Master 2
Linguistique, Informatique et Technologies du Langage

Rapport du projet collectif sur Vikidia

Encadrant-e-s : *Cécile Fabre, Mai Ho-Dac, Ludovic Tanguy*

Etudiant-e-s :
Mathilde Bacqué, Ariyanidevi Dharma Gita, Silvia Federzoni,
Damien Gouteux, Manon Lalanne et Beatriz Martínez Tornés

2018-2019

Table des matières

1 Introduction

Ce document a pour but de présenter le travail que les étudiant-e-s du master 2 de Linguistique, Informatique et Technologies du Langage, de l'université Jean-Jaurès (Toulouse) ont réalisé pour l'association Vikidia. Le travail réalisé consiste à évaluer la difficulté de la ressource encyclopédique en ligne Vikidia. Pour ce faire, nous avons établi une liste d'indices linguistiques révélant la complexité d'un texte. À partir des résultats de ceux-ci, nous avons comparé les articles de Vikidia entre eux ainsi que la ressource entière avec d'autres types de textes.

Ce rapport est structuré de telle sorte qu'il présente l'organisation générale du projet, puis l'organisation technique (la chaîne de traitement) et les apports théoriques utilisés pour aboutir à la liste des indices retenus. Nous présenterons également les différents types de textes qui nous servent à la comparaison ainsi que les outils et les ressources utilisés dans notre travail. Les résultats de cette comparaison seront présentés et commentés.

1.1 Traitement automatique du langage

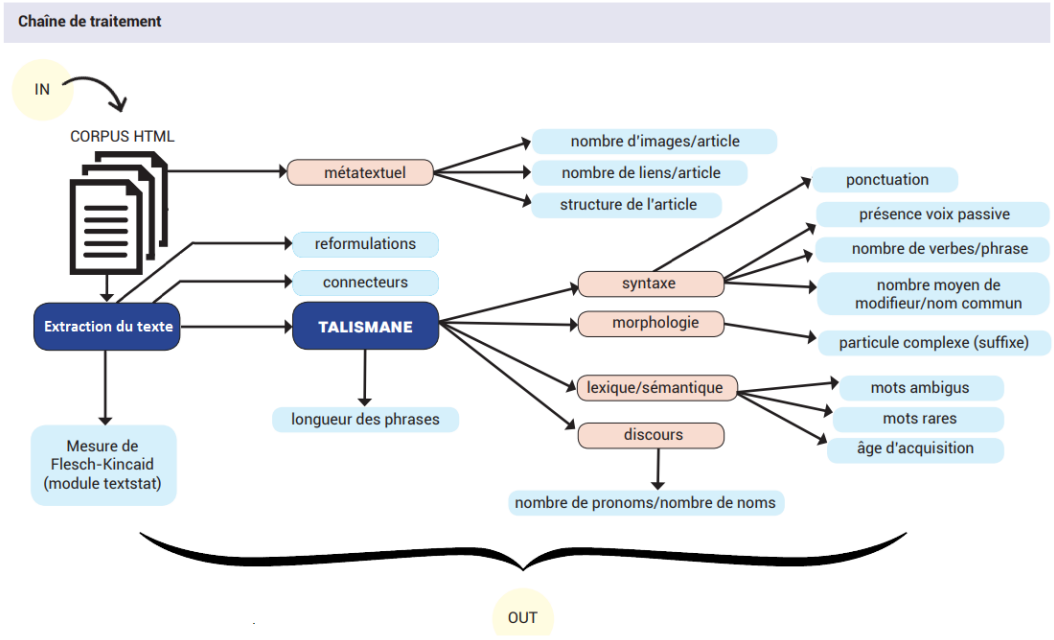
Pour mener à bien ce projet, nous utilisons des méthode de traitement automatique du langage naturel (TAL désormais). Le TAL "s'intéresse aux traitements informatisés mettant en jeu du matériau linguistique : analyse de textes, génération des textes, traduction automatique sont parmi les grands types de traitements (Jacquemin & Zweigenbaum, 2000). "Le TALN est un champ de savoir et de techniques élaborées autour de problématiques diverses. Les concepts et techniques qu'il utilise se trouvent à la croisée de multiples champs disciplinaires : l'IA « traditionnelle », l'informatique théorique, la logique, la linguistique, mais aussi les neuro-sciences, les statistiques, etc" (Yvon, 2007).

Comme le montrent Jacquemin et Zweigenbaum (2000), nous pouvons distinguer différents domaines du TAL :

- **Morphologie** : étude de la formation des mots et de leurs variations de forme ;
- **Syntaxe** : étude de l'agencement des mots et de leurs relations structurelles dans un énoncé ;
- **Sémantique** : étude du sens des énoncés ;
- **Pragmatique** : étude du contexte d'énonciation.

Dans le cadre de ce projet, nous avons mobilisé les différents domaines présenté ci-dessus, à différents niveaux d'analyse. Les différents indices que nous avons utilisés pour mesurer la complexité des articles Vikidia, comme nous le montrons dans la section 3, relèvent des différents domaines, allant de la morphologie à l'analyse du discours.

1.2 Chaîne de traitement



Pour Vikidia, nous partons d'un corpus d'articles dont le texte est compris dans du code HTML. Notre chaîne est également capable de traiter du texte brut.

Certains indices vont travailler sur l'HTML directement pour repérer des informations comme les alinéas ou la mise en gras du texte. Pour les autres, le texte est extrait du HTML et transmis à Talismane, un outil qui va fournir pour chaque mot sa forme canonique, sa catégorie du discours, des informations complémentaires comme le genre ou le nombre pour un nom, le temps pour un verbe, ainsi que les relations de dépendances entre les mots dans la phrase. À partir de ces informations supplémentaires, les autres indices peuvent être calculés. Nous verrons plus en détails nos outils dans le chapitre 4, nous arrêtons dans le chapitre suivant sur les apports théoriques à notre travail.

2
Apports théoriques

Pour mener à bien ce projet d'évaluation des articles de Vikidia, nous avons dans un premier temps établi un état de l'art afin d'obtenir une vue d'ensemble des techniques existantes dans la littérature scientifique. Évaluer la difficulté d'un texte de manière automatique est indissociable de la lisibilité d'un texte. Mesurer la lisibilité d'un texte consiste à évaluer la complexité de lecture et de compréhension d'un texte à partir de variables.

2.1 Les débuts de la lisibilité textuelle : les mesures statistiques

À l'origine, les premières mesures de lisibilité se concentraient essentiellement sur des aspects statistiques du texte. Ces premières approches méthodologiques apparaissent dans les années 20 : Lively et Pressey (1923) ou encore Vogel et Washburne (1928). Viennent ensuite à partir des années 40 des mesures statistiques se basant sur la longueur des mots, des phrases ou le pourcentage de mots dits "difficiles" dans le texte. La mesure la plus connue est celle de Flesch (1948) et repose sur le nombre moyen de syllabes par 100 mots et la longueur moyenne des phrases en mots.

Cette mesure a été ensuite adaptée pour devenir la mesure de Flesch-Kincaid (Kincaid, Fishburne, Rogers, & Chissom, 1975) et est toujours utilisée aujourd'hui. Cette nouvelle mesure calcule un score de lisibilité permettant un classement selon les grades scolaires du système américain. Il se fonde sur le nombre total de mots divisé par le nombre total de phrases additionné au nombre total de syllabes divisé par le nombre total de mots. Nous avons choisi d'utiliser cette mesure dans notre chaîne de traitement car il existe déjà un module effectuant son calcul pour le langage Python, **textstat**, de plus, d'autres études récentes utilisent également cette mesure dans leurs processus. Pour davantage d'informations sur ces mesures statistiques, l'étude de Conquet et Richaudeau (1973) analyse de manière plus précise cinq méthodes.

Ces formules ont été adaptées pour le français à partir de la fin des années 60. Nous retiendrons notamment les travaux de De Landsheere (1963) pour son adaptation du test de Flesch, les travaux de Henry (1975) destinés au traitement automatique de textes et prenant en compte six variables : nombre de mots par phrase, ratio des types de tokens, pourcentage de mots absents d'une liste de référence, la liste Gougenheim, pourcentage de pronoms person-

nels, pourcentage d'indices de dialogue, pourcentage de noms présents dans une liste de mots concrets.

Cependant, ces différentes mesures montrent rapidement des limites. En effet, ces dernières considèrent le texte comme un ensemble de mots ou de phrases. Les relations entre les mots ou les phrases ne sont pas prises en compte. Nous allons détailler dans la section suivante différentes études qui se penchent sur cet aspect et plus précisément sur la notion de cohésion textuelle, complétant celle de la lisibilité.

2.2 Lisibilité et cohésion textuelle

La cohésion textuelle correspond aux relations entre les différents composants d'un texte. Ces relations garantissent une homogénéité du texte. Pour évaluation cette cohésion, beaucoup d'indices ont été observés et étudiés dans la littérature scientifique. Halliday et Hasan (1976) ont par exemple choisi de distinguer cohésion grammaticale et cohésion lexicale. La cohésion grammaticale regroupe des indices tels que la référence, la substitution ou encore les conjonctions. La cohérence lexicale regroupe des indices plus sémantiques comme la présence d'hyponymie. Nous allons nous servir de certains indices de cohésion textuelle pour notre évaluation de la lisibilité. Nous allons essentiellement nous baser sur la présence de pronoms et de références, comme nous allons le décrire dans la section suivante.

« La notion de cohésion textuelle est complémentaire au concept de lisibilité, car elle se rapporte au lexique morphosyntaxique. Au nombre de ces marqueurs de cohésion, on compte l'ordre normatif des mots dans la phrase et le respect des règles d'accord, la distribution (raisonnée) des temps verbaux, les connecteurs, les phénomènes d'anaphorisation, de renominalisation, les marqueurs d'intégration linéaire (temps, espace, progression) » (Maingueneau, 1991), les marqueurs configurationnels (paragraphes, organisateurs métadiscursifs). Ensemble, les marqueurs de cohésion sont vus comme des mécanismes de textualisation, qui « consistent en la création de séries isotopiques qui contribuent à l'établissement de la cohérence thématique. » Nous nous intéressons aux connecteurs et à l'anaphore pronominale.

Pronoms et Référence

La reprise d'un nom par des pronoms constitue une marque de cohésion textuelle. L'utilisation d'un pronom permet de référer à une même entité. Les reprises effectuées créent ce que l'on appelle une chaîne de référence, c'est-à-dire un ensemble d'expressions linguistiques qui se rapportent à un référent commun. Todirascu et al. (2017) étudient les chaînes en relation avec la complexité des textes. Pour mener cette étude, ils utilisent deux corpus s'adressant à différents types de public : enfant et adultes (Vikidia vs Wikipedia), apprenants du FLE à différents niveaux (corpus FLE).

L'idée qui est à la base de cette étude et qui est confirmée sur l'échantillon étudié, est que la reprise à l'identique du référent facilite la compréhension du texte, alors que l'utilisation des pronoms peut rendre le texte plus complexe à lire. En effet, quand on lit un texte, on stocke les informations au fur et à mesure dans la **mémoire de travail**. Ces informations sont ensuite récupérées lorsque c'est nécessaire. Ainsi, dans l'exemple suivant :

- *Le lion est un animal.*
- *Cet animal est un mammifère et il appartient à la famille des félins.*

Nous stockons dans la mémoire de travail en premier le *lion* et *animal*, cette information est exploitée pour comprendre que *cet animal* réfère à *lion* et de même pour *il*. Or, les pronoms ont une portée sémantique et informationnelle faible, c'est pourquoi la présence du référent est indispensable.

Dans le cas où les pronoms se trouvent à une distance brève de ses référents, il n'y a pas de problèmes de compréhension, et retrouver le référent est assez facile. Plus les pronoms sont utilisés loin du référent, moins celui-ci est accessible, ce qui peut causer de problèmes de compréhension, surtout pour des enfants. De plus, dans un passage textuel, plusieurs référents peuvent être instanciés, et la reprise par un pronom pourrait causer des difficultés et créer des ambiguïtés.

Nous avons pris en compte le nombre de pronoms sur le nombre de noms, communs ou propres, pour voir la proportion des pronoms utilisés, mais il serait nécessaire et intéressant d'aller plus loin, en s'intéressant aux chaînes de référence complètes. Si par exemple pour un référent donné on a une alternance de pronoms et redénomination ou reprise à l'identique, la présence de pronoms ne sera pas considérée comme indice de complexité, car le reprise à l'identique ou la redénomination rendent le référent toujours accessible, donc actif dans la mémoire de travail.

Les connecteurs et les marqueurs (ou patrons) de reformulation sont également pris en compte pour mettre en évidence les aspects positifs des articles Vikidia. En effet, si un article présente ces marques, cela peut signifier que le texte est cohérent et cohésif.

Les connecteurs de discours

Nous avons pris la liste disponible sur Lexconn et nous l'avons projetée sur les corpus à analyser. Cette liste contient 328 connecteurs, également appelés marqueurs de discours, regroupés par Roze, Danlos, et Muller (2012) dans le but d'améliorer des applications du domaine du traitement Automatique du Langage. Ces connecteurs sont considérés comme de marques de cohésion textuelle. De ce fait, la présence de ces connecteurs dans les articles Vikidia est un indicateur de cohésion. Il resterait à étudier comment ces connecteurs sont utilisés, c'est-à-dire si leur enchaînement rend le texte cohérent où si leur utilisation crée de problèmes d'incohérence. Nous avons décidé d'intégrer ce type d'indice après la lecture de l'article correspondant au projet Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004).

La reformulation

A partir d'une lecture de quelques articles, nous nous sommes aperçus que les articles Vikidia ne manquaient pas de mots dits complexes. Cependant, ceux-ci sont souvent accompagnés d'explications ou de reformulations, ce qui rend le mot compréhensible et accessible pour l'enfant, en lui permettant de familiariser avec le sens de ce mot et de l'apprendre.

Repérer ces passages, permet de mettre en évidence les aspects positifs qui rendent un texte adapté, compréhensible et enrichissant pour des enfants. Dans le cadre de ce projet, nous nous sommes intéressés à l'extraction et à la mise en évidence de passages qui présentent des patrons de reformulation. Nous appelons patron une structure ou une construction spécifique qui se répète plusieurs fois dans un texte. Pour établir une liste de patrons de reformulation, nous nous sommes inspirés des travaux de Rebeyrolle et Tanguy (2000) qui se sont concentrés sur les énoncés définitoires et qui ont pu fournir des patrons linguistiques, applicables à notre projet.

2.3 Les variables syntaxiques

Lors de nos différentes lectures pour la mise en place des différentes variables d'analyse de la difficulté d'un texte, nous avons pris en compte une catégorie à part entière, regroupant elle-même plusieurs indices. Cette catégorie appartient à la syntaxe.

- **voix passives** : cette variable est souvent prise en compte dans la littérature sur la lisibilité des textes. Une phrase construite sur la base de la voix passive ne respecte pas le schéma classique de la phrase Sujet-Verbe-Objet, comme pourrait s'y attendre le lecteur, ce qui peut engendrer davantage de difficulté de compréhension François, Müller, Degryse, et Fairon (2018). Nous avons sélectionné cette variable d'une part car elle est redondante dans la littérature mais aussi parce qu'elle est facile à mettre en place : en effet la voix passive se repère facilement de manière automatique, notamment grâce à l'étiquetage morpho-syntaxique.
- **constructions verbales** : Collins-Thompson (2014) a sélectionné cette variable afin de repérer les constructions verbales au sein de chaque phrase. L'auteur a pu démontrer l'impact de ces constructions sur la complexité d'un texte en utilisant le coefficient de Pearson : ces constructions présentent un fort coefficient de corrélation avec la difficulté du texte.
- **les modifieurs** : cette variable prend compte les modifieurs de noms afin de mesurer la densité de ces derniers dans un texte. Un modifieur est un élément facultatif qui permet la description de la propriété de la tête de phrase : ici le nom commun ou le nom propre. C'est une des métriques principales pour la syntaxe dans le projet Coh-Metrix (Graesser et al. (2004) et nous avons estimé que cette variable était également pertinente pour nos propres mesures.
- **moyenne de la longueur des phrases** : Nous citerons Richaudeau (1979) pour expliciter l'importance de la longueur des phrases :

> « Il apparaît à tous que les phrases comprenant un nombre important de mots (plusieurs dizaines) sont peu lisibles ; ce qui est exact, et nous y reviendrons plus loin. Et l'on est alors tenté de penser que plus une phrase est courte plus elle est lisible. Or, l'expérience prouve qu'un texte trop segmenté en très courtes phrases est moins bien retenu qu'un texte équivalent écrit en phrases moyennes. »

Cette variable, plus classique en mesure de lisibilité, permet une comparaison de la longueur des phrases entre les articles de Vikidia mais aussi entre d'autres ressources comme Wikipédia.

2.4 Lisibilité et contexte éducatif

Au cours de nos différentes lectures, nous avons observé une problématique liée à la compréhension des textes du point de vue pédagogique. Certains chercheurs se sont penchés sur de nouveaux critères, plus fins et davantage tournés vers des indices linguistiques complexes.

Zeid, Foulonneau, et Atéchian (2012) ont choisi de mêler à la fois des mesures statistiques avec des variables syntaxiques et sémantiques afin d'évaluer des collections de ressources issues d'un environnement éducatif. Cette étude nous a semblé pertinente pour notre travail puisqu'elle prend en compte des paramètres différents pour évaluer la difficulté de lecture d'un texte comme le profil du lecteur ou encore le contexte d'apparition des mots (ce qui diffère des méthodes statistiques). De plus, les auteurs prennent en compte les relations entre les constituants du texte, évaluant ainsi la cohésion de ce dernier comme nous l'avions développé précédemment.

Autre élément caractéristique de la lisibilité que nous avons retenu pour notre chaîne de traitement : l'utilisation de mots rares. Les chercheurs ont choisi d'utiliser une approche basée sur modules créés par Google. Pour notre travail, nous avons choisi de repérer les mots rares grâce à la comparaison avec une liste de référence : le lexique de Manulex de Lété, Sprenger-Charolles, et Colé (2004). Cette liste regroupe les mots et leur fréquence issus de 54 manuels scolaires. Elle présente l'avantage de classer la fréquence des mots selon le niveau scolaire, ainsi nous avons pu viser une tranche d'âge proche de celle des lecteurs de Vikidia.

Pour terminer sur le contexte pédagogique de la mesure de complexité des textes, nous avons abordé la question de l'âge d'acquisition des mots des enfants. C'est un aspect qui influe sur la compréhension du texte, notamment pour des élèves ou des apprenant d'une langue. Nous avons choisi d'intégrer cette variable d'après le projet Coh-Metrix, lui-même inspiré par l'étude de Gilhooly et Logie (1980) et de la création d'une liste de référence pour l'âge d'acquisition des mots en langue anglaise.

2.5 Un indice de surface particulier : la ponctuation

Il existe également dans la littérature scientifique une influence de la ponctuation sur la complexité d'un texte. Nous avons retenu deux études, du même auteur : Timbal-Duclaux (1985, 1986). L'auteur présente une métrique simple pour calculer l'impact de la ponctuation et de la typographie sur la lisibilité d'un texte. Pour cela, il s'est basé sur les travaux de Flesch. Cette métrique prend en compte :

- les majuscules
- les mots en gras, soulignés ou en italique
- les chiffres
- les signes de ponctuation
- des symboles typographiques courants comme &, @, #, ou %
- les alinéas

Chacun de ces éléments valent un point, et sur 100 mots, on obtient un score entre 0 et 35 permettant d'accorder un score de lisibilité.

3

Liste des indices

TYPE	Outils / Ressources	INDICES	CODE INDICE	ARTICLE	EXEMPLES
	annotation Talismane auxiliaire	Proportion de voix passive	pass	Computational Assessment of Text Readability: A Survey of Current and Future Research / Simplification syntaxique de phrases pour le français	*Ces puissances orientales furent bientôt dépassées par celles apparues plus à l'ouest. Au IVe siècle av. J.-C., l'ancienne colonie phénicienne de Carthage constitua un empire qui regroupait la plupart des comptoirs phéniciens de la région.*
Syntaxe	annotation Talismane	Proportion de phrases/syntagmes verbales (verb phrases per sentence)		Computational Assessment of Text Readability: A Survey of Current and Future Research	
		Nombre moyen de modifieurs par nom commun	modParNC	coh-metrix	*Mais, dans les années 1930, la Grande Dépression arriva,ce pays où regnait l'un des plus haut niveaux de vie du monde malgré ses pertes pendant la Première Guerre Mondiale subit la crise de plein fouet, faisant chuter les prix du blé et de la laine. (Vikidia)*
	annotation Talismane	Utilisation de la ponctuation		Timbal-Duclaux, L. (1986). La ponctuation, outil de lisibilité. Communication & Langages, 69(1), 26-38. & Timbal-Duclaux, L. (1985). Textes «inlisable» et lisible. Communication & Langages, 66(1), 13-31.	https://fr.vikidia.org/wiki/Groupe_salafiste_de_la_prédication_et_du_combat
		Mesurer la moyenne de la longueur des phrases		Réutiliser des textes dans un contexte éducatif – évaluation de la difficulté	
Morphologique	demonext, Manulex (pas possible pour Manulex, c'est plutôt orienté association graphème/phonème finalement donc pas très utile)	Particules morphologiques rares ou complexes		Computational Assessment of Text Readability: A Survey of Current and Future Research	*La systole auriculaire, comme son nom l'indique, fait une contraction des oreillettes qui ont été précédemment remplies de sang. Vient alors la systole ventriculaire, consistant en la contraction du muscle cardiaque (myocarde) qui va faire propulser le sang des ventricules dans l'artère aorte ou dans les veines pulmonaires.*
Lexical/Sémantique		Proportion de mots rares/ambigus par rapport à une liste de référence / par rapport au mots du texte (type-token ratio)		Coh Metrix	*Une hémorragie méningée est un épanchement de sang important qui a lieu dans les méninges donc dans le cerveau.*
		Fréquence relative d'un mot par rapport à d' autres mots dans une base de données		Réutiliser des textes dans un contexte éducatif – évaluation de la difficulté	
		âge d'acquisition			
Discours	liste connecteurs + niveau d'analyse paragraphe	Utilisation de connecteurs et autres features de cohesion => coh metrix			*Du coup, Osiris est extrêmement chaude, au point que son atmosphère s'évapore. Les scientifiques imaginent que dans quelques millions d'années, elle pourrait avoir totalement disparu, et il ne resterait **alors que** le cœur de la planète, constitué de roches et de métal : Osiris sera devenu une planète chtonienne.*
		patrons énonces définitoires			*La sédentarité est le fait de vivre toujours au même endroit, **c'est-à-dire** d'habiter dans un endroit précis, et ne pas changer d'endroit tous les jours. Ceci implique un aménagement des lieux, et souvent la sédentarité va avec l'agriculture, l'élevage, et la construction de maisons. La sédentarisation est apparue à l'époque néolithique.*
Lexical/Psycholinguistique		Caractéristiques lexicales comprenant l'âge moyen d'acquisition d'un mot, son caractère concret, imageabilité et son degré de polysémie.		Computational Assessment of Text Readability: A Survey of Current and Future Research /cohmetrix	
Statistique	Direct avec textstat	Mesure de Flesch-Kincaid		Réutiliser des textes dans un contexte éducatif – évaluation de la difficulté	

4

Corpus, outils et ressources lexicales utilisées

4.1 Corpus

Corpus écartés

Lors de la demande initiale, il nous a été demandé de comparer Vikidia avec d'autres ressources similaires. Nous nous sommes alors intéressées aux ressources encyclopédiques adressées aux enfants disponibles sur internet. Une piste donnée était d'utiliser des manuels scolaires et particulièrement Lelivrescolaire.fr, un éditeur indépendant qui élabore des manuels scolaires collaboratifs aux formats papier et numérique. Son contenu ne nous a pas permis de constituer un corpus étant donné que les manuels scolaires sont gratuitement et librement consultables en ligne, sous format pdf mais ne sont pas téléchargeables sans abonnement.

Ce document a pour but de décrire les ressources approchées mais qui n'ont pas pu être intégrées à notre étude.

Encyclopédie Universalis Ressource en ligne mais payante

Incroyable encyclopédie Junior (Larousse) Encyclopédie téléchargeable en ligne mais payante.

Encyclopédie Junior dot com Encyclopédie téléchargeable obtenue dans des circonstances frisant l'illégalité mais sous format pdf inexploitable (même après océrisation car de trop mauvaise qualité).

Wikimini Ressource trop similaire et écrite exclusivement par des enfants.

Manuels scolaires libres Manuel Sésamath seul disponible pour le téléchargement. Ce type de manuel ayant peu de textes, il a été écarté.

Nous avons écarté, pour la comparaison avec des corpus différents, les corpus sur les tweets de la présidentielle de 2017. Ce format très spécifique des tweets semble nous éloigner des phénomènes qu'on cherche à analyser. Ce corpus est disponible sur la plateforme Ortolang, un site internet regroupant des corpus oraux et écrits. Nous nous sommes intéressées, toujours sur le même site, à la ressource TermIT, qui regroupe des textes du domaines des sciences humaines. Seul le lexique est accessible ce qui écarte cette ressource de nos corpus pour la comparaison.

Corpus retenus

Nous avons donc décidé d'effectuer une comparaison de Vikidia avec plusieurs autres corpus de textes de genre et de niveau de difficulté divers. L'objectif de cette comparaison est de pouvoir observer des scores différents selon les indices que nous avons mis en place et de pouvoir attribuer des spécificités en fonction de ces derniers, dans le but de caractériser davantage le niveau d'écriture et de lecture de Vikidia.

ÉMA Corpus d'écrits scolaires de niveau primaire, disponible sur Ortolang

Corpus Monde Diplomatique Corpus d'articles journalistiques issus du Monde Diplomatique de 1998 à 2008

OrthoCorpus Corpus d'articles scientifiques issus de la revue Rééducation Orthophonique, disponible sur Ortolang

Corpus Wikipédia Un corpus que nous avons constitué, contenant une sélection d'articles Wikipédia en français

Corpus Vikibest Les 76 meilleurs articles de Vikidia (à date de mars 2019)

Corpus Vikidia à simplifier Les 17 articles à simplifier de Vikidia (à date de mars 2019)

Corpus Vikidia Un large extrait des articles de Vikidia

Corpus Wikipédia Un large extrait des articles de Wikipédia

Corpus Littéraire 6^{e}-5^{e} Corpus d'œuvres littéraires destinées aux élèves en 6^{e} et en 5^{e}

Corpus Maupassant Corpus d'œuvres littéraires de Maupassant, romans et nouvelles.

Pour mieux étudier Vikidia, nous nous sommes intéressées aux *super articles* de Vikidia et aux *articles à simplifier*, les deux présentant les deux extrémités du spectre de ce qui est lisible sur Vikidia. Nous faisons l'hypothèse qu'un super article correspond, en terme de lisibilité, à ce qui est attendu pour les lecteurs de Vikidia.

4.2 Outils

4.2.1 Talismane

Pour obtenir les informations relatives nécessaires à notre étude, nous avons utilisé un analyseur syntaxique : Talismane [1]. Ce programme a été développé au sein du laboratoire CLLE-ERSS de Toulouse par Assaf Urieli dans le cadre d'une thèse. Il permet d'analyser tous les mots d'une phrase. On obtient alors pour chaque mot sa forme canonique, sa catégorie grammaticale (nom, adjectif, adverbe, etc.), les deux informations formant son *lemme*, et des informations supplémentaires :

Pour les noms : genre et nombre

Pour les verbes : temps et personne

1. http ://redac.univ-tlse2.fr/applications/talismane.html

Ainsi, Talismane est capable pour *« Tu danses bien. »* et *« La danse est ma passion. »* d'obtenir pour le premier *danses* le lemme (danser, verbe) et pour le second (danse, nom). Talismane analyse également les relations de dépendances syntaxiques entre les mots dans une phrase. Talismane se présente comme un serveur auquel est soumis du texte. En réponse, il renvoie du texte avec les informations qu'il a calculées que nous pouvons utiliser avec un autre outil.

4.2.2 Python

Nous avons utilisé le langage de script Python [2], dans sa version 3.6, pour le calcul de nos indices et la transformation de nos données. Un script lit les données retournées par Talismane et les charge en mémoire. Ensuite, différents scripts calculent nos indices, parfois en interrogeant des ressources supplémentaires comme les lexiques GLAWI [3] ou Manulex [4]. Un script rassemble ensuite les résultats produits pour générer une page web au format HTML ou un tableur au format Excel (XLS).

Nous avons développé deux approches complémentaires : dans l'une, nous calculons nos indices par corpus, en produisant une page web par corpus et un tableur avec une ligne par corpus. Dans l'autre, destinée aux encyclopédies Vikidia et Wikipédia, nous calculons nos indices par article, en produisant une page web par article et un tableur avec une ligne par article. La première permet de comparer différents corpus, la seconde de comparer les articles composant un même corpus. Les résultats sont ensuite transmis à un autre outil.

4.2.3 R

Le langage R [5] et son environnement R Studio permet d'analyser de nombreuses variables statistiques pour en obtenir des représentations ou des mesures de corrélations.

Nous avons présenté dans ce chapitre les corpus et les outils que nous avons utilisés. Le prochain chapitre détaille nos résultats.

2. https ://www.python.org/
3. http ://redac.univ-tlse2.fr/lexiques/glawi.html
4. http ://www.manulex.org/
5. https ://www.r-project.org/

5
Résultats des analyses

5.1 La méthode : Analyse en Composantes Principales (ACP)

La chaîne de traitement précédemment explicitée nous a permis d'obtenir des indices calculés sur chaque corpus et chaque article. On obtient alors une base de données comprenant un grand nombre d'individus (les textes analysés) décrits par un grand nombre de variables (les indices calculés). Afin de synthétiser et structurer toutes ces informations, nous avons procédé à une analyse factorielle : l'Analyse en Composantes Principales (ACP).

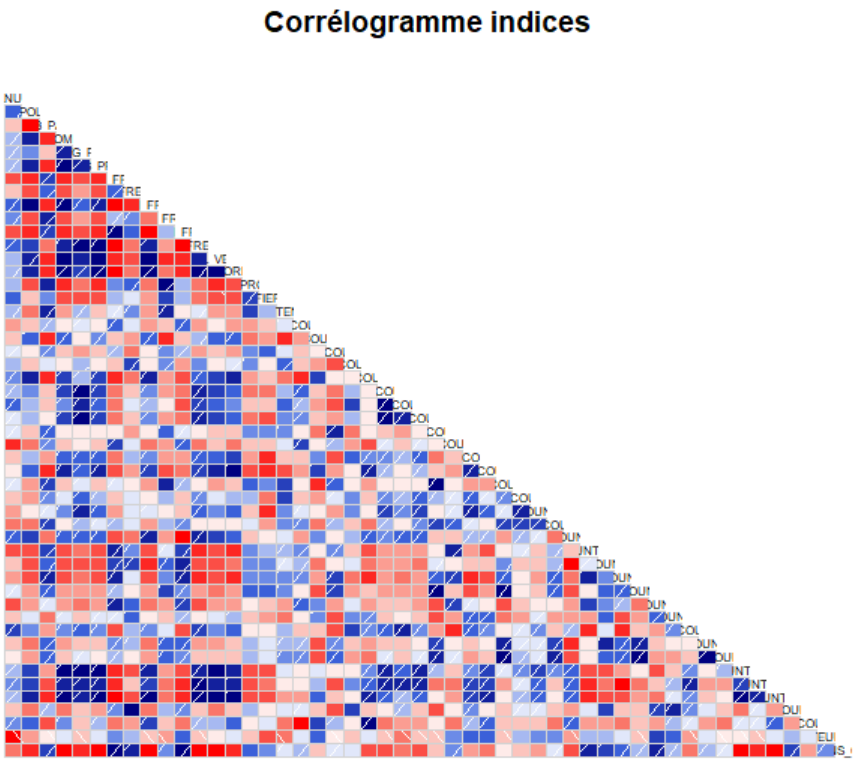
L'ACP sur unités statistiques, ou sur individus, est une méthode statistique utilisée pour mener des investigations au sein d'une population décrite par des variables quantitatives, souvent en grande quantité.

Si une population de n individus statistiques est décrite par deux variables numériques, par exemple une taille et un poids, il est facile de la représenter par un nuage de n points-individus dans un espace vectoriel à deux dimensions et d'analyser graphiquement cette représentation en repérant immédiatement s'il existe un lien entre les deux variables, selon que le nuage a une forme définie ou qu'il ressemble à une tempête de neige. Mais si l'on ajoute d'autres variables quantitatives, par exemple une date de naissance convertie en âge ou une pointure, plus un nombre de dents, plus je ne sais quoi, il devient impossible de distinguer des sous-groupes parmi nos n individus. On pourrait alors standardiser les variables (en enlevant la moyenne puis en divisant par l'écart-type) afin de disposer de critères comparables puis de présenter le nuage de points dans tous les plans construits avec les axes représentatifs des variables. Le problème de cette démarche réside dans la quantité de plans construits, la plupart étant probablement inutiles car ne permettant pas de déceler grand chose. L'ACP permet de visualiser ce nuage de points non pas sur tous les axes représentant les variables standardisées de départ mais, et c'est là son principal intérêt, sur de nouveaux axes (sous-espaces vectoriels).

Graphiquement, ceux-ci sont représentés deux à deux pour que les proximités entre points-individus soient visibles dans des plans mais il y en a beaucoup moins que dans notre idée de départ : ces axes sont calculés afin de représenter au mieux les données initiales. La direction du premier axe épouse le plus fort allongement du nuage de points. En d'autres termes, il est placé de façon à

absorber un maximum d'inertie. Un second axe orthogonal au premier absorbe le maximum d'inertie restante et ainsi de suite. Les axes sont ainsi triés par ordre décroissant d'importance dans leur rôle à épouser la forme du nuage : il est désormais possible de ne choisir qu'un petit nombre d'axes (typiquement les 2 ou 3 premiers) et omettre les axes restants.

Cette méthode permet ainsi de réduire la dimensionnalité des données en les synthétisant selon les premiers axes qui sont les plus explicatifs de la dispersion du nuage de points-individus. Pour une présentation plus approfondie de l'ACP, nous proposons (Levshina, 2015, Chap. 18) ainsi que le site de Jean-Yves Baudot pour une présentation plus accessible [1]. En ce qui concerne notre démarche, cette approche a été choisie afin d'analyser le grand nombre d'individus (articles et corpus) sur le grand nombre d'indices linguistiques calculés. En effet, l'étude des corrélations des indices deux par deux aurait été trop longue et coûteuse à mettre en place en faisant le lien entre les observations sur les indices et les articles ou corpus évalués, comme souligne le corrélogramme des indices calculés sur les corpus présenté-ci dessous :



Corrélogramme des indices calculés sur les différents corpus

Nous présentons tout d'abord les analyses effectuées sur les différents corpus menées afin de caractériser Vikidia dans sa globalité, puis les analyses menées sur ses différents articles.

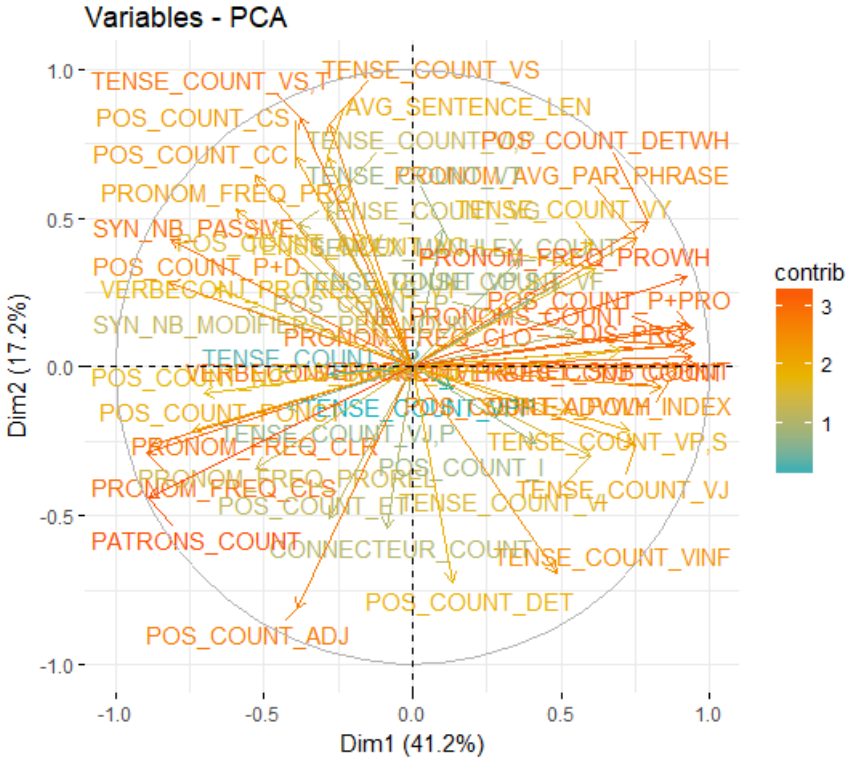
1. http ://www.jybaudot.fr/Analdonnees/acpvar.html

5.2 Analyse comparée de Vikidia avec les autres corpus

5.2.1 ACP sur les variables

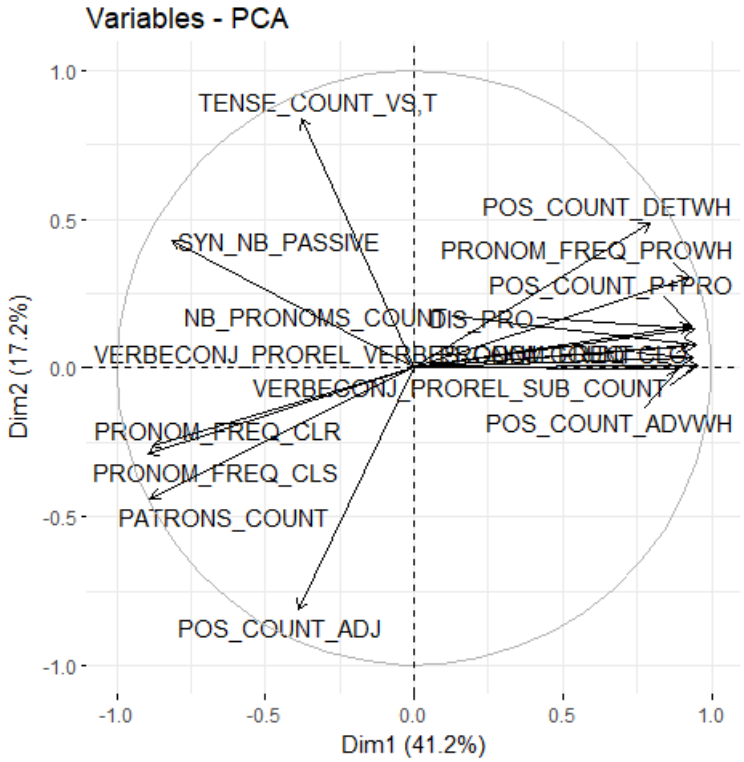
Nous avons commencé par une ACP sur les variables afin d'analyser les liens entre les différents indices calculés (considérés comme des variables statistiques). Pour ce faire, nous avons considéré les nouvelles variables calculées par l'ACP. Ces nouvelles variables, nommées composantes principales, correspondent aux axes construits par la méthode et sont triées de la plus significative à la plus moins significative, ce qui permet de n'en retenir qu'un petit nombre. Toutefois, ces composantes principales ne mesurent pas quelque chose de prédéfini, et c'est un des principes d'une analyse factorielle de leur trouver une signification.

La figure ci-dessous représente le cercle des corrélations des variables considérées sur les deux premiers axes. Nous avons choisi de ne considérer que ces deux axes étant donné qu'ils permettent d'expliquer presque 60% de l'inertie des données. Le légende en couleurs rend compte de la contribution des différentes variables dans la construction de ces deux axes.



Le grand nombre de variables complique la lecture de ce graphique. Il est pourtant possible de remarquer que le premier axe (Dim1) semble assez sensible à la fréquence des pronoms ainsi qu'à leur type. En effet, cet axe est corrélé négativement à un fort ratio de pronoms clitiques réflexifs et sujet et positivement au ratio de pronoms interrogatifs et des P + PRO (prépositions + pronom comme *duquel*...). Afin de faciliter (dans la mesure du possible) la

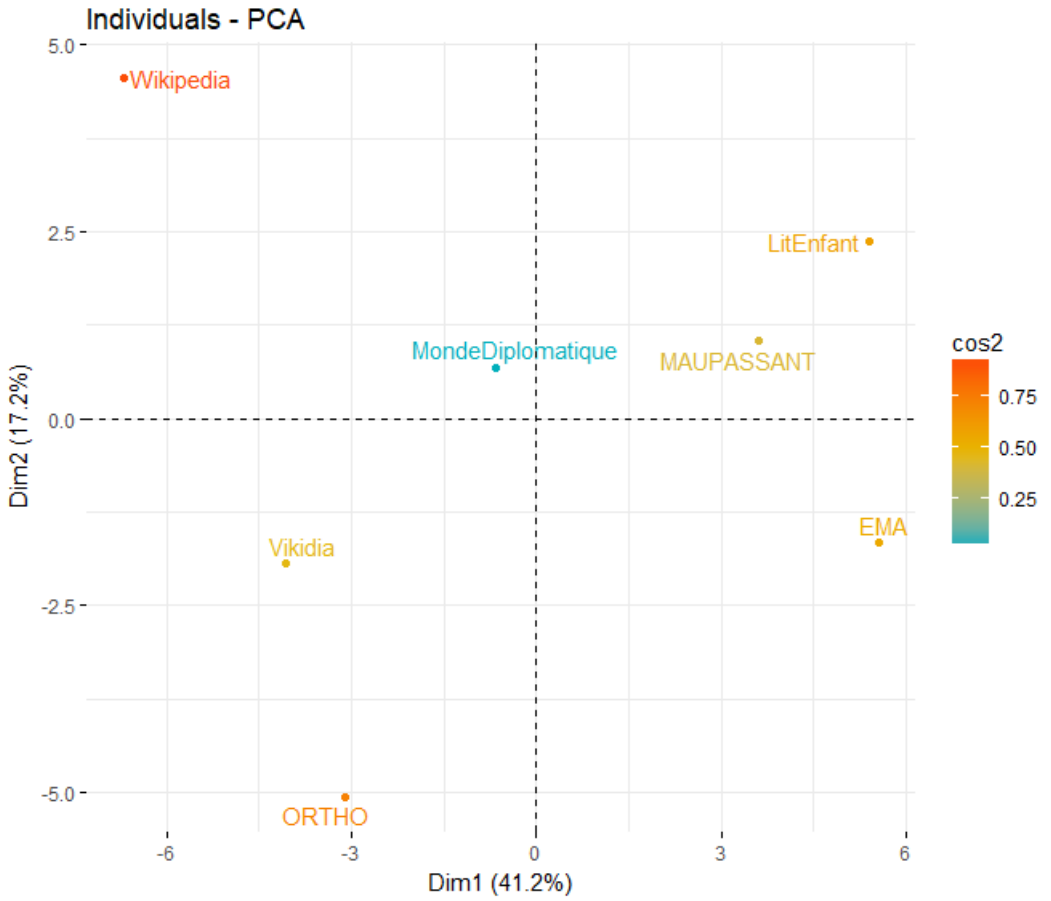
lecture du graphique, nous proposons une figure qui ne représente que les 15 variables ayant contribué le plus à la construction des axes.



On remarque ainsi que le deuxième axe est influencé par le ratio d'adjectifs et de voix passive. On remarque aussi la forte influence de TENSE_COUNT_VS,T qui comme tout les temps verbaux étiquettés par Talismane avec une virgule correspond à un temps verbal ambigu pour cet outil : ils s'agit soit du passé simple soit du subjonctif présent. Nous avons également pu remarquer le lien entre la présence de pronoms clitiques sujet et réflexifs avec la quantité relative de patrons de reformulation repérés dans un corpus.

ACP sur les individus

Nous avons également effectué une ACP sur les individus (ici, les différents corpus). Nous avons ainsi représenté les individus sur les mêmes axes que nous avons essayé d'analyser précédemment. La légende de couleurs correspond cette fois-ci au cosinus carré. Le cosinus carré permet de juger la qualité de la qualité de représentativité d'un point vis-à-vis d'un axe. Géométriquement, tout point de l'espace considéré forme un triangle rectangle avec sa projection sur l'axe et avec l'origine. L'éloignement du point peut donc être mesuré par son cosinus carré. Si la représentation est parfaite, le point se confond avec l'axe, l'angle est nul et le cosinus carré est égal à 1. Les fortes contributions et les cosinus carré proches de 1 sont des éléments susceptibles d'attribuer une signification à un axe. Dans notre cas le cosinus carré le plus élevé correspond au corpus Wikipedia.



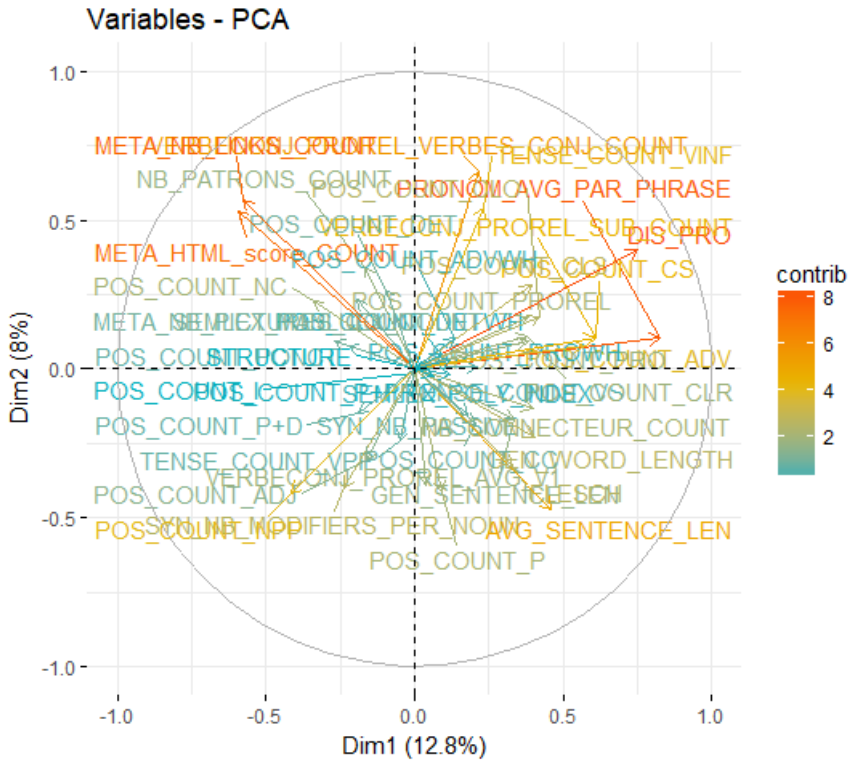
On remarque la distribution des corpus selon le premier axe (Dim1) qui semble distinguer les corpus plus littéraires (MAUPASSANT et LitEnfant) ou narratifs (EMA) des corpus plus techniques ou encyclopédiques. Ces remarques semblent aller en accord avec les variables associées à cette dimension : on s'attend à retrouver plus de tournures interrogatives dans des narrations, et plus de patrons de reformulation dans les textes explicatifs.

Le deuxième axe (Dim2) a attiré notre attention car il semblait distinguer, à première vue, les corpus écrits par des adultes des corpus écrits par des enfants (ou à vocation de vulgarisation comme ORTHO). Ceci pourrait s'expliquer par les variables associées à cette dimension (décompte de voix passives, d'adjectifs et de patrons de reformulation). Même si l'inertie expliquée par cette dimension n'est pas très élevée (17,2%), nous pensons qu'une analyse plus approfondie de ces trois indices pourrait être intéressante pour caractériser avec plus de finesse Vikidia par rapport à ces différents types de corpus. Pourtant, même avec les indices linguistiques assez superficiels que nous avons utilisés pour comparer ces différents corpus, il est assez encourageant de visualiser Vikidia comme étant proche de Wikipedia et de ORTHO en selon la première dimension, mais de conserver une ordonnée comparable au corpus EMA, écrit par des enfants : cette comparaison entre corpus semble souligner la simplicité et le caractère explicatif de Vikidia.

5.3 Analyse des articles de Vikidia

La même méthode d'analyse a été menée avec pour objet d'étude les articles de Vikidia. Dans ce traitement article par article, les indices métatextuels ont pu être pris en compte à partir de la page html de chaque article. Afin de comprendre au mieux la distribution des articles de Vikidia et de pouvoir les caractériser, nous avons pris en compte les articles des sous-corpus Vikibest (76 super articles) et Vikisimply (17 articles "à simplifier"). Nous avons ainsi considéré une variable qualitative supplémentaire dans l'ACP dénotant le type de chaque article : s'agit-il d'un article normal, super ou à simplifier ? Les articles considérés comme "normaux" correspondent à 13% des articles de Vikidia.

Nous avons commencé par effectuer une ACP sur les variables, dont nous présentons le cercle des corrélations avec toutes les variables ci-dessous.



Le code couleur permet de visualiser les variables ayant une contribution plus élevée dans la construction des axes. On remarque ainsi que les variables permettant d'absorber le plus d'inertie des données sont deux indices métatextuels (nombre de liens internes et score HTML) ainsi que la distribution des pronoms (nombre moyen de pronoms par phrase et nombre de pronoms / taille de l'article en nombre de mots). Les deux composantes principales présentées en dimension 1 et 2 de la figure, n'expliquent pourtant que 20,8% de l'inertie, ou de la distribution des indices calculés pour les différents articles. Nous avons ainsi estimé que cet espace n'était pas assez pertinent pour discriminer les articles en matière de complexité, mais qu'il reste de l'ordre de l'intuition de la bonne formation d'un article (présence de liens internes et d'un typographie travaillée).

Nous avons tout de même essayé de représenter les individus dans cet espace afin d'essayer de voir si des zones se dégageaient pour les super articles ou les articles "à simplifier". Le faible nombre de super articles et d'articles à simplifier a rendu cette représentation trop peu fiable pour être analysée. Ces considérations soulignent certaines limites de notre travail, qui reste un travail de défrichage et d'analyse superficielle de la complexité de Vikidia. Nous avons réfléchi à certaines pistes à suivre afin d'approfondir et de continuer les analyses.

6
Pour aller plus loin

Tout d'abord, dans l'optique de rendre possible et plus représentative une analyse article par article de la ressource, il serait intéressant de prendre en compte une évaluation quantitative de la complexité, ou des attentes en termes de bon article (ou exemples de mauvais articles). Ceci pourrait permettre d'approfondir les aspects perçus comme difficiles par les sujets, plutôt que rester contraints aux simples indices linguistiques et leur valeur discriminante dans la caractérisation des articles. Pour ce faire, nous avons envisagé de proposer un questionnaire de compréhension à des enfants, permettant d'évaluer pour différents articles ou extraits d'articles la difficulté par rapport à la tranche d'âge visée.

De plus, il serait envisageable d'ajouter dans la plateforme même de Vikidia un petit questionnaire à la fin de chaque article sous forme d'une échelle d'évaluation que tout lecteur pourrait remplir afin de jauger la complexité perçue de l'article qu'il vient de consulter. Ces données pourraient être utilisées pour compléter les analyses effectuées en fournissant des points de repère sur ce qui est trop complexe. D'autre part, les indices que nous avons utilisés peuvent être améliorés, approfondis et étendus sur de nombreux aspects.

Premièrement, penchons nous sur la cohésion et la cohérence textuelles. S'intéresser à la cohérence signifie s'intéresser aux relations du discours exprimées par les marqueurs que nous avons compté dans les indices. Les études sur les relations de discours de basent sur les schémas proposés par la RST (Mann & Thompson, 1986). Ce cadre descriptif est utilisé pour rendre compte de la structure d'un texte, tout en essayant de faire ressortir la structure hiérarchique qui sous-tend. La description de la structure s'appuie sur la prise en compte des relations sémantiques pouvant exister entre les différentes propositions du texte. Ces relations font souvent l'objet d'une "signalisation" dans les textes, c'est-à-dire qu'elles sont signalées dans le texte à l'aide d'expressions lexicales. Les connecteurs correspondent donc à ces expressions, et peuvent être considérés des "marqueurs de cohérence". La présence de ces marqueurs permet alors d'interpréter plus aisément les relations qui sous-tendent, et le lecteur n'est pas obligé de les inférer.

S'intéresser à ces marqueurs et aux relations de cohérence pour les articles Vikidia, permettrait de faire ressortir les aspects positif, donc mettre en évidence les articles qui présentent une bonne cohérence, ainsi que repérer les articles qui manquent de cet aspect, pour pouvoir les améliorer.

Une autre perspective de développement que nous souhaitions mettre en valeur concerne les indices basés sur les patrons de reformulation. En effet, il serait intéressant de mener des études plus approfondies. Il serait possible, et intéressant, de repérer les termes difficiles (ou spécifiques) et analyser les cas où ceux-ci sont suivis ou non d'une explication, explicitation ou définition.

Les énoncés définitoires ou de reformulation, peuvent être également caractérisés par l'utilisation de termes plus ou moins techniques/spécifiques. Il serait pertinent de procéder à une extraction et une observation qualitative des ces termes pour analyser à quel niveau de généralité ils se situent.

Par exemple, si on utilise des termes généraux (appelés hyperonymes) pour décrire des termes plus spécifiques, on peut avoir différents niveau des généralisation. Pour mieux comprendre, considérons les trois phrases suivantes :

1. Le lion est un félin.
2. Le lion est un mammifère.
3. Le lion est un animal.

Les trois définitions ont un degré de spécificité différent, et "animal" est le terme (l'hyperonyme) le plus générique, ce qui pourrait rendre plus accessible le sens. Ici nous avons donné un exemple assez simple, mais cela vaut pour d'autre termes plus compliqués et qui pourraient s'avérer difficiles d'accès pour des enfants entre 8 et 13 ans.

L'objectif, comme pour les connecteurs, serait alors de faire ressortir les articles qui contiennent des explications, pour en mettre en évidence l'aspect positif, et indiquer les articles qui contiennent des termes "spécifiques" et "compliqués" qui ne sont pas suivi d'explication, ce qui en rendrait la lecture et la compréhension plus difficile.

Concernant la liste de référence Manulex : il serait possible d'envisager la prise en compte de la fréquence des mots dans les manuels. Si le mot rencontré a une fréquence élevée dans Manulex, il ne sera pas pondéré de la même manière qu'un mot ayant une fréquence faible car plus rare et susceptible d'être moins connu par un enfant.

Il serait également envisageable de prendre en compte la complexité morphologique des mots. Pour ce faire, l'idée est de repérer les mots complexes morphologiquement à l'aide d'une liste de suffixes français tirée d'une ressource lexicale telles que Glawi [1]. S'intéresser à cet indice, permet de repérer des mots qui, étant donné leur compositionnalité, pourraient complexifier la tâche de lecture et compréhension, ou qui mériteraient d'être expliqués.

Pour tous les indices que nous venons de présenter, il serait envisageable de les calculer en s'intéressant à une unité d'analyse plus petite : le paragraphe.

1. http://redac.univ-tlse2.fr/lexiques/glawi.html

Cela permettrait d'avoir une vision plus précise au niveaux de la structure interne du texte, rendant possible l'identification des paragraphes "complexes". En effet, il se peut que dans sa globalité l'article est un bon article et qu'il n'y ait que des paragraphes qui posent problèmes.

Enfin, intéressons nous aux variables de l'âge d'acquisition et des mots concrets ou abstraits : nous n'avons pas pu mettre en place ces deux variables car il n'existe pas, à notre connaissance, d'équivalent français d'une liste de mots concrets ou un lexique des mots classés en fonction de l'âge d'acquisition des lecteurs. Sans doute, des travaux sur cette thématique vont émerger et des ressources seront mises à disposition pour de futures études.

7
Discussion

En nous intéressant à la complexité des textes, nous nous sommes également questionné sur leur accessibilité pour des enfants et les capacités de rédaction de ceux-ci. Bien sûr, aucun article ne se targue de décrire précisément ce que comprennent et produisent des enfants mais nous avons relevés quelques éléments pertinents qui nous permettent de prendre du recul sur notre travail.

À 7 ans, l'enfant a acquis l'essentiel de ses capacités linguistiques. Son lexique (l'ensemble des mots qu'il peut mobiliser) comporte plusieurs milliers de mots et les structures grammaticales les plus courantes sont maîtrisées. Les capacités discursives (savoir organiser et construire son discours) et pragmatiques (utiliser le langage pour un but précis) quant à elles, sont opérationnelles mais vont connaître un développement significatif dans les année de scolarité primaire et secondaire (Hickmann 2000, cité par Schelstraete, Zesiger, et Bragard (2006)).

De manière générale, la littérature scientifique s'intéresse particulièrement aux troubles du langage ou au jeunes de moins de 10 ans. Pour expliquer le manque de recherches sur une population adolescente, la chercheuse Claire Placial explique que l'écriture peut être considéré comme culpabilisante. En effet, plus les jeunes grandissent, plus les disparités socio-culturelles se creusent et se remarquent dans les écrits (la maîtrise du français n'est pas homogène). Le cadre scolaire à tendance à se concentrer sur la forme plus que sur le contenu, ce qui ne permet pas de valoriser les écrits des adolescents qui ont plus de difficultés avec le français écrit standard. Cette observation est aussi faite par Gromer (1996) qui précise que la logique interne des productions d'élèves du CM2 à la 5^{eme} est cohérente même si elle ne correspond pas à l'attendu (aux consignes du professeur).

En ce sens, Vikidia fait partie des projets qui permettent aux enfants et jeunes adolescents de s'approprier l'écriture. Il parait alors contre-productif d'imaginer un système qui corrigera en rouge les tournures considérées comme excentriques ou les fautes d'orthographes, à l'instar de l'école.

> Comme dans le domaine de l'oral, l'activité langagière des élèves dépasse largement ce qui fait l'objet d'un enseignement explicite"
> Bautier et Bucheton (1995)

A

Notes techniques du script

Notre code est disponible librement sur la plate-forme Github à l'adresse : https ://github.com/m2litl2019/Projet-conception-Vikidia
Nous présentons dans cette annexe les principaux fichiers qui le composent.

- **presentation.py** : Un fichier qui offre un objet Presentation qui permet de représenter un résultat de nos analyses. Une présentation peut concentrer les résultats de plusieurs analyses, chacune ayant un numéro différent (0, 1, 2...). Les résultats sont des dictionnaires, des couples indice - valeur. L'objet peut ensuite fournir une représentation HTML à partir d'un template ou XLS (tableur Excel).
- **reperage_def_con.py**
  reperage_passive.py
  reperage_pronoms.py
  reperage_tpsV.py
  reperage_verbeconj_prorel_sub.py
  repere_connecteurs.py
  indices_html.py
  lexique.py : Différents fichiers pour calculer des indices. Ils fonctionnent tous de la même manière : ils prennent en entrée un nom de fichier ou de répertoire et fournissent un dictionnaire de résultats, sous la forme indice-valeur. Certains indices fonctionnent à partir de :
  - des fichiers textes talismanés (.tal ou .txt, exemples dans vikibest_tal et ema.tal) :
    - reperage_def_con.py
    - reperage_passive.py
    - reperage_pronoms.py
    - reperage_tpsV.py
    - reperage_verbeconj_prorel_sub.py
    - lexique.py
  - des fichiers html bruts (.html ou .txt, exemples dans vikibest_html) :
    - indices_html (via une requête)
  - des fichiers html simplifiés (.txt, exemples dans vikibest_div) :
    - avec uniquement des balises h2, h3, h4, h5 et p. C'était dans l'optique de garder la structure du texte. Nous n'avons pas exploré cette voie par manque de temps.

  - des fichiers html simplifiés talismanés (.txt, pas d'exemples). C'était dans l'optique de garder la structure du texte tout en ayant les informations de Talismane. Nous n'avons pas exploré cette voie par manque de temps.

  Certains indices ont besoin de ressources externes dans le répertoire **resources** :

  - lexique.py
    * manulex.csv
    * GLAWI.txt

- **multitests.py :** Calcule les indices sur plusieurs corpus et stocke les résultats dans du HTML (1 page par corpus) et du XLS (1 ligne par corpus).
- **multitests-articles.py :** Calcule les indices pour chaque élément d'un corpus et stocke les résultats dans du HTML (1 page par élément) et du XLS (1 ligne par élément).
- **texteval.py** : Bibliothèque de base qui peut lire :
  - un fichier texte talismané ⇒ 1 part
  - un répertoire de fichiers textes talismanés => 1 part ou une liste de parts
  - un répertoire de fichiers html simplifiés talismanés ⇒ 1 multifile avec autant de multipart qu'il y a de fichiers et les stocke dans un modèle de données qui compte 5 niveaux :
    1. **Multifile :** un répertoire de fichiers
    2. **Multipart :** un fichier avec plusieurs parties
    3. **Part :** une partie
    4. **Sentence :** une phrase
    5. **Word :** un mot

  Une partie est assimilable à un paragraphe. Nos fichiers textes talismanés n'avaient pas conservé la structure en paragraphe nous n'utilisons donc pas Mutlifile et Multipart. Un corpus est assimilé à une seule Part dans la comparaison entre corpus et un article est assimilé à une seule Part dans la comparaison entre articles.

- **texteval.py** peut également sérialiser (stocker sur le disque dur) sa représentation en mémoire dans un fichier binaire avec l'extension .bin.

- **ema.tal :** Un fichier texte talismané petit pour expérimenter rapidement.

- **ema.bin : texteval.py** est capable de stocker sur le disque sur une représentation en objets d'un fichier tal pour un chargement plus rapide. Il y a donc une équivalence entre **ema.tal** et **ema.bin**.

- **vikibest_filter.py :** Un script pour transformer du HTML en HTML simplifié. Ce script a été fait pour **vikibest_html** mais produit des résultats ayant des défauts (2 fichiers vides, chat des sables et fermes des animaux, non gestion des tags h5, non suppression automatique des tags img). Ce fichier n'est pas utilisé pour nos productions.

- **server.py :** Un serveur Python pour notre outil de démonstration pour évaluer les indices sur un article.

- **index.py :** Notre outil de démonstration pour évaluer les indices sur un article. Il nécessite un serveur Python (fichier server.py), un interpréteur Python pour lancer le serveur et un serveur Talismane s'exécutant sur la même machine sur le port 7272.

- **templates (répertoire) :** Répertoire contenant des maquettes HTML pour presentation.py. Une maquette est un fichier HTML avec des clés de la forme ___KEY___. presentation.py se chargera de remplacer la clé par sa valeur.

- **results (répertoire) :** Répertoire contenant les résultats de multitests.py.

- **vikibest (répertoire) :** Fichiers textes des 76 meilleurs articles.

- **vikibest_tal (répertoire) :** Fichiers textes talismanés des 76 meilleurs articles.

- **vikibest_html (répertoire) :** Fichiers HTML des 76 meilleurs articles.

- **vikibest_div (répertoire) :** Fichiers HTML simplifiés, non utilisés.

Obtention des différents formats :

1. Pour obtenir les corpus vikibest et vikibest_html, il faut utiliser : spikes/scrappingWikiViki.py.La méthode process_target se charge de rapatrier le contenu d'une page. Si son paramètre strip est à True, on obtient le texte comme le contenu de vikibest sinon on obtient l'html complet comme le contenu de vikibest_html. Le script fonctionne avec une liste d'URLs à aller chercher. Il y en a deux :
   - targets_all.txt
   - targets_best.txt

   Une variable restart permet de redémarrer à une url précise. Il faut la mettre à None si on faire toutes les URLs.
2. Pour obtenir vikibest_tal, il faut passer Talismane sur vikibest
3. Pour obtenir vikibest_div, il faut passer le script base/vikibest_filter.py sur l'html complet.

Références

Bautier, E., & Bucheton, D. (1995). L'écriture : qu'est-ce qui s' enseigne, qu'est-ce qui s' apprend, qu'est-ce qui est déjà là. *Le français aujourd'hui*(111), 26–35.

Collins-Thompson, K. (2014). Computational assessment of text readability. *Recent Advances in Automatic Readability Assessment and Text Simplification*, *165*(2), 97–135.

Conquet, A., & Richaudeau, F. (1973). Cinq méthodes de mesure de la lisibilité. *Communication Et Langages*, *17*, 5-16. doi: 10.3406/colan.1973.3978

De Landsheere, G. (1963). Pour une application des tests de lisibilité de Flesch a la langue française. *Travail Humain*, *26*(1-2), 141-154.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221.

François, T., Müller, A., Degryse, B., & Fairon, C. (2018). Amesure, une plateforme web pour soutenir la rédaction simple de textes administratifs. *Repères-Dorif*, *15*, 1.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*(4), 395–427.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Cohmetrix : Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202.

Gromer, B. (1996). Le texte de l'enfant et l'écrit littéraire. *Repères. Recherches en didactique du français langue maternelle*, *13*(1), 147–163.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. Longman.

Henry, G. (1975). *Comment mesurer la lisibilité*. Paris : Nathan.

Jacquemin, C., & Zweigenbaum, P. (2000). Traitement automatique des langues pour l'accès au contenu des documents. In J. Le Maître, J. Charlet, & C. Garbay (Eds.), *Le document multimédia en sciences du traitement de l'information* (p. 71-110). Toulouse : Cépaduès Éditions.

Kincaid, J., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas : (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Chief of Naval Technical Training, Naval Air Station Memphis.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). Manulex : A grade-level lexical database from french elementary school readers. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 156–166.

Levshina, N. (2015). *How to do linguistics with R : Data exploration and statistical analysis*. John Benjamins Publishing Company.

Lively, B., & Pressey, S. (1923). *A method for measuring the "vocabulary burden" of textbooks*.

Maingueneau, D. (1991). *L'analyse du discours, introduction aux lectures de l'archive*. Paris : Hachette Supérieur.

Mann, W. C., & Thompson, S. A. (1986). *Rhetorical structure theory : Description and construction of text structures*. Marina del Rey : Information Sciences Institute.

Rebeyrolle, J., & Tanguy, L. (2000). Repérage automatique de structures lin-

guistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, *25*, 153-174.

Richaudeau, F. (1979). Une nouvelle formule de lisibilité. *Communication et langages*, *44*(1), 5–26.

Roze, C., Danlos, L., & Muller, P. (2012). LEXCONN : A french lexicon of discourse connectives. *Discours*(10). doi: 10.4000/discours.8645

Schelstraete, M.-A., Zesiger, P., & Bragard, A. (2006). Bilan de la lecture chez l'enfant et l'adolescent. *Les bilans de langage et de voix*, 139–162.

Timbal-Duclaux, L. (1985). Textes « inlisable » et lisible. *Communication et langages*, *66*(1), 13–31.

Timbal-Duclaux, L. (1986). La ponctuation, outil de lisibilité. *Communication et langages*, *69*(1), 26–38.

Todirascu, A., François, T., Bernhard, D., Gala, N., Ligozat, A.-L., & Khobzi, R. (2017, septembre). Chaînes de référence et lisibilité des textes : Le projet ALLuSIF. *Langue française*, *195*(3), 35-52. Consulté sur `https://halshs.archives-ouvertes.fr/halshs-01665316`

Vogel, M., & Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *Elementary School Journal - ELEM SCH J*, *28*. doi: 10.1086/456072

Yvon, F. (2007). Une petite introduction au traitement automatique des langues naturelles.

Zeid, E. A., Foulonneau, M., & Atéchian, T. (2012). Réutiliser des textes dans un contexte éducatif : L'évaluation de la difficulté. *Document numérique*, *15*(3), 119–142.