: Implications for dyslexia. *Philosophical Transactions of the Royal Society B*.

[i] https://alectorsite.wordpress.com/

[ii] https://easy-to-read.eu

[iii] https://www.bdadyslexia.org.uk/advice/employers/creating-a-dyslexia-friendly-workplace/dyslexia-friendly-style-guide

[iv] http://www.lire-et-ecrire.be/latraversee

[v] http://obvil.lip6.fr/medite/

[vi] https://collatex.net

[vii] Publications en cours, plus d'information dans les pages du projet Alector, cf. note i.

[viii] https://cental.uclouvain.be/resyf/