UCLouvain
Faculté de philosophie, arts et lettres – FIAL

Faculté de philosophie, arts et lettres

Les apports du Traitement Automatique du Langage (TAL) à l'évaluation des facteurs de lisibilité du néerlandais en tant que langue étrangère

Auteur : Victorine Colin
Promoteur(s) : Liesbeth Degand
Année académique : 2024-2025
Master en linguistique à finalité
Traitement Automatique du Langage

Déclaration sur l'honneur, mémoires/TFE de la Faculté de philosophie, arts et lettres, UCLouvain

Je déclare qu'il s'agit d'un travail original et personnel, que toutes les sources référencées ont été indiquées dans leur totalité et ce, quelle que soit leur provenance, et que l'utilisation des outils d'intelligence artificielle est conforme aux consignes générales d'utilisation publiées par l'UCLouvain et disponible à cette adresse : https://cdn.uclouvain.be/groups/cms-editors-p1/portail/reglement/Consignes-ChatGPTversion-etudiante.pdf?itok=SebXXm87

Je suis conscient·e que le fait de ne pas citer une source, de ne pas la citer clairement et complètement, de ne pas annoncer dans une partie spécifique en début du mémoire l'utilisation que j'ai faite des outils d'intelligence artificielle ou de les utiliser de manière contraire à ce qui est prévu dans la note institutionnelle constituent des irrégularités graves au regard du Règlement des études et des examens de l'Université catholique de Louvain (Chapitre 4, section 7, article 107 à 114) au sein de l'Université. J'ai notamment pris connaissance des risques de sanctions académiques et disciplinaires encourues.

Au vu de ce qui précède, je déclare sur l'honneur ne pas avoir commis de plagiat ou toute autre forme de fraude et de n'avoir pas utilisé les outils d'intelligences artificielles en dehors de ce qui est accepté à l'UCLouvain.

Nom, Prénom : Colin, Victorine

Date : 12/08/2025

Signature de l'étudiant·e :

Abstract

This thesis explores the contributions of Natural Language Processing (NLP) to the assessment of the readability of Dutch as a foreign language. Given the growing importance of language learning and the status of Dutch as an official language of the European Union, which is particularly relevant in the Belgian context, this study aims to fill a gap in research. The main objective is to identify the most relevant linguistic variables (lexical and syntactic) for predicting the difficulty of a text for French-speaking learners.

The methodology is based on a quantitative and empirical approach. A corpus of 178 texts, classified according to CEFR levels, was compiled from textbooks and authentic sources. NLP techniques were used to extract 18 linguistic variables, the relevance of which was assessed using statistical analyses (ANOVA, multiple linear regression).

The results show that lexical variables, in particular the proportion of words not found in reference lists and concreteness, are robust predictors of difficulty. In terms of syntax, the proportion of sentence-final verb (SOV) structures is also significant.

The multiple linear regression model developed explains approximately 65% of the variance in text difficulty, confirming the validity of a multidimensional approach. This research constitutes the first systematic modeling of the readability of Dutch as a foreign language and lays the foundations for the future development of objective and appropriate assessment tools.

Résumé

Ce mémoire explore les apports du Traitement Automatique du Langage (TAL) à l'évaluation de la lisibilité du néerlandais en tant que langue étrangère. Face à l'importance croissante de l'apprentissage des langues et au statut du néerlandais comme langue officielle de l'Union Européenne, particulièrement pertinente dans le contexte belge, cette étude vise à combler une lacune dans la recherche. L'objectif principal est d'identifier les variables linguistiques (lexicales et syntaxiques) les plus pertinentes pour prédire la difficulté d'un texte pour des apprenants francophones.

La méthodologie repose sur une approche quantitative et empirique. Un corpus de 178 textes, classés selon les niveaux du CECR, a été constitué à partir de manuels et de sources authentiques. Des techniques de TAL ont été utilisées pour extraire 18 variables linguistiques, dont la pertinence a été évaluée par des analyses statistiques (ANOVA, régression linéaire multiple).

Les résultats montrent que les variables lexicales, notamment la proportion de mots absents de listes de référence ainsi que la concrétude, sont des prédicteurs robustes de la difficulté. Sur le plan syntaxique, la proportion de structures à verbe final (SOV) se révèle également significative.

Le modèle de régression linéaire multiple développé explique environ 65% de la variance dans la difficulté des textes, confirmant la validité d'une approche multidimensionnelle. Cette recherche constitue une première modélisation systématique de la lisibilité du néerlandais langue étrangère et jette les bases pour le développement futur d'outils d'évaluation objectifs et adaptés.

Remerciements

Je tiens à exprimer ma gratitude et mes remerciements les plus sincères à Madame Degand, promotrice de ce mémoire, pour son accompagnement attentif, sa disponibilité, ses conseils et sa confiance. Sa capacité à allier exigence et bienveillance a été déterminante dans la conduite de ce travail. Je remercie également mes parents et ma marraine, dont les relectures et les avis ont contribué à améliorer ce mémoire, ainsi que pour le soutien constant et patient tout au long du parcours. Ma reconnaissance va à Justine, qui préparait elle aussi un mémoire en traitement automatique du langage (TAL) : dans un master où nous sommes peu nombreux, notre soutien mutuel a vraiment aidé à maintenir le cap. Enfin, merci à mes amies Zoé et Camille, avec qui tout a commencé à l'UCLouvain ; bien que nous ayons emprunté des voies différentes, c'est avec elles que je termine à présent mon parcours universitaire. Merci pour les blocus groupés, le soutien émotionnel et les sessions de mémoire intensives.

Utilisation de l'intelligence artificielle comme outil d'assistance

Avant de rentrer dans le vif du sujet, nous aimerions signaler que l'intelligence artificielle a été utilisée dans le cadre de la réalisation de ce mémoire. *Mistral* et *Perplexity* ont été utilisé comme outil d'assistance méthodologique et rédactionnelle, rédactionnelle, tout en préservant l'intégrité académique et l'originalité de la recherche. Les outils d'IA ont principalement servi à reformuler certaines phrases et paragraphes afin d'améliorer la clarté et la fluidité de l'expression écrite, les idées originales et le contenu demeurant entièrement de ma conception.

L'intelligence artificielle a également été mobilisée pour faciliter la recherche de sources pertinentes et l'identification de références bibliographiques en lien avec la problématique étudiée. Toutefois, l'ensemble des sources ainsi identifiées a fait l'objet d'une vérification minutieuse de ma partt, incluant la validation de leur pertinence de leur fiabilité et de leur adéquation avec les objectifs de recherche.

Par ailleurs, l'IA a apporté une assistance technique pour l'optimisation des codes Python utilisés dans l'analyse de données, ainsi que pour la correction orthographique et grammaticale.

Il convient de souligner que cette utilisation de l'intelligence artificielle s'inscrit dans une démarche de transparence méthodologique et que l'ensemble des contenus, analyses et conclusions présentés dans ce travail résultent d'une réflexion et d'un travail de recherche personnels.

Table des matières

Introduction générale

Dans un monde où l'apprentissage des langues étrangères occupe une place de plus en plus importante, autant dans les parcours académiques que professionnels, la question de l'accessibilité des textes pour les apprenants se pose. Comment déterminer si un texte est adapté au niveau des apprenants ? Comment mesurer objectivement la difficulté d'un texte en langue étrangère ? Ces questions constituent le cœur de la problématique de la lisibilité des textes. Depuis plus d'un siècle, ce domaine tente d'apporter des réponses scientifiques à ces interrogations fondamentales.

La lisibilité, définie comme "l'ensemble des caractéristiques d'un texte qui affectent la réussite qu'un groupe de lecteurs peut avoir avec ce texte" (Dale et Chall, 1949) (notre traduction), représente un enjeu majeur pour l'enseignement des langues étrangères.

En effet, proposer des textes dont le niveau de difficulté correspond aux compétences des apprenants constitue un facteur déterminant pour favoriser leur progression. Un texte trop simple risque d'engendrer de l'ennui et du désintérêt, et un texte trop complexe pourrait provoquer découragement et abandon (Krashen, 1985). L'évaluation objective de la lisibilité des textes apparaît donc comme un outil précieux pour les enseignants, les concepteurs de matériel pédagogique, mais aussi pour les apprenants eux-mêmes.

Ce travail de fin d'études s'inscrit dans cette perspective en se concentrant sur le néerlandais en tant que langue étrangère. Si de nombreuses études ont été menées sur la lisibilité de l'anglais, et plus récemment du français langue étrangère, notamment grâce aux travaux de François (2011), force est de constater que le néerlandais demeure relativement peu exploré dans ce domaine. Pourtant, avec plus de 24 millions de locuteurs natifs aux Pays-Bas et en Belgique, et son statut de langue officielle de l'Union européenne, le néerlandais représente une langue d'importance sur la scène internationale, particulièrement dans le contexte belge où il constitue l'une des trois langues nationales.

L'objectif principal de ce mémoire est de faire avancer la recherche sur la lisibilité en néerlandais langue étrangère en explorant et testant le potentiel discriminatoire de différentes variables linguistiques. Cette étude vise à identifier quelles variables sont les plus pertinentes pour évaluer le niveau de difficulté d'un texte destiné aux apprenants de cette langue, contribuant ainsi au développement futur d'outils d'évaluation de la complexité textuelle. L'ambition est d'approfondir notre compréhension des facteurs linguistiques qui influencent la compréhension en néerlandais L2,

offrant de cette manière des perspectives précieuses tant pour la recherche que pour la pratique pédagogique.

Ce travail s'appuie principalement sur deux fondements méthodologiques solides. D'une part, la thèse de François (2011), qui a mené en 2011 une recherche doctorale approfondie sur la lisibilité en FLE (Français Langue Étrangère), nous fournit un cadre méthodologique rigoureux pour l'identification et l'analyse des variables linguistiques pertinentes. D'autre part, les travaux de Kleijn (2018), qui a étudié la lisibilité du néerlandais dans un contexte de langue maternelle, constituent une base empirique précieuse pour adapter ces approches au contexte spécifique du néerlandais. En combinant l'expertise méthodologique de François dans le domaine de la L2 et les connaissances spécifiques de Kleijn sur le néerlandais, notre étude se positionne à l'intersection de ces deux domaines de recherche pour explorer un territoire encore peu défriché, celui de la lisibilité du néerlandais langue étrangère.

L'originalité de notre démarche réside dans l'utilisation des techniques du Traitement Automatique du Langage (TAL) pour analyser et mesurer ces différentes variables linguistiques dans le contexte spécifique du néerlandais L2. Le TAL, en permettant l'automatisation de l'analyse de phénomènes linguistiques complexes tels que la syntaxe, la morphologie ou la sémantique, offre des perspectives prometteuses pour l'évaluation objective de la difficulté des textes. Cette approche computationnelle permet de dépasser les limites des formules traditionnelles, souvent critiquées pour leur simplicité excessive et leur manque de fondement théorique (Bailin et Grafstein, 2001).

La méthodologie adoptée dans ce mémoire repose sur une approche empirique et quantitative. Elle comprend plusieurs étapes clés, à savoir :

1. L'identification des variables linguistiques potentiellement pertinentes pour évaluer la lisibilité du néerlandais langue étrangère ;
2. L'élaboration et l'assemblage d'un corpus de textes classés en différents niveaux de langue ;
3. L'extraction automatique des variables linguistiques sélectionnées à partir du corpus ;
4. L'analyse statistique des données extraites, comprenant dans un premier temps des analyses descriptives, suivies d'une analyse de la variance (ANOVA) afin d'identifier les variables présentant des différences significatives entre les niveaux, et enfin une régression linéaire multiple visant à évaluer le potentiel discriminant de ces variables dans la classification des niveaux de langue.

Ce mémoire s'organise en trois parties principales. La première partie établit les bases théoriques nécessaires à notre étude en explorant le processus de lecture en langue étrangère et en présentant un état de l'art détaillé sur la lisibilité. La deuxième expose la méthodologie adoptée, en détaillant les variables linguistiques retenues et la création du corpus d'entraînement utilisé. Enfin, la troisième partie présente les résultats de nos expérimentations et propose une analyse critique de ceux-ci.

À travers cette recherche, nous visons à faire progresser la compréhension de la lisibilité en langue étrangère, en particulier pour le néerlandais. Si ce travail ne

débouche pas directement sur un outil opérationnel pour les enseignants ou les apprenants, il jette les bases d'une modélisation objective de la lisibilité, susceptible de nourrir, à terme, le développement d'instruments d'aide à la sélection de textes adaptés aux différents niveaux de compétence. Plus largement, cette étude s'inscrit dans une réflexion sur l'apport des technologies numériques à l'enseignement et à l'apprentissage des langues.

Première partie

Les bases théoriques

Chapitre 1

La lecture en néerlandais langue étrangère

1.1 Introduction

La lecture constitue une compétence fondamentale dans l'apprentissage d'une langue étrangère. Elle représente non seulement un objectif pédagogique en soi, mais aussi un moyen privilégié d'accès à la langue et à la culture cible. Dans le contexte spécifique de l'apprentissage du néerlandais comme langue étrangère, cela revêt une importance particulière en Belgique, ce pays étant également néerlandophone.

Ce premier chapitre vise à établir les fondements théoriques nécessaires à notre étude sur la lisibilité du néerlandais langue étrangère. Pour ce faire, nous explorerons d'abord les processus cognitifs impliqués dans la lecture, tant en langue première qu'en langue seconde en mettant en évidence leurs similitudes et leurs différences. Nous nous intéresserons ensuite aux spécificités de la lecture en néerlandais, en tenant compte des particularités linguistiques de cette langue qui peuvent influencer le processus de compréhension. Enfin, nous examinerons la place accordée à la lecture et au texte écrit dans les différentes approches didactiques en néerlandais langue étrangère.

Cette exploration théorique nous permettra de mieux comprendre les mécanismes sous-jacents à la compréhension écrite de textes en néerlandais langue étrangère et, par conséquent, d'identifier les facteurs susceptibles d'influencer la lisibilité de ces textes. Ce cadre conceptuel offre ainsi les fondements nécessaires à la conduite rigoureuse de notre étude.

1.2 Les processus de lecture en langue première et seconde

La lecture n'est pas une activité simple ou linéaire. Au contraire, il s'agit d'un processus cognitif complexe qui requiert la mobilisation coordonnée de multiples

opérations interdépendantes. Comme nous pouvons le lire chez Barberán et al. (2024, p. 6306) : « La lecture est un processus cognitif complexe impliquant le décodage et l'interprétation du langage écrit. Au-delà de la simple reconnaissance des mots, les lecteurs construisent du sens en s'appuyant sur leurs connaissances préalables, en formulant des inférences et en s'engageant dans une démarche d'analyse approfondie du texte » (notre traduction). Cette définition met en lumière la double dimension de la lecture : d'une part, les processus de décodage permettant de reconnaître les mots écrits ; d'autre part, les processus de compréhension qui ouvrent l'accès au sens du texte.

Dans cette section, nous proposons d'explorer ces différents processus. Nous commencerons par préciser ce qu'est la lecture, puis nous adopterons une approche componentielle pour examiner les principales problématiques liées à cette activité cognitive. Enfin, nous présenterons les principales théories de la lecture, en accordant une attention particulière à celles qui concernent la lecture en langue seconde.

1.2.1 Essai de définition : la lecture

La lecture est une activité cognitive complexe ayant fait l'objet de nombreuses définitions au fil des décennies. Selon Goodman (1967, pp. 126-135), elle peut être considérée comme un "jeu de devinettes psycholinguistiques" où le lecteur construit le sens du texte en formulant des hypothèses qu'il vérifie ensuite à partir des indices fournis par le texte. Cette conception, qui s'inscrit dans une approche descendante (top-down) de la lecture, met l'accent sur le rôle actif du lecteur dans le processus de compréhension.

À l'opposé, Gough (1972, p. 291) définit la lecture comme un processus séquentiel et linéaire de décodage, où le lecteur convertit les graphèmes en phonèmes pour accéder au sens des mots. Cette vision, qui relève d'une approche ascendante (bottom-up), insiste sur l'importance des processus de bas niveau dans l'activité de lecture.

Ces deux conceptions, longtemps considérées comme contradictoires, sont aujourd'hui envisagées comme complémentaires. Ainsi, selon le modèle interactif proposé par Rumelhart (1985, pp. 731-732), la lecture implique à la fois des processus ascendants et des processus descendants interagissant de manière dynamique. Le lecteur s'appuie à la fois sur ses connaissances linguistiques (lexicales, syntaxiques, sémantiques) et sur ses connaissances du monde pour construire un sens au texte.

Dans le contexte spécifique de la lecture en langue étrangère, Bernhardt (2011, p. 10) propose une définition qui tient compte des particularités de cette situation : "La lecture en L2 est un processus de construction de sens qui implique l'interaction entre le lecteur et le texte, et qui est influencé par les connaissances linguistiques et conceptuelles du lecteur dans sa L1 et sa L2, ainsi que par les différences entre ces deux langues" (traduction libre de l'anglais). Cette définition met en évidence l'influence de la langue première sur le processus de lecture en langue seconde, un aspect réellement pertinent pour notre étude sur la lisibilité des textes en néerlandais langue étrangère.

Pour le néerlandais particulièrement, cette interaction entre L1 et L2 prend différentes formes selon la langue maternelle du lecteur. En effet, pour les locuteurs de langues germaniques comme l'allemand ou encore l'anglais, les similitudes lexicales et syntaxiques avec le néerlandais peuvent faciliter la compréhension. Par contre, pour les locuteurs de langues romanes comme le français ou l'espagnol, les différences typologiques peuvent constituer un obstacle supplémentaire (Hiligsmann et Rasier, 2006, pp. 10-11). Il s'agit là de considérations essentielles pour bien comprendre les défis spécifiques que pose la lecture en néerlandais langue étrangère.

De plus, il est important de noter que le néerlandais présente quelques particularités linguistiques pouvant influencer le processus de lecture. Parmi celles-ci, nous pouvons citer :

1. La formation de mots composés : le néerlandais est une langue réputée pour la productivité de la composition lexicale, ce qui signifie que de nouveaux mots composés peuvent être aisément, mais aussi fréquemment créés. Cette caractéristique engendre la présence de mots très longs et complexes dans les textes écrits (par exemple, *Kindercarnavalsoptochtvoorbereidingswerkzaamhedenplan* pour *plan des activités de préparation du défilé de carnaval pour enfants*). Cette morphologie agglutinante pose un défi cognitif particulier aux lecteurs non natifs, qui doivent être en mesure de segmenter mentalement ces mots en unités significatives pour accéder à leur sens global. de Jong et al. (2002) ont démontré que ces constructions complexes engendrent une charge cognitive spécifique chez le lecteur, notamment lors de tâches de décision lexicale. Leurs expérimentations ont montré que la reconnaissance de ces mots nécessite un traitement non seulement morphologique, mais également positionnel des constituants internes.
2. L'ordre des mots : la syntaxe du néerlandais présente des spécificités qui peuvent perturber la compréhension en lecture, notamment pour les apprenants dont la langue maternelle ne partage pas ces caractéristiques. L'une des difficultés majeures réside dans la position finale du verbe dans les subordonnes (verbe en fin de proposition) et la séparation des particules verbales, qui exigent du lecteur un maintien prolongé de l'information en mémoire de travail jusqu'à l'apparition du verbe principal. Ce phénomène complique la construction en temps réel du sens de la phrase, car le lecteur doit souvent attendre la fin de la proposition pour intégrer l'information verbale essentielle.
3. Les faux amis : le néerlandais partage de nombreux mots apparentés avec d'autres langues germaniques et, dans un moindre mesure, avec les langues romanes. Cependant, la présence de faux amis (des mots qui se ressemblent mais qui ont un sens qui diffère) constitue un obstacle supplémentaire pour les apprenants. Ces similitudes trompeuses peuvent induire des erreurs de compréhension, car le lecteur pourrait attribuer à un mot néerlandais le sens qu'il possède dans sa langue maternelle alors qu'il en a un autre en néerlandais. Chuquet et Paillard (1987, p. 224) soulignent que les réelles difficultés apparaissent quand une certaine parenté sémantique s'ajoute à une ressemblance graphique, sans que les deux termes aient le même sens et puissent être traduits l'un par l'autre. La prise en compte de ces interférences lexicales est donc essentielle dans toute réflexion sur la lisibilité des textes néerlandais

pour un public non natif.

Ces spécificités linguistiques du néerlandais doivent être prises en compte dans chaque réflexion sur la lisibilité des textes dans cette langue, particulièrement dans un contexte d'apprentissage comme langue étrangère.

Pour résumer, la lecture en néerlandais langue étrangère peut être définie comme un processus complexe de construction de sens qui implique l'interaction entre les connaissances linguistiques et conceptuelles du lecteur dans sa langue maternelle et en néerlandais, et qui est influencée par les spécificités linguistiques du néerlandais, mais aussi par les différences entre cette langue et sa langue maternelle.

1.2.2 Les grandes problématiques de la lecture : une approche componentielle

Afin de mieux comprendre les processus impliqués dans la lecture, il est utile d'adopter une approche componentielle distinguant les différentes opérations cognitives mises en œuvre par le lecteur. Cette approche, notamment développée par Hoover et Gough (1990) permet d'identifier les composantes essentielles de la lecture et d'analyser leurs interactions.

Les processus de bas niveau : la reconnaissance des mots.

La reconnaissance des mots constitue la première étape du processus de lecture. Elle implique plusieurs opérations cognitives, dont l'analyse visuelle des lettres, l'accès au lexique mental et l'activation des représentations phonologiques et sémantiques associées aux mots.

En néerlandais, la reconnaissance des mots présente certaines particularités liées aux caractéristiques orthographiques et morphologiques de la langue. Contrairement au français, le néerlandais possède une orthographe relativement transparente. En effet, la correspondance entre les graphèmes et phonèmes est assez régulière. Cette transparence orthographique peut faciliter le décodage pour les apprenants débutants, mais elle n'élimine pas pour autant toutes les difficultés.

En effet, comme mentionné précédemment, la tendance du néerlandais à former des mots composés longs peut compliquer la reconnaissance des mots pour les apprenants. Selon de Jong et al. (2002), la décomposition morphologique joue un rôle important dans la reconnaissance des mots composés en néerlandais, ce qui implique une charge cognitive supplémentaire pour les lecteurs non natifs.

Par ailleurs, la recherche a montré que la fréquence des mots influence significativement leur reconnaissance, que ce soit en L1 ou en L2 (Dijkstra et Van Heuven, 2002, p. 183). Les mots fréquents sont reconnus plus rapidement et avec moins d'effort que les mots rares. Cette observation est très pertinente pour notre étude sur la lisibilité, car celle-ci suggère que la fréquence lexicale constituerait un prédicteur important de la difficulté d'un texte en néerlandais langue étrangère.

Les processus de haut niveau : la compréhension

Au-delà de la simple reconnaissance des mots, la lecture implique des processus de haut niveau qui permettent d'accéder à la signification globale du texte. Nous présenterons ici l'analyse syntaxique, l'intégration sémantique, l'inférence et la construction d'un modèle mental cohérent du texte.

L'analyse syntaxique consiste à identifier les relations grammaticales entre les mots pour construire des unités de sens plus larges (des propositions, des phrases). En néerlandais, on repère certaines structures particulières telles que la position du verbe dans les propositions subordonnées (par exemple "*Ik weet* ***dat*** *hij morgen* ***komt***" pour "*Je sais qu'il vient demain*") et la séparation des particules verbales (par exemple "*Hij* ***belt*** *zijn moeder* ***op***" pour "*Il appelle sa mère au téléphone*"). Si ces caractéristiques peuvent initialement sembler complexes, leur impact sur la compréhension dépend fortement du niveau de maîtrise syntaxique du lecteur et de sa familiarité avec la structure du néerlandais.

En effet, pour un lecteur qui maîtrise la syntaxe du néerlandais, la position finale du verbe dans les subordonnées (SOV) ou la séparation des particules verbales ne constitue pas nécessairement un obstacle majeur. Les recherches de Jordens (2006, p. 11) ont montré que l'ordre des mots en néerlandais, particulièrement la position finale du verbe dans les propositions subordonnées, est un obstacle majeur pour les apprenants ayant pour L1 une langue à ordre SVO, comme c'est le cas du français. Néanmoins, des travaux montrent que les lecteurs expérimentés anticipent souvent la position du verbe grâce à une bonne connaissance des contraintes syntaxiques du néerlandais, ce qui facilite l'analyse et la compréhension de la phrase (Koster, 1975, pp. 111-112). Ainsi, la difficulté n'est réelle que pour les apprenants ou les lecteurs non familiers de ces structures, qui doivent maintenir en mémoire les éléments de la proposition jusqu'à l'apparition du verbe, ce qui augmente la charge cognitive (Collard et al., 2019).

Concernant la séparation des particules verbales, typiques des verbes à particule, bien qu'on puisse supposer qu'elle représente une difficulté importante pour les apprenants, les données de Herbay et al. (2018) montrent que cette difficulté dépend fortement de deux facteurs : la connaissance lexicale des verbes à particules séparables et la capacité de mémoire de travail. Les participants combinant une bonne connaissance lexicale des verbes à particule séparable et une haute capacité de traitement présentent des temps de lecture similaires à ceux des locuteurs natifs lorsqu'ils doivent traiter des structures où la particule est séparée du verbe, notamment lorsque le groupe nominal est long ou que la dépendance sémantique est élevée (Herbay et al., 2018, pp. 9-13).

L'intégration sémantique, quant à elle, consiste à combiner les significations des mots et des phrases afin de construire une représentation du texte cohérente. Ce processus s'appuie sur les connaissances linguistiques et conceptuelles du lecteur, ainsi que sur sa capacité à établir des liens entre différentes parties du texte (Sanders et Spooren, 2012). Dans une langue étrangère, l'intégration sémantique peut être entravée par des lacunes lexicales ou encore par une maîtrise insuffisante des connecteurs et autres marqueurs de cohésion. Les connecteurs jouent un rôle crucial dans

la compréhension des relations logiques entre les propositions et les phrases.

Comme l'a montré Perrez (2006), la compréhension et l'usage approprié des connecteurs causaux et contrastifs en néerlandais langue étrangère (comme "*omdat*", "*want*", "*maar*" ou encore "*hoewel*") constituent un défi majeur pour les apprenants. Cette difficulté s'explique d'une part par la complexité conceptuelle des relations logiques qu'ils expriment, comme le souligne la *conceptuelecomplexiteitshypothese* (ou hypothèse de complexité conceptuelle), et d'autre part par l'absence de correspondance directe entre les connecteurs néerlandais et leurs équivalents dans la langue maternelle, ce qui décrit l'*equivalentiehypothese* (ou hypothèse d'équivalence) (Perrez, 2006, pp. 243-245).

De leur côté, Degand et Sanders (2002) ont montré que les connecteurs jouent un rôle essentiel dans l'établissement de la cohérence textuelle. Leur présence facilite la construction d'une représentation mentale intégrée du texte, tant en langue première qu'en langue seconde, en particulier dans les textes expositifs. Leur étude démontre que les marqueurs relationnels, tels que les connecteurs causaux et contrastifs, améliorent la compréhension globale du texte, et ce même chez les lecteurs de L2, pour autant que leur niveau de compétence soit suffisant (Degand et Sanders, 2002, pp. 740–743, 750-753).

Enfin, la construction d'un modèle mental cohérent du texte implique la capacité à faire des inférences, à déduire des informations qui ne sont pas explicitement mentionnées dans le texte. Cette capacité dépend non seulement des connaissances linguistiques du lecteur, mais aussi de ses connaissances du monde et de sa familiarité avec le sujet. En langue étrangère, la charge cognitive associée au décodage et à l'analyse syntaxique peut limiter les ressources disponibles pour les processus inférentiels, ce qui peut entraver la compréhension globale des textes (Koda, 2005, p. 147).

L'interaction entre processus de bas et de haut niveau

Les processus de bas et de haut niveau ne fonctionnent pas isolément, mais interagissent constamment tout au long de la lecture. Selon le modèle interactif-compensatoire de Stanovich (1980), les lecteurs sont capables de compenser des déficiences dans un niveau de traitement en s'appuyant davantage sur d'autres niveaux. Un lecteur peut par exemple s'appuyer sur un contexte et ses connaissances préalables pour comprendre un texte, malgré le fait qu'il rencontre des difficultés dans la reconnaissance des mots.

Cependant, comme le souligne Bernhardt (2005, pp. 138-142), cette compensation a ses limites en langue étrangère, surtout chez les débutants et intermédiaires où les connaissances linguistiques sont encore insuffisantes pour permettre une compréhension fluide. C'est pourquoi il est important de tenir compte du niveau de compétence des apprenants dans l'évaluation de la lisibilité des textes en néerlandais langue étrangère.

En résumé, la lecture en néerlandais langue étrangère implique une interaction complexe entre processus de bas et de haut niveau, influencée par les spécificités

linguistiques du néerlandais et par les différences entre cette langue et la langue maternelle des apprenants. La compréhension de tous ces processus est fondamentale pour identifier les facteurs déterminants de la lisibilité des textes en néerlandais langue étrangère.

1.3 La place de la lecture et du texte écrit dans les différentes approches didactiques en néerlandais langue étrangère

1.3.1 Évolution des modèles théoriques de la lecture : vers une compréhension intégrée

Les recherches en psycholinguistique ont considérablement évolué depuis les premiers modèles théoriques de la lecture. Comme nous l'avons vu précédemment, les approches ascendantes, descendantes et interactives ont posé les bases de notre compréhension des processus de lecture. Cependant, l'émergence de nouvelles technologies et méthodologies a permis de développer des modèles plus sophistiqués et pertinents pour comprendre la lecture en langue étrangère et donc, par extension, pour développer des outils d'évaluation automatique de la lisibilité.

La prochaine section se concentrera sur les développements théoriques récents qui dépassent les modèles traditionnels présentés précédemment, tout en mettant l'accent sur leurs implications pour l'évaluation de la lisibilité du néerlandais langue étrangère.

Les modèles connexionnistes et l'apprentissage automatique.

Les modèles connexionnistes représentent une avancée majeure dans la compréhension des processus de lecture, en particulier pour leur capacité à simuler l'apprentissage et l'adaptation du système cognitif. Ces modèles sont inspirés du fonctionnement des réseaux neuronaux et offrent une perspective particulièrement riche pour comprendre l'acquisition de la lecture en langue étrangère.

Les développements récents en modélisation bayésienne de l'apprentissage de la lecture, notamment les travaux de Steinhilber (2023), apportent une perspective nouvelle sur la façon dont les lecteurs intègrent différentes sources d'information de manière probabiliste. Ces modèles suggèrent que les lecteurs développent simultanément des attentes probabilistes sur la structure des mots (orthographique, phonologique, lexicale) et sur leur interprétation contextuelle, et qu'ils ajustent ces prédictions en continu en fonction des nouvelles informations rencontrées dans le texte. Cette approche est particulièrement éclairante pour comprendre comment les apprenants du néerlandais développent progressivement la capacité à traiter les spécificités de cette langue, comme les mots composés complexes ou les structures syntaxiques particulières.

L'application de l'apprentissage automatique aux modèles de lecture a également ouvert de nouvelles perspectives pour l'évaluation automatique de la lisibilité. Les

techniques de TAL permettent désormais d'extraire et d'analyser automatiquement un large éventail de variables linguistiques qui étaient auparavant difficiles à quantifier.

François et Fairon (2013) ont démontré que l'utilisation de variables TAL, comme les mesures de cohésion lexicale basées sur l'analyse sémantique latente (LSA) ou les ratios de catégories grammaticales, peut considérablement améliorer la prédiction de la difficulté des textes en français langue étrangère. Ces approches computationnelles dépassent largement les limites des formules traditionnelles de lisibilité, qui se contentaient souvent de mesures plus superficielles telles que la longueur des mots ou des phrases par exemple. Nous aborderons ces formules traditionnelles en profondeur au chapitre suivant.

1.3.2 Modèles spécifiques à la lecture bilingue et multilingue.

La compréhension de la lecture en langue étrangère a été considérablement enrichie par le développement de modèles spécifiquement conçus pour rendre compte des particularités du traitement bilingue. Ces modèles sont essentiels pour bien comprendre les défis spécifiques auxquels sont confrontés les apprenants du néerlandais langue étrangère.

Le modèle BIA+ (*Bilingual Interactive Activation Plus*) de Dijkstra et Van Heuven (2002) est une des contributions les plus significatives dans ce domaine. Ce modèle propose que les représentations lexicales des deux langues d'une personne bilingue sont intégrées dans un système lexical unique, où les mots des deux langues sont activés en parallèle lors de la reconnaissance. Cette activation parallèle explique pourquoi les cognats (mots similaires dans les deux langues) facilitent la compréhension, mais également pourquoi les "faux amis" peuvent induire en erreur. Ce modèle permet de comprendre, pour les apprenants du néerlandais, comment des mots comme "informatie" (information) ou "universiteit" (université) peuvent être facilement reconnus par des francophones grâce à leur similarité avec les mots français équivalents. Alors que de faux amis comme "de bol" (la sphère) ou "de sfeer" (l'ambiance) peuvent activer des représentations sémantiques incorrectes.

Le modèle BIA+ a été étendu pour inclure des facteurs contextuels et pragmatiques qui influencent la sélection de la langue appropriée. Ces extensions sont particulièrement pertinentes dans le contexte belge, où les apprenants du néerlandais sont souvent exposés à des environnements multilingues où le français et le néerlandais coexistent.

Le modèle de transfert cross-linguistique de Koda (2005) apporte une perspective complémentaire en se concentrant sur la façon dont les compétences de lecture en langue première influencent le développement des compétences en langue seconde. Ce modèle distingue plusieurs types de transfert : les transferts de stratégies métacognitives (comme les stratégies de compréhension globale), les transferts de connaissances linguistiques (comme la conscience phonologique), et les transferts de connaissances procédurales (comme les automatismes de reconnaissance des mots). Pour les apprenants du néerlandais, l'efficacité de ces transferts dépend largement de la distance typologique entre la langue maternelle et le néerlandais. Les locuteurs

de langues germaniques peuvent bénéficier de transferts positifs plus importants que les locuteurs de langues romanes comme le français.

1.3.3 Implications pour l'évaluation automatique de la lisibilité

Les développements théoriques récents ont des implications directes pour l'évaluation automatique de la lisibilité. Les modèles connexionnistes et les approches d'apprentissage automatique suggèrent que la difficulté d'un texte ne peut pas être réduite à quelques variables superficielles, mais doit être évaluée à travers un ensemble complexe d'interactions entre différents niveaux linguistiques.

Les recherches de François et Fairon (2013) ont démontré l'efficacité des approches basées sur les machines à vecteurs de support (SVM) pour l'évaluation automatique de la lisibilité en français langue étrangère. Leur modèle, qui intègre 46 variables représentatives des niveaux lexical, syntaxique et sémantique, surpasse significativement les formules traditionnelles de lisibilité.

Au niveau lexical, les variables prédictives n'incluent pas seulement la fréquence et la longueur des mots, mais également des mesures plus sophistiquées comme la densité des mots composés, la proportion de verbes à particules séparables, ou encore la diversité lexicale. Pour le néerlandais, la formation productive de mots composés représente un défi particulier, car ces formations peuvent créer des mots très longs et complexes.

Au niveau syntaxique, les variables pertinentes incluent la complexité des structures phrastiques, mesurée par des indices comme la profondeur d'enchâssement syntaxique, la fréquence des constructions passives, ou encore la proportion de propositions subordonnées avec inversion verbale (caractéristique du néerlandais). En effet, comme nous l'avons vu plus haut, les recherches de Jordens (2006) ont montré que l'ordre des mots en néerlandais, particulièrement la position finale du verbe dans les propositions subordonnées, est un obstacle majeur pour les apprenants ayant pour L1 une langue à ordre SVO comme le français.

Au niveau sémantique, les mesures de cohésion textuelle basées sur l'analyse sémantique latente (LSA) (Bestgen, 2004) ou sur des modèles de représentation distributionnelle plus récents comme *Word2Vec* (Crossley et al., 2019) ou *BERT* (Lei et al., 2021) permettent d'évaluer la cohérence thématique et la progression informationnelle du texte. Ces mesures sont particulièrement importantes pour évaluer la charge cognitive imposée par le traitement sémantique, qui peut être considérablement augmentée en langue étrangère.

Vers une approche intégrée de la lisibilité en néerlandais langue étrangère

L'intégration des perspectives théoriques récentes suggère qu'une évaluation efficace de la lisibilité en néerlandais langue étrangère doit adopter une approche multidimensionnelle qui tient compte des spécificités de cette langue et des défis particuliers auxquels sont confrontés les apprenants selon leur langue maternelle.

Les modèles connexionnistes soulignent l'importance de considérer les interactions entre différents niveaux de traitement. Par exemple, un texte contenant de nombreux mots composés complexes peut imposer une charge cognitive plus importante au niveau de la reconnaissance lexicale, ce qui peut limiter les ressources disponibles pour les processus de compréhension de niveau supérieur. Cette perspective suggère que l'évaluation de la lisibilité doit tenir compte non seulement de la difficulté intrinsèque de chaque variable linguistique, mais aussi de leurs interactions potentielles.

Les modèles bilingues mettent en évidence l'importance de considérer la langue maternelle des apprenants dans l'évaluation de la lisibilité. Un texte contenant de nombreux cognats sera probablement plus facile pour un apprenant germanophone que pour un apprenant francophone, même si d'autres variables (comme la complexité syntaxique) restent constantes. Cette perspective suggère que les formules de lisibilité pourraient bénéficier d'une adaptation en fonction du profil linguistique des apprenants cibles.

Enfin, les approches d'apprentissage automatique offrent la possibilité de développer des modèles adaptatifs qui peuvent être continuellement améliorés à mesure que de nouvelles données deviennent disponibles. Cette capacité d'adaptation est importante dans le contexte de l'enseignement du néerlandais langue étrangère, où les profils d'apprenants et les contextes d'apprentissage peuvent considérablement varier.

Pour conclure, l'évolution des modèles théoriques de la lecture vers des approches plus intégrées et sophistiquées ouvre de nouvelles perspectives pour l'évaluation automatique de la lisibilité en néerlandais langue étrangère. Ces développements suggèrent qu'une formule de lisibilité efficace pour cette langue se doit d'aller au-delà des mesures traditionnelles pour intégrer une gamme plus large de variables linguistiques, tenir compte des spécificités du néerlandais, et considérer les particularités du traitement bilingue. Cette approche multidimensionnelle et théoriquement informée constitue la base méthodologique nécessaire pour développer des outils d'évaluation de la lisibilité qui soient à la fois scientifiquement rigoureux et pratiques pour l'enseignement du néerlandais langue étrangère.

Chapitre 2

Une approche particulière de la lecture : la lisibilité

2.1 Définition

2.1.1 La lisibilité

La lisibilité est un concept qui a fait l'objet de nombreuses définitions au fil des décennies, reflétant l'évolution des perspectives théoriques sur la lecture et la compréhension des textes. Dans cette section, nous proposons d'explorer ces différentes définitions et de préciser celle que nous adopterons dans le cadre de notre étude sur la lisibilité du néerlandais langue étrangère.

L'une des définitions les plus citées est celle de Dale et Chall (1949), qui définissent la lisibilité comme "l'ensemble des caractéristiques d'un texte qui affectent la réussite qu'un groupe de lecteurs peut avoir avec ce texte. La réussite est le degré avec lequel ils comprennent le texte, le lisent à une vitesse optimale et le trouvent intéressant." (notre traduction). Cette définition met en évidence trois dimensions essentielles de la lisibilité : la compréhension, la fluidité de lecture et l'intérêt suscité par le texte.

Klare (1963) propose une définition plus concise, centrée sur la facilité de compréhension : "La lisibilité concerne la facilité de compréhension due au style d'écriture." Cette définition met l'accent sur les caractéristiques stylistiques du texte, par opposition à son contenu ou à sa structure.

McLaughlin (1969), quant à lui, définit la lisibilité comme le degré auquel un lecteur donné peut comprendre un texte donné à un niveau spécifique de lecture (notre traduction et reformulation). Cette définition introduit explicitement la notion de niveau de lecture et souligne l'interaction entre les caractéristiques du texte et les compétences du lecteur.

Dans le contexte spécifique des langues étrangères, Crossley et al. (2008a) proposent une définition qui tient compte des particularités de la lecture en L2. Selon

eux, la lisibilité en L2 ne peut être comprise qu'en tenant compte de la manière dont les apprenants mobilisent leurs compétences linguistiques pour décoder, analyser syntaxiquement et construire le sens d'un texte. Cette approche souligne que la lisibilité d'un texte dépend de son adéquation avec les capacités de traitement linguistique et cognitives des lecteurs L2, plutôt que de caractéristiques formelles seules.

Dans le cadre de notre étude, nous adopterons une définition de la lisibilité qui s'inspire de ces différentes perspectives, tout en tenant compte des spécificités du néerlandais langue étrangère. Ainsi, nous définirons ici la lisibilité comme "le degré de facilité avec lequel un texte en néerlandais peut être lu et compris par des apprenants de cette langue, en fonction de leurs compétences linguistiques en néerlandais, de leurs connaissances préalables et de l'influence de leur langue maternelle." Notre définition se concentre spécifiquement sur les aspects linguistiques et cognitifs qui déterminent la facilité avec laquelle un texte peut être lu et compris des apprenants du néerlandais.

2.1.2 Les formules de lisibilité

Les formules de lisibilité sont des outils qui visent à quantifier la difficulté d'un texte en se basant sur certaines de ses caractéristiques linguistiques. Elles constituent une application pratique du concept de lisibilité et ont été développées pour répondre à divers besoins pédagogiques et éditoriaux.

Historiquement, les premières formules de lisibilité ont été développées pour l'anglais dans les années 1920, avec des travaux pionniers comme ceux de Lively et Pressey (1923) qui ont établi une corrélation entre la fréquence des mots et la difficulté des textes (il ne s'agit donc pas encore d'une formule au sens strict). Par la suite, de nombreuses autres formules ont été proposées, dont les plus célèbres sont dues à Lorge (1944), Dale et Chall (1948) et Gunning (1952)

Pour le néerlandais, les premières formules ont été développées dans les années soixante avec l'adaptation de formules anglo-saxonnes aux spécificités de la langue néerlandaise. La première formule significative est la *Flesch-Douma* (Douma, 1960), directement inspirée de la formule *Flesch Reading Ease* anglaise. Trois ans plus tard apparaît la *Leesindex A* (Brouwer, 1963), également basée sur la formule *Flesch Reading Ease*. Cette formule est utilisée jusqu'en 2008 pour évaluer le niveau de lecture des enfants dans l'enseignement primaire (niveau AVI : *Analyse van Individualiseringsvormen*.

L'évolution vers des formules plus spécialisée s'est concrétisée dans les années nonante avec le développement par l'organisation *Cito* (Cito, Centraal Instituut voor Toets Ontwikkeling, 2023) de deux nouvelles formules : le *CLIB* (*Cito leesbarheidsindex voor het basinsonderwijs*), conçu en 1994, spécifiquement pour évaluer la compréhension de lecture des enfants dans l'enseignement primaire (Staphorsius, 1994), et le *CILT* (*Cito leesindex technisch lezen*) en 1997 (Staphorsius et Verhelst, 1997), qui se concentre sur les compétences techniques de lecture et qui remplace la *Leesindex* A pour le calcul des niveaux AVI à partir de 2008.

Les formules de lisibilité traditionnelles se caractérisent généralement par leur simplicité : elles se basent sur un nombre limité de variables qui sont faciles à mesurer. La longueur moyenne des phrases ou des mots, ou encore la proportion de mots difficiles par exemple. La formule de Flesch-Douma (Douma, 1960) se calcule comme suit :

$$RE = 206.84 - 77 \times \left(\frac{syl}{w}\right) - 0.93 \times \left(\frac{w}{sen}\right)$$

Où RE est l'indice de lisibilité (Readability Ease), $\frac{syl}{w}$ est la longueur moyenne des mots en syllabes (nombre total de syllabes divisé par le nombre de mots) et $\frac{w}{sen}$ est la longueur moyenne des phrases en mots (nombre total de mots divisés par le nombre de phrases). Le nombre "206.84" représente la constante de départ de la formule. Elle représente la valeur maximale théorique de lisibilité pour un texte composé uniquement de mots monosyllabiques et de phrases très courtes (c'est-à-dire les textes les plus faciles à lire). Il est suivi par le coefficient "-0.77" qui indique l'impact de la longueur moyenne des mots (en syllabes) sur la lisibilité. Plus les mots sont longs, plus la lisibilité diminue. Ce coefficient représente en fait la "pénalité" appliquée à la lisibilité pour chaque syllabe supplémentaire en moyenne par mot. Quant au coefficient "-0.93", il indique l'impact de la longueur moyenne des phrases (en mots) sur la lisibilité. Plus les phrases sont longues, plus la lisibilité diminue. Cette formule utilise une pente négative, ce qui signifie que plus le score est élevé plus le texte est facile à lire.

Ces formules traditionnelles présentent des avantages. En effet, elles sont faciles à calculer, à comprendre et à appliquer, et elles fournissent une estimation rapide de la difficulté d'un texte. Cependant, elles font également l'objet de nombreuses critiques que nous allons énumérer ici :

1. Une simplicité excessive : en se basant sur un nombre très limité de variables, ces formules ne peuvent pas refléter toute la complexité des facteurs qui influencent la lisibilité d'un texte.
2. Un manque de fondement théorique : la plupart de ces formules ont été développées de manière empirique, sans s'appuyer sur une théorie solide de la lecture et de la compréhension.
3. Une négligence des aspects sémantiques et pragmatiques : en se concentrant sur des caractéristiques de surface comme la longueur des mots et des phrases, ces formules ignorent certains aspects importants tels que la cohérence du texte, la densité conceptuelle ou la familiarité du sujet.

Face à ces limites, des approches plus récentes de la lisibilité ont cherché à développer des modèles plus sophistiqués, qui prennent en considération un plus grand nombre de variables linguistiques, et qui s'appuient sur des théories plus solides de la lecture et de la compréhension. Ces approches utilisent des techniques avancées de traitement automatique du langage et d'apprentissage automatique pour analyser les textes et prédire leur difficulté. On pourrait qualifier ces approches de "computationnelles".

Dans le contexte du néerlandais, Kleijn (2018) a exploré l'utilisation de telles approches computationnelles pour évaluer la lisibilité des textes pour les adolescents

néerlandophones. Son modèle, appelé U-Read, prend en compte une variété de caractéristiques linguistiques, allant de la fréquence lexicale à la complexité syntaxique, en passant par la cohésion textuelle.

Néanmoins, ces approches n'ont pas encore été entièrement appliquées au contexte spécifique du néerlandais langue étrangère. C'est précisément cette lacune que notre étude vise à combler, en cherchant à faire progresser la recherche sur l'évaluation de la difficulté textuelle pour ce public particulier.

Notre contribution s'articule autour d'une démarche de recherche quantitative en empirique. L'objectif n'est pas de proposer une formule de lisibilité définitive, mais plutôt de jeter les bases scientifiques qui pourraient, à terme, mener à son développement. Pour ce faire notre approche se distingue par plusieurs aspects fondamentaux :

— Elle s'appuie sur un cadre théorique solide, en tenant compte des spécificités de lecture en langue étrangère et des particularités linguistiques du néerlandais.
— Elle intègre un large éventail de variables linguistiques, au-delà des simples mesures de longueur, en incluant des aspects lexicaux et syntaxiques, dont l'influence sur la difficulté est analysée statistiquement. Une attention particulière est portée à des caractéristiques propres au néerlandais.
— Elle met à l'épreuve une méthodologie fiable et reproductible, en employant des techniques de traitement automatique du langage pour extraire les variables.

En somme, cette étude a pour ambition de fournir à la communauté scientifique des données, des outils et des analyses approfondies sur ce qui constitue la difficulté d'un texte en néerlandais pour un non-natif. Nos résultats ont vocation à paver la voie pour de futures recherches et, potentiellement, pour la création d'outils d'évaluation plus justes et plus efficaces.

2.2 Conclusion

L'évolution des recherches sur la lisibilité, des premières formules empiriques jusqu'aux approches computationnelles contemporaines, révèle une constante progression vers des modèles de plus en plus sophistiqués et théoriquement fondés. Cette évolution s'accompagne d'une prise de conscience progressive des limites des approches traditionnelles et de la nécessité de développer des outils spécialement adaptés aux contextes d'apprentissage des langues étrangères.

Les formules classiques, bien qu'utiles pour une première estimation de la difficulté textuelle, présentent des limites importantes, notamment lorsqu'elles sont appliquées aux apprenants de langues étrangères. Le fait qu'elles se focalisent sur des variables de surface comme la longueur des mots et des phrases, la fréquence lexicale générale, ... ne permet pas de saisir les défis spécifiques auxquels sont confrontés les apprenants de L2 : par exemple l'influence de la langue maternelle des apprenants, la complexité des structures syntaxiques ou encore l'importance des cognats et des faux amis.

L'émergence des approches computationnelles ouvre de nouvelles perspectives en permettant l'analyse automatique d'un large éventail de variables linguistiques et

l'utilisation de techniques d'apprentissage automatique pour la modélisation de la complexité de la lecture en langue étrangère. Les travaux pionniers de François (2011) sur le français langue étrangère (FLE) ont démontré la faisabilité et la pertinence de formules de lisibilité fondées sur le traitement automatique du langage. En combinant des prédicteurs linguistiques variés et des techniques d'apprentissage statistique, ces approches se sont révélées plus précises que les formules traditionnelles fondées sur un nombre limité de variables, notamment dans un cadre aligné sur les niveaux du CECR.

Pour le néerlandais langue étrangère, le domaine reste inexploré. Les spécificités linguistiques de la langue créent des défis spécifiques pour les apprenants qui ne sont pas pris en compte par les formules de lisibilité existantes, conçues pour des lecteurs natifs.

Cette lacune justifie le développement d'un modèle de lisibilité spécifiquement adapté au néerlandais langue étrangère. Un tel modèle doit s'appuyer sur une compréhension fine des processus de lecture en L2, intégrer des variables linguistiques pertinentes pour les apprenants du néerlandais, et être validé sur des corpus de textes annotés selon des échelles de compétence reconnues comme le CECR. A travers ce travail, notre objectif est de faire un pas dans cette direction. En testant la pertinence de multiples variables linguistiques sur un corpus adapté, nous cherchons à établir une base empirique solide qui guidera et facilitera le développement futur de modèles de lisibilité pour le néerlandais langue étrangère.

Dans les chapitres suivants, nous présenterons notre démarche qui s'articulera en deux temps principaux.

Dans un premier temps, nous exposerons en détail la méthodologie mise en œuvre. Ce chapitre sera consacré à la présentation des fondations de notre étude : nous y définirions précisément l'ensemble des variables linguistiques que nous avons choisi d'examiner, en justifiant leur pertinence théorique pour évaluer la difficulté textuelle en néerlandais langue étrangère. Ensuite, nous décrirons le processus de constitution de notre corpus de textes, spécifiquement assemblé pour les besoins de la recherche.

Dans un second temps, nous nous concentrerons sur l'analyse quantitative de nos données. Nous présenterons d'abord les résultats de nos statistiques descriptives pour offrir une vue d'ensemble du corpus et des variables. Par la suite, nous emploierons des tests ANOVA, pour comparer les groupes et identifier des différences significatives. Enfin, nous réaliserons une analyse de régression afin de déterminer quelles variables linguistiques sont les plus prédictives de la difficulté textuelle.

Cette approche structurée nous permettra de fournir des conclusions empiriques robustes, contribuant ainsi à une meilleure compréhension des facteurs de lisibilité en néerlandais langue étrangère et ouvrant la voie à de futures recherches appliquées.

Deuxième partie

Méthodologie : un modèle pour la mesure de la difficulté

Chapitre 3

Quelles variables pour un modèle en néerlandais langue étrangère ?

3.1 Introduction

Maintenant que nous avons exploré les fondements théoriques de la lecture et de la lisibilité dans les précédents chapitres, nous allons aborder la question centrale de notre recherche : quelles caractéristiques linguistiques permettent de prédire la difficulté d'un texte en néerlandais pour les apprenants de cette langue ? Dans ce chapitre, nous allons identifier et décrire les variables linguistiques qui seront intégrées dans notre modèle de lisibilité.

Le choix des variables constitue l'étape fondamentale dans la construction d'un modèle de lisibilité efficace. Avant de décrire les variable sélectionnées, il est utile de répondre à la question "Que fait un bon indice de la difficulté des textes ?". Une variable linguistique ne peut être considérée comme un bon prédicteur de la difficulté textuelle si elle ne satisfait pas un ensemble de critères, que nous décrivons dans la section suivante. Nous nous appuyons sur les critères présentés par François (2011, pp. 217-218).

3.1.1 Les critères d'un prédicteur de qualité en lisibilité

Corrélation avec le critère de difficulté

Un prédicteur efficace de la difficulté textuelle doit présenter une corrélation élevée avec le critère retenu pour évaluer cette difficulté. Une telle corrélation est souvent perçue comme un indicateur direct de la pertinence d'une variable, c'est pourquoi de nombreux chercheurs s'en sont servis comme outil simple pour estimer l'importance relative des paramètres impliqué dans la compréhension d'un texte. Il faut toute fois relativiser car, comme l'explique notamment Carrell (1987), lorsque l'on restreint la variété des textes ou l'hétérogénéité des lecteurs, la capacité prédictive des formules, bien que les corrélations puissent rester élevées, chute de manière significative.

Complémentarité et non-redondance

De plus, afin d'avoir un modèle qui soit efficace, les différentes variables retenues doivent capturer des aspects différents et complémentaires de la difficulté textuelle. Par exemple, la longueur moyenne des phrases et la complexité syntaxique peuvent sembler liées, mais elles mesurent des phénomènes distincts : une phrase peut être longue mais syntaxiquement simple (coordination multiple) ou courte mais complexe (subordination dense). Cette complémentarité permet d'éviter la redondance entre les variables. En effet, un bon prédicteur doit non seulement être pertinent, mais aussi faiblement corrélé aux variables indépendantes, sans quoi il n'apporte que peu d'information nouvelle au modèle de lisibilité.

Fiabilité méthodologique

La fiabilité d'une mesure concerne sa stabilité et sa reproductibilité dans les temps et entre différents évaluateurs ou outils. Une variable fiable doit produire des résultats constants lorsqu'elle est mesurée à plusieurs reprises sur le même texte dans des conditions identiques. Elle doit donc pouvoir être mesurée de manière objective, et ce, qu'elle soit calculée manuellement ou par différents outils de traitement automatique du langage. Par exemple, le nombre de mots par phrase doit donner le même résultat, qu'il soit calculé manuellement ou de manière automatique grâce au TAL.

Fréquence suffisante

De plus, même si une variable est objectivement mesurable, elle n'a d'intérêt pour un modèle que si elle apparaît fréquemment. Nous reprenons ici l'exemple de François (2011, p. 217) : savoir si un texte a été écrit par Proust peut effectivement indiquer une certaine difficulté pour les textes en français. Mais comme ce cas est très rare, surtout dans un contexte de FLE, cette information sera trop peu présente pour être réellement utile dans une formule de lisibilité. Un bon prédicteur doit donc être à la fois pertinent et suffisamment fréquent pour avoir un impact dans le modèle.

Ancrage théorique et validation empirique

Enfin, une variable pertinente doit être théoriquement fondée. En effet, Bormuth (1969) recommande de n'utiliser que des variables supportées par des arguments théoriques raisonnables, c'est-à-dire, qui entretiennent probablement une relation causale avec la difficulté (François, 2011, p. 218). Par exemple, la fréquence lexicale constitue un prédicteur théoriquement solide car elle est directement liée à la facilité d'accès aux mots dans le lexique mental (Brysbaert et New, 2009, p. 997). Plus un mot est fréquent, plus son traitement cognitif est rapide et automatique.

3.1.2 Organisation du chapitre

Dans les sections suivantes, nous passerons en revue les différentes catégories de variables linguistiques potentiellement pertinentes pour évaluer la lisibilité du néerlandais langue étrangère : les variables lexicales, syntaxiques et celles spécifiques

à la lecture en néerlandais. Pour chaque catégorie, nous décrirons les variables retenues en justifiant leur pertinence au regard des critères exposés ci-dessus et en précisant comment elles peuvent être mesurées à l'aide d'outils de TAL. Nous présenterons ensuite une synthèse des variables sélectionnées pour notre modèle. Enfin, nous proposerons différentes pistes d'améliorations.

3.2 Les variables lexicales

3.2.1 Variables de fréquence lexicale

Justification théorique

L'intégration de variables basées sur la fréquence lexicale dans l'évaluation de la lisibilité des textes repose sur des fondements théoriques et empiriques solides. L'hypothèse fondamentale est qu'il existe une association forte entre la fréquence d'emploi des mots dans une langue et la facilité de leur lecture (François, 2011, p. 219). Autrement dit, un texte qui comporte une proportion élevée de mots rares est plus difficile à décoder et à comprendre.

Cette corrélation entre la fréquence objective des mots et la facilité lexicale a été constamment observée dans la recherche sur la lisibilité (François, 2011, p. 219). Des études psychologiques pionnières, comme celles de Howes et Solomon (1951), ont démontré une corrélation linéaire négative significative entre le logarithme de la fréquence des mots et le temps d'exposition minimal nécessaire à leur reconnaissance. Ce phénomène, connu sous le nom d'effet de fréquence, indique que les mots plus fréquents sont reconnus plus rapidement et plus facilement.

L'effet de fréquence ne se limite pas à la simple reconnaissance des mots : il influence également diverses tâches cognitives telles que la décision lexicale, l'identification perceptive, la prononciation et la catégorisation sémantique, ainsi que les mouvements oculaires lors de la lecture (Lupker, 2005, p. 42). La raison en est que les représentations des mots courants dans le lexique mental sont plus accessibles que celles des mots moins courants (Brysbaert et al., 2000, p. 66). Un décodage plus automatisé des mots fréquents libère des ressources cognitives qui peuvent alors être allouées à des processus de compréhension de niveau supérieur, améliorant ainsi la compréhension globale du texte (Crossley et al., 2008b, p. 488).

En L2, l'effet de fréquence mérite d'être nuancé car la familiarité réelle de l'apprenant avec le vocabulaire est plus pertinente que la fréquence objective dans la langue (Laroche, 1979, p. 132). Cependant, si cette objection semble invalider la mesure de la fréquence objective des mots d'un texte comme cause de la complexité lexicale de celui-ci dans le cadre de la L2, cette information peut tout de même constituer un prédicteur efficace. En effet, pour les apprenants entrant principalement en contact avec des matériaux linguistiques via des manuels, il est évident que leur familiarité avec les mots de la L2 étudiée est dirigée par les ressources pédagogiques qu'ils utilisent, et par la progression établie dans celles-ci. Or, il est fort probable que ces listes reposent sur des listes de fréquences qui ont elles-mêmes été établies à partir des fréquences en L1.

Par conséquent, l'utilisation de mesures de fréquence lexicale est jugée essentielle dans les formules de lisibilité, car elles reflètent directement la charge cognitive associée au décodage des mots et, par extension, la compréhension du texte.

Variables retenues

Nous retenons les variables suivantes, reprises directement de François (2011, p. 296) :

- **mean_fl** (Fréquence moyenne des lemmes) : Cette mesure capture la difficulté lexicale globale du texte. Un texte avec une fréquence moyenne élevée sera plus accessible aux apprenants.
- **median_fl** (Médiane des fréquences des lemmes) : La médiane est moins sensible aux valeurs extrêmes que la moyenne et fournit ainsi une indication plus robuste de la difficulté lexicale typique du texte.
- **75FL et 90FL** (75e et 90e percentiles) : Ces valeurs indiquent les seuils au-delà desquels se situent respectivement les 25 % et les 10 % de mots les plus fréquents du texte. Elles permettent d'évaluer la proportion et le poids des mots rares ou fréquents, et ainsi de mieux cerner l'accessibilité lexicale d'un texte au-delà de la tendance centrale.

Dans son travail, François (2011) a testé un large éventail de variables, incluant les fréquences des lemmes (FL), des formes fléchies (FFF), et des formes fléchies désambiguïsées (FFFD). Notre recherche, bien qu'inspirée par ses travaux, adopte une approche plus ciblée en se concentrant exclusivement sur la fréquence des lemmes. Ce choix nous permet de mieux distinguer la difficulté conceptuelle de la difficulté stylistique (Kleijn, 2018, p. 42). La connaissance du lemme est liée au concept fondamental, alors que la reconnaissance de ses formes relève davantage de la variation stylistique ou de la maîtrise grammaticale. En nous concentrant sur le lemme, nous cherchons à modéliser la difficulté lexicale fondamentale, une approche plus robuste pour évaluer la compréhension réelle d'un texte par des non-natifs.

Adaptation au néerlandais et limites

Pour le néerlandais, nous utilisons les fréquences issues du corpus SoNaR-500 (Oostdijk et al., 2013), qui constitue la ressource de référence la plus complète pour le néerlandais contemporain avec ses 500 millions de mots. Cette ressource nous a semblé préférable aux listes plus anciennes car elle reflète un usage plus actuel de la langue.

Compte tenu des contraintes techniques et des objectifs de notre recherche, nous avons appliqué deux filtres spécifiques à notre extraction des données de fréquence.

Premièrement, nous nous sommes concentrés sur la sous-collection "journaux" et "Belgique" du corpus SoNaR-500, représentant environ 200 millions de mots de textes journalistiques belges. Cette sélection est justifiée par plusieurs considérations. D'une part, les textes journalistiques sont reconnus comme une source fiable et représentative du néerlandais écrit standard (Biber et Conrad, 2009). D'autre part, ce type de discours correspond aux documents que les apprenants du néerlandais rencontrent fréquemment au cours de leur apprentissage.

Il est important de noter que, bien que la fiction représente également un domaine linguistique pertinent pour l'apprentissage des langues, son intégration dans cette analyse n'a pas été possible. Cette limitation est due à l'absence de listes de fréquences de mots issues de corpus de fiction facilement accessibles, téléchargeables ou manipulables dans le cadre de notre étude. Nous reconnaissons que l'inclusion de données issues de la fiction aurait enrichi la portée de notre analyse en offrant une perspective plus complète sur l'usage lexical. L'intégration de corpus de fiction pour le calcul de fréquences lexicales serait à considérer pour de futures recherches afin d'affiner la représentativité des données de fréquence pour les apprenants.

Quant au choix de la variante belge du néerlandais, celui-ci présente un intérêt particulier pour notre étude. En effet, le néerlandais de Belgique, tout en conservant les structures fondamentales de la langue standard, présente certaines spécificités lexicales qui peuvent faciliter la compréhension des locuteurs francophones (Geeraerts, 2001, p. 234).

Deuxièmement, nous avons limité notre analyse aux 50 000 lemmes les plus fréquents de cette sous-collection. Cette limitation se justifie par la loi de Zipf (1949), selon laquelle la distribution des fréquences lexicales suit une loi de puissance où un nombre relativement restreint de mots couvre la majorité des occurrences textuelles. Cette approche méthodologique est d'ailleurs corroborée par les recherches en linguistiques appliquée, notamment par les travaux de Nation (2006) et Schmitt (2008) sur l'acquisition du vocabulaire. Leurs études démontrent qu'un noyau lexical bien défini, même s'il est quantitativement limité, est suffisant pour atteindre une couverture textuelle très élevée. Cette approche présente également l'avantage d'éliminer les mots très rares qui pourraient introduire du bruit dans nos mesures de fréquence. Les mots au-delà des 50 000 premiers pourraient inclure des néologismes éphémères, ou des termes techniques très spécialisés qui ne sont pas pertinents pour l'évaluation de la lisibilité en langue étrangère. En choisissant un seuil de 50 000 lemmes, nous nous assurons donc de capturer de manière exhaustive les phénomènes lexicaux les plus pertinents et représentatifs, tout en optimisant l'efficacité de notre analyse computationnelle.

3.2.2 Variables basées sur les listes de référence

Justification théorique

L'approche par liste de référence permet d'évaluer dans quelle mesure le vocabulaire d'un texte correspond au vocabulaire supposé connu par des apprenants d'un niveau donné. Cette approche est particulièrement pertinente pour la L2, car elle permet d'aligner l'évaluation de la difficulté sur les progressions pédagogiques réelles.

Pour évaluer la difficulté lexicale, nous nous inspirons de l'approche multi-niveaux utilisée par François (2011, p. 298). Plutôt que d'utiliser notre liste de référence comme un bloc, nous la segmentons en plusieurs sous-listes de tailles croissantes. Cette démarche fait suites aux remarques de Dale et Chall (1948) et Harris et Jacobson (1974), qui font remarquer la relation entre le pouvoir discriminant des variables basées sur le nombre d'absents et la taille de la liste utilisée.

Ces sous-listes seront générées en sélectionnant les N premiers mots de notre liste de référence principale, qui est ordonnée par fréquence d'usage décroissante.

En créant des variables basées sur des seuils différents, nous pouvons modéliser la difficulté lexicale avec une granularité plus fine. Chaque sous-liste agit comme un "filet" différent, nous permettant de capturer des niveaux de complexité variés et d'obtenir des indicateurs distincts et complémentaires pour un même texte.

En bref, la division de notre listes en plusieurs sous-listes est une démarche méthodologique visant à créer un ensemble de prédicteurs lexicaux plus riches et plus nuancés que ne le permettrait une seule liste de référence.

Variables retenues

- **PA_SUBTLEX_1500, PA_SUBTLEX_3000, PA_SUBTLEX_MAX** (Proportion d'absents des sous-listes et de la liste complète SUBTLEX) : Ces variables mesurent le pourcentage de mots d'un texte qui sont absents de différentes listes de référence. Ces listes ne sont pas constituées de mots aléatoires, mais bien des 1500, 3000 et 4621 mots les plus fréquents de la langue néerlandaise, classés par ordre de fréquence décroissant.
- **PA_SUBTLEX_1500_U, PA_SUBTLEX_3000_U, PA_SUBTLEX_MAX_U** (Versions "uniques") : Ces variables comptent chaque lemme absent une seule fois, indépendamment de sa fréquence dans le texte. Elles capturent ainsi la diversité du vocabulaire difficile, plutôt que sa densité.

Adaptation au néerlandais et limites

Nous utilisons ici une liste provenant de Wiktionary (Wiktionary contributors, n.d.) contenant 4621 mots les plus utilisés d'après OpenSubtitles (OpenSubtitles.org contributors, n.d.). Ce corpus est une vaste base de données de sous-titres de films et séries télévisées. Nous avons récupéré cette liste et l'avons convertie au format CSV. Un nettoyage a ensuite été effectué. Les entrées contenant des chiffres ont été supprimées, celles-ci ne correspondant pas à des unités lexicales pertinentes mais à des dates, codes ou abréviations techniques. Leur maintien aurait introduit du bruit et biaisé la représentativité linguistique du corpus, notamment en raison de la nature spécifique et contextuelle des occurrences chiffrées dans les sous-titres.

Le choix d'une approche fondée sur des listes de fréquence pour évaluer la difficulté lexicale implique une simplification méthodologique dont nous reconnaissons les limites inhérentes. Cette méthode, par sa nature même, ne permet pas de capturer toutes les nuances de la compréhension lexicale. Deux cas illustrent particulièrement ces limites. Premièrement, le traitement de mots composés (*autodeur* par exemple) : un tel mot peut être absent des listes de fréquence et donc classé comme "difficile", alors même que ses composants (*auto* et *deur*) sont très fréquents et que son sens est sémantiquement transparent pour le lecteur. Deuxièmement, cette approche n'évalue pas directement l'effet facilitateur des cognats (*portier* par exemple), qui, bien qu'éventuellement peu fréquents, sont immédiatement transparents pour un public francophone.

Néanmoins, le recours à cette méthodologie demeure justifié au regard des objectifs de notre étude. Conformément aux travaux fondateurs en lisibilité (Dale et Chall, 1948) et aux synthèses plus récentes (François, 2011 ; Kleijn, 2018), l'analyse fréquentielle constitue une première approximation robuste et objective de la difficulté textuelle. Elle offre un cadre d'analyse systématique et reproductible, essentiel pour établir une base de comparaison fiable entre les textes. En choisissant de ne pas intégrer ces phénomènes complexes (transparence morphologique, proximité inter-langues) à ce stade, nous faisons le choix délibéré de privilégier une mesure standardisée et contrôlée.

3.2.3 Variables de diversité lexicale

Justification théorique

La diversité lexicale d'un texte est un facteur déterminant de la charge cognitive imposée au lecteur. Un texte qui présente une faible diversité, et donc une forte répétition des mêmes unités lexicales, facilite la reconnaissance des mots par un effet d'amorçage, réduisant ainsi l'effort de décodage (François, 2011, p. 233). A l'inverse, un texte lexicalement diversifié, qui introduit constamment de nouveaux termes, augmente la charge cognitive en requérant un effort lexical plus soutenu de la part du lecteur (McCarthy et Jarvis, 2010, p. 381). Mesurer cette dimension est donc essentiel pour évaluer la difficulté d'un texte.

Variables retenues

Pour quantifier la diversité lexicale, le Type-Token Ratio (TTR) est la mesure la plus traditionnellement utilisée. Cependant, sa sensibilité à la longueur du texte est une limite méthodologique bien connue : sa valeur diminue mécaniquement à mesure que la longueur du texte augmente, ce qui rend les comparaisons entre des textes de tailles différentes peu fiables.

C'est pourquoi nous avons retenu la variable suivante, qui a été spécifiquement conçue pour surmonter les faiblesses du TTR :

— **MTLD** (Measure of Textual Diversity) : En calculant le TTR moyen sur des segments de textes successifs jusqu'à ce qu'un seuil de variance soit atteint, le MTLD fournit un indice de diversité lexicale qui est intrinsèquement robuste à la variation de la longueur du texte (McCarthy et Jarvis, 2010, p. 386). Cette robustesse garantit une évaluation juste et comparable de la diversité lexicale, que les textes analysés soient courts où longs.

3.2.4 Variables de concrétude

Justification théorique

Les concepts concrets sont généralement plus faciles à comprendre et à mémoriser que les concepts abstraits, car ils peuvent être reliés à des expériences sensorielles directes. Cette idée, fondamentale en psychologie cognitive, est souvent associée à la théorie du double codage (Paivio, 1971). Selon cette théorie, les mots concrets

bénéficient d'un traitement privilégié car ils peuvent être représentés à la fois dans un système verbal et dans un système imaginal (sous forme d'images mentales), alors que les mots abstraits ne disposent que d'un codage verbal. Les concepts concrets profitent ainsi d'un double accès au sens, ce qui facilite leur traitement et leur mémorisation.

Comme le soulignent Paivio et al. (1988), ce double codage implique l'existence de trois systèmes de représentation interconnectés : un système verbal pour chacune des deux langues (L1 et L2) et un système imaginal unique. Les mots concrets peuvent dès lors être traduits soit directement, par un lien entre les deux systèmes verbaux, soit indirectement via le système imaginal (Hell et Groot, 1998, p. 42–43). Ce chemin additionnel explique leur traitement plus aisé, ce qui constitue une étape cruciale pour les lecteurs en L2, qui ne peuvent pas toujours s'appuyer sur la même masse de connaissances linguistiques et culturelles qu'un lecteur natif (François, 2011, p. 264-265).

La concrétude d'un texte, c'est-à-dire la proportion de mots concrets qu'il contient, influe donc directement sur sa lisibilité. Un texte riche en concepts abstraits impose une charge cognitive plus importante, particulièrement pour un lecteur en langue seconde, pour qui l'accès au sens des mots est moins automatisé et qui ne peut pas toujours compenser par des connaissances linguistiques ou culturelles préexistantes.

Variables retenues

Pour évaluer la concrétude de nos textes, la variable suivant est utilisée

— **Meanconc** (Score moyen de concrétude) : Cette mesure évalue la proportion de noms concrets dans un texte. Un score élevé indique un texte concret, tandis qu'un score faible signale un texte plus abstrait, potentiellement plus difficile à comprendre pour les apprenants.

Adaptation au néerlandais

La mesure présentée ci-dessus s'appuie sur les scores de concrétudes développés par Brysbaert et al. (2014) pour près de 30 000 mots néerlandais. L'utilisation de ces scores est particulièrement pertinente pour notre étude car ces scores ont été spécifiquement développé pour le néerlandais.Ces évaluations ont été collectées auprès de locuteurs natifs néerlandophones qui ont évalué chaque mot sur une échelle de 1 à 5, où 1 correspond à des concepts abstraits (basés sur le langage) et 5 à des concepts concrets (basés sur l'expérience). Les participants devaient juger dans quelle mesure le sens d'un mot pouvait être acquis par l'expérience directe à travers les sens (vue, ouïe, toucher, goût, odorat) et les actions, par opposition à une acquisition purement linguistique. Cette approche méthodologique, inspirée de la théorie du double codage de Paivio, permet de distinguer les mots selon leur degré d'ancrage perceptuel et moteur, offrant ainsi une mesure fiable de la complexité conceptuelle pour l'évaluation de la lisibilité en néerlandais L2.

3.3 Les variables syntaxiques

3.3.1 Fondement théorique de la complexité syntaxique

La syntaxe, souvent décrite comme la "grammaire" d'une langue, est le système de règles qui gouverne la manière dont les mots se combinent pour former des unités de sens plus larges, comme des syntagmes et des phrases. Elle ne se limite pas à l'ordre des mots, mais englobe également les relations hiérarchiques et fonctionnelles entre eux.

L'importance de la syntaxe dans l'apprentissage d'une langue étrangère réside dans le fait qu'elle est le squelette sur lequel le sens est construit. Elle joue donc un rôle très important dans la compréhension textuelle, tout particulièrement en L2 où les structures peuvent différer significativement de celles de la L1.

Pour les apprenants francophones du néerlandais, cette problématique est particulièrement aiguë en raison des différences d'ordre des mots entre les deux langues. Comme l'a montré Hulstijn (2001, p. 258), les différences dans l'ordre des mots posent des difficultés spécifiques. Celles-ci vont au-delà d'une simple opposition entre l'ordre SVO du français et les structures de surface du néerlandais.

En effet, la syntaxe du néerlandais engendre deux difficultés majeures pour les apprenants francophones :

- **L'ordre V2 dans la proposition principale :** Il ne s'agit pas d'un ordre SVO strict. Le verbe conjugué est placé en deuxième position, mais le premier constituant peut être un sujet, un complément ou un adverbe, forçant une inversion du sujet qui est cognitivement coûteuse pour un locuteur natif d'une langue SVO.
- **Le groupe verbal en fin de subordonnée :** L'ordre SOV apparent des subordonnées est la manifestation la plus visible de cette structure à verbe final. L'apprenant doit maintenir l'ensemble des compléments en mémoire de travail avant de rencontrer le groupe verbal qui leur donne un sens, ce qui, conformément à la théorie de la localité de Gibson (1998), augmente significativement la charge cognitive.

3.3.2 Variables de longueur des phrases

Justification théorique

La longueur des phrases est un prédicteur robuste de la complexité syntaxique car elle corrèle généralement avec le nombre de propositions et la complexité structurelle. La longueur des phrase est considérée comme "la meilleure mesure unique de la complexité grammaticale, et est incorporée dans la plupart des formules de lisibilité existantes" (Bormuth, 1966, p. 92) (notre traduction).

Selon Smith (1961) et Kemper et al. (1993, p. 418), une augmentation de la longueur des phrases est généralement due à l'inclusion de plusieurs propositions subordonnées, ce qui accroît la complexité syntaxique de la phrase.

Bien que Bormuth (1966, p. 122) ait suggéré une relative indépendance entre

la longueur des phrases et la complexité syntaxique, Chall et Dale (1995, p. 5) indiquent que "la longueur des phrases reste tout à fait valide comme prédicteur de la complexité syntaxique, et même davantage que des mesures plus compliquées reposant sur des théories linguistiques sophistiquées".

De plus, les travaux de Just et Carpenter (1980, p. 342) ont montré que les phrases longues imposent une charge plus importante sur la mémoire de travail, ce qui peut entraver la compréhension, notamment chez les lecteurs apprenants. En effet, si le nombre de mots excède les capacités limitées de la mémoire de travail, le lecteur ne serait pas capable de mémoriser une phrase dans son ensemble, et, si l'on accepte qu'il existe une association entre taux de mémorisation et compréhension, il faut dès lors en conclure qu'une phrase plus longue serait alors plus difficile à comprendre (Labasse, 1999, p. 90). Cette hypothèse est également soutenue par Graesser et al. (2004, p. 194)

Variables retenues

— **MeanSentL** (Longueur moyenne des phrases en mots) : Cette mesure capture la complexité syntaxique globale du texte. Elle est pertinente notamment dans le contexte de l'apprentissage du néerlandais, car les structures typiques de cette langue favorisent la création de dépendances à longue distance qui constituent une source de difficulté.

Limites

Bien que la longueur des phrases soit un indicateur utile, elle est critiquée pour ne pas apporter une information propre sur la complexité syntaxique, agissant plutôt comme un intermédiaire. Un défaut majeur est qu'elle ne tient pas compte de l'ordre des mots ou de la structure grammaticale interne des phrases, ce qui peut conduire à des scores de lisibilité identiques pour des textes de complexité structurelle différente (François, 2011, p. 248). Pour dépasser cette limite, il est suggéré de combiner cette mesure avec des analyses plus fines des structures syntaxiques.

3.3.3 Variables de complexité syntaxique

Justification théorique

Au-delà de simplement la longueur, la complexité syntaxique dépend de la structure hiérarchique des phrases. Les propositions subordonnées créent des niveaux d'imbrication qui augmentent la charge cognitive. Selon Labasse (1999, p. 89), cité par François (2011, p. 247) : "la compréhension d'une phrase impliquerait la reconstitution de sa structure syntaxique : plus la phrase sera longue, donc compliquée, plus il sera difficile d'en reconstruire l'arborescence grammaticale".

Cette perspective est renforcée par les travaux sur la lisibilité qui soulignent l'importance de la complexité structurelle des phrases. François (2011, p. 246) note que les mesures de la longueur des phrases, bien que pratiques, sont souvent considérées comme un indice de la complexité syntaxique sous-jacente. L'augmentation de la

longueur des phrases s'effectue généralement via l'inclusion de plusieurs propositions subordonnées, ce qui accroît la complexité syntaxique de la phrase.

Kleijn (2018, pp. 81-85) dans ses travaux sur la lisibilité du néerlandais met en évidence comme les dépendances syntaxiques, en particulier les dépendances non-locales affectent le traitement et la compréhension du texte. La difficulté de traitement des phrases complexes est liée à la charge cognitive imposée par la nécessité de maintenir en mémoire de travail les éléments syntaxiques en attente de leur résolution, comme c'est le cas avec les structures à verbe final, par exemple.

Variables retenues

— **NbSubord** (Nombre de propositions subordonnées moyen par texte) : Cette variable capture une source majeure de complexité syntaxique. Les propositions subordonnées en néerlandais sont la manifestation la plus claire de la structure fondamentale à verbe final (SOV) de la langue. En effet, en néerlandais, les verbes forment un groupe verbal dans les propositions à verbe final. Comme l'a montré Perrez (2006, p. 189), cette structure où le groupe verbal est rejeté en fin de proposition, est particulièrement difficile pour les apprenants francophones habitués à un ordre SVO plus direct. Le lecteur est contraint de maintenir en mémoire de travail l'ensemble des compléments avant de rencontrer le verbe, ce qui augmente la charge cognitive.

3.3.4 Variables spécifiques au néerlandais

Justification théorique

Les difficultés majeures rencontrées par les apprenants francophones du néerlandais relèvent d'un phénomène syntaxique fondamental : la structure SOV, modulée par la contrainte V2. Cette organisation syntaxique, caractéristique du néerlandais, entraîne des effets structuraux qui n'existent pas en français, notamment la position du verbe en fin de proposition subordonnée et la séparation des particules dans les verbes à particule séparable.

Dans les propositions principales néerlandaises, la contrainte V2 impose que le verbe conjugué occupe systématiquement la deuxième position, ce qui a pour conséquence la séparation de la particule du verbe dans les constructions à particule séparable ("*Ik bel je op*" pour "*opbellen*") (Perrez, 2006, p. 234). En revanche, dans les subordonnées, la structure SOV impose le placement du verbe (ou de l'élément verbal) en fin de proposition, ce qui contraste fortement avec l'ordre SVO du français et constitue une source d'erreurs fréquente chez les apprenants francophones.

Ces deux manifestations syntaxiques ne doivent pas être considérées isolément, mais comme des symptômes d'une même difficulté structurelle propre au néerlandais.

Variables retenues

Afin de quantifier et d'illustrer cette difficulté, deux variables spécifiques ont été conceptualisées et retenues. Leur sélection est directement motivée par les observa-

tions issues de la littérature scientifique et les défis linguistiques identifiés. Ces variables visent à capturer les manifestations concrètes des phénomènes de la contrainte V2 et de l'ordre SOV.

— **PropSepVerb** (Proportion de verbes à particule séparable) : Cette variable correspond à la proportion de verbes à particule séparable présents dans un texte, par rapport au nombre total de verbe, mesurant ainsi la fréquence des constructions où la particule verbale est détachée du verbe conjugué. L'idée de cette variable découle directement de l'analyse de la contrainte V2 en néerlandais. Ce phénomène, où le verbe conjugué occupe la deuxième position et force la particule à se déplacer en fin de proposition principale, constitue une difficulté syntaxique majeure et spécifique au néerlandais L2. L'implémentation de cette variable est justifiée par le fait qu'elle représente une divergence structurelle fondamentale avec le français, et qu'elle n'est généralement pas prise en compte par les modèles génériques de lisibilité. Sa quantification permet d'évaluer la densité de cette structure dans un texte. Cette mesure représente une lacune dans la littérature existante et souligne l'importance de notre contribution à la recherche en lisibilité.
— **PropSOV** (Proportion de propositions avec ordre SOV) : Cette variable a été élaborée pour quantifier la proportion de phrase contenant au moins une proposition subordonnée. L'inspiration pour cette variable provient de la reconnaissance de l'ordre SOV comme structure fondamentale des propositions subordonnées en néerlandais, un point de contraste majeur avec le français. L'implémentation de cette variable est justifiée par sa capacité à identifier les textes riches en subordonnées ou en constructions verbales complexes, qui sont des sources avérées de difficultés pour les apprenants francophones en raison de la charge cognitive accrue liée à la réorganisation de l'information verbale. Cette variable, en comparaison avec la variable **NbSubord** permet de normaliser les données, et donc de comparer la densité des structures entre des textes de longueurs différentes.

Ces deux variables, bien que distinctes, sont complémentaires et représentent des indices ciblés d'une même contrainte syntaxique sous-jacente au néerlandais. Leur combinaison permet une évaluation plus nuancée de la difficulté textuelle, en se focalisant sur les aspects les plus problématiques pour les apprenants francophones du néerlandais.

3.3.5 Variables de connecteurs

Justification théorique

Les connecteurs, en tant qu'explicitateurs des relations logiques entre les propositions, sont fondamentaux pour assurer la cohérence textuelle. Leur maîtrise représente cependant un défi majeur dans le cadre de l'acquisition des langues secondes (L2). La recherche a mis en évidence une hiérarchie de difficulté dans l'acquisition de ces marqueurs, y compris chez les locuteurs natifs. Il est généralement admis que les apprenants acquièrent les connecteurs selon un ordre spécifique : les relations additives et temporelles sont maîtrisées plus précocement que les relations causales

et contrastives/concessives (Wetzel et al., 2020). Cette hiérarchie s'explique par la complexité cognitive sous-jacente : les connecteurs causaux et, plus encore, concessifs (une forme de contraste), encodent des relations sémantiques et pragmatiques plus abstraites, qui exigent du lecteur une charge inférentielle plus élevée. La présence et le type de connecteurs dans un texte constituent par conséquent des indicateurs pertinents de sa difficulté syntaxique et sémantique.

Au-delà de cette hiérarchie générale, plusieurs facteurs théoriques spécifiques influencent la difficulté d'acquisition des connecteurs pour les apprenants francophones du néerlandais. Premièrement, l'absence d'équivalents directs en langue maternelle constitue un obstacle majeur, forçant les apprenants à développer de nouvelles catégories conceptuelles plutôt que de simplement transférer des connaissances de leur L1 (Perrez, 2006, pp. 176–177). Deuxièmement, la transparence sémantique réduite de certains connecteurs nécessite un apprentissage explicite plutôt qu'une acquisition intuitive (Wetzel et al., 2020). Troisièmement, les nuances pragmatiques subtiles, notamment les distinctions de volition et de subjectivité, représentent des défis conceptuels particuliers (Perrez, 2006, pp. 111–124). La présence et le type de connecteurs dans un texte constituent par conséquent des indicateurs pertinents de sa difficulté syntaxique et sémantique.

Variables retenues

En nous basant sur cette hiérarchie de difficulté, nous avons choisi de nous concentrer sur les deux catégories de connecteurs qui posent le plus de problèmes aux apprenants, et qui sont donc les plus discriminantes pour évaluer la complexité des textes. Plutôt que d'inclure tous les connecteurs causaux et contrastifs, notre sélection se fonde sur des critères théoriques précis : la complexité conceptuelle, l'absence d'équivalents directs en français, la fréquence d'usage discriminante, et la transparence sémantique réduite.

— **RatioConnCaus** (Ratio de connecteurs causaux) : Cette variable mesure la proportion de connecteurs causaux spécifiquement sélectionnés pour leur complexité d'acquisition par raport au nombre total de mots du texte. Nous nous concentrons sur trois connecteurs particulièrement problématiques pour les apprenants francophones :

1. *Doordat* : Ce connecteur présente une spécificité sémantique cruciale car il est "quasi-exclusivement utilisé dans des contextes non-volitionnels" où le sujet n'a pas de contrôle sur l'action (Perrez, 2006, p. 124). Cette nuance pragmatique subtile échappe systématiquement aux apprenants francophones qui n'ont pas d'équivalent direct en français. Perrez (2006, p. 177) documente le fait que ce connecteur est sous-utilisé, probablement pour cette raison.
2. *Daardoor* : Similaire à *doordat*, ce connecteur encode également des relations non-volitionnelles et présente un sous-emploi notable par les apprenants (Perrez, 2006, p. 178). Son acquisition tardive en fait un excellent indicateur de maîtrise avancée.
3. *Aangezien* : Ce connecteur code un degré d'implication du locuteur (subjectivité) intermédiaire entre *omdat* (plus objectif) et *want* (plus subjectif)

(Degand et Pander Maat, 2003). Sa maîtrise dépend de nuances pragmatiques complexes que les apprenants acquièrent tardivement (Perrez, 2006, pp. 124-125).

Nous excluons délibérément des connecteurs comme *omdat* et *want*, bien qu'ils présentent certaines difficultés, car leur fréquence élevée et leur acquisition relativement précoce en font des indicateurs moins discriminants de sophistication textuelle (Perrez, 2006, pp. 175–176).

— **RatioConnContr** (Ratio de connecteurs contrastifs) : Cette variable mesure la proportion de connecteurs de contraste ou de concession sélectionnés pour leur complexité cognitive maximale. Notre sélection se concentre sur quatre connecteurs représentant les défis les plus importants :

1. *Echter* : Perrez (2006, p. 198) documente un sous-emploi massif de *echter* par les apprenants, qui le remplacent souvent par *maar*. Sa complexité réside dans son registre plus formel.
2. *Hoewel* : Ce connecteur concessif marque des relations où une attente est contrariée. Perrez (2006, p. 140) note que ces connecteurs sont également difficiles à maitriser pour les apprenants et que leur acquisition est tardive.
3. *Ook al* : Ce connecteur marque une concession encore plus forte que *hoewel* (Perrez, 2006, p. 141) et est particulièrement intéressant car il est encore moins fréquent.
4. *Desondanks* : Très peu utilisé par les apprenants, ce connecteur marque une attente fortement contredite et représente un niveau de sophistication pragmatique élevé (Perrez, 2006, p. 143).

Nous excluons *maar*, connecteur de base acquis précocement et sur-utilisé par les apprenants, ainsi que *toch* dont la polysémie importante rend l'analysee ambiguë (Perrez, 2006, pp. 194–199)

Limites

Bien que les ratios de connecteurs causaux et contrastifs soient des indicateurs précieux de la difficulté textuelle pour les apprenants en L2, il est important de reconnaitre certaines limites. La simple présence ou proportion de ces connecteurs ne capture pas toujours la subtilité de leur usage ou la complexité des relations qu'ils encodent. Par exemple, la polyfonctionnalité de certains connecteurs peut rendre leur classification ambigüe, et leur impact sur la compréhension peut varier en fonction du contexte discursif et des connaissances préalable du lecteur (Perrez, 2006, pp. 97-98).

De plus, notre approche sélective, bien que théoriquement motivée, présente le risque de sous-estimer la complexité de textes utilisant d'autres connecteurs sophistiqués non inclus dans notre sélection. Wetzel et al. (2020, p. 2) soulignent également que la maîtrise des connecteurs représente une étape importante vers l'acquisition d'un haut niveau de compétence linguistique, mais que celle-ci reste difficile pour les apprenants L2 même aux niveaux avancés, suggérant que nos variables pourraient nécessiter des ajustements selon les populations d'apprenants étudiées.

Enfin, comme le note Perrez (2006, p. 182), l'usage des connecteurs peut refléter des stratégies d'apprentissage spécifiques plutôt que la maîtrise linguistique pure, ce qui nécessite une interprétation prudente des résultats obtenus avec ces variables.

3.4 Synthèse des variables utilisées pour notre modèle de lisibilité

Nous avons donc au total 18 variables différentes classées en deux grandes catégories : lexicales et syntaxiques. Nous proposons ici d'en faire un résumé clair.

Tableau 3.1 – Synthèse des variables retenues

Type	**Sous-type**	**Variable**	**Mesure**
Lexicale	Fréquence lexicale	mean_fl	Fréquence moyenne des lemmes
Lexicale	Fréquence lexicale	median_fl	Médiane des fréquences
Lexicale	Fréquence lexicale	75FL	75^{e} percentile des fréquences
Lexicale	Fréquence lexicale	90FL	90^{e} percentile des fréquences
Lexicale	Proportion d'absents	PA_SUBTLEX_1500	Proportion d'absents de la sous-liste de fréquence de 1500 mots
Lexicale	Proportion d'absents	PA_SUBTLEX_3000	Proportion d'absents de la sous-liste de fréquence de 3000 mots
Lexicale	Proportion d'absents	PA_SUBTLEX_MAX	Proportion d'absents de la liste de fréquence complète
Lexicale	Proportion d'absents	PA_SUBLTEX_1500_U	Proportion d'absents uniques de la sous-liste de fréquence de 1500 mots
Lexicale	Proportion d'absents	PA_SUBLTEX_3000_U	Proportion d'absents uniques de la sous-liste de fréquence de 3000 mots
Lexicale	Proportion d'absents	PA_SUBLTEX_MAX_U	Proportion d'absents uniques de la liste de fréquence complète
Lexicale	Diversité lexicale	MTLD	Mesure de la diversité lexicale (alternative au TTR)
Lexicale	Concrétude	MeanConc	Moyenne des scores de concrétude des mots
Syntaxique	Longueur de phrases	MeanSentL	Longueur moyenne des phrases (en mots)

Tableau 3.1 – Synthèse des variables retenues (suite)

Type	Sous-type	Variable	Mesure
Syntaxique	Complexité hiérarchique	NbSubord	Nombre de subordonnées par texte
Syntaxique	Spécificités NL	PropSepVerb	Proportion de verbes à particule séparable
Syntaxique	Spécificités NL	PropSOV	Proportion de phrase contenant au moins une proposition subordonnée (ordre SOV)
Syntaxique	Connecteurs logiques	RatioConnCaus	Proportion de connecteurs causaux
Syntaxique	Connecteurs logiques	RatioConnContr	Proportion de connecteurs contrastifs

3.5 Pistes d'amélioration

Afin de faire preuve de véritable rigeur scientifique, nous présentons ici plusieurs pistes d'améliorations liées aux limites méthodologiques de notre étude, afin d'ouvrir la voie aaux futures recherches.

3.5.1 Intégration de variables basées sur les N-grammes

Les N-grammes, séquences contiguës de N éléments (mots ou caractères), sont des outils puissants en traitement automatique du langage pour capturer des informations contextuelles et des dépendances linguistiques qui dépassent le niveau du mot isolé. L'intégration de variables basées sur les N-grammes aurait pu enrichir notre étude en permettant de mieux appréhender la complexité syntaxique et sémantique des textes. Par exemple, des N-grammes de mots pourraient révéler des collocations ou des expressions idiomatiques difficiles pour les apprenants, tandis que des N-grammes de parties du discours (POS N-grammes) pourraient offrir une mesure plus fine de la complexité structurelle des phrases (Razon et Barnden, 2015).

Cependant, l'inclusion de telles variables a été limitée par des contraintes computationnelles. Bien que l'interface N-grams d'OpenSonar permette d'extraire des modèles n-grammes de taille 2 à 5, nous n'avons pas été en mesure d'implémenter ces variables, faute de puissance de traitement suffisante. Cette limitation technique a empêché une exploration exhaustive de ces variables dans le cadre de cette recherche, mais représente une avenue prometteuse pour des études futures disposant de ressources informatiques plus importantes.

3.5.2 Variables spécifiques à l'interférence linguistique : faux-amis et cognats

L'apprentissage d'une langue étrangère est intrinsèquement lié aux phénomènes d'interférence linguistique, où la langue maternelle (L1) de l'apprenant influence l'acquisition de la langue seconde (L2). Dans le contexte néerlandais-français, les faux-amis (mots de

langues différentes ayant une forme similaire mais un sens distinct) et les cognats (mots de langues différentes ayant une forme et un sens similaires) jouent un rôle crucial dans la compréhension et la production (Brenders et al., 2011). La proportion de faux-amis ou de cognats dans un texte pourrait être un indicateur pertinent de sa difficulté ou de sa facilité pour un apprenant francophone du néerlandais.

Malgré leur pertinence théorique, l'intégration de variables spécifiques aux faux-amis et cognats néerlandais-français n'a pas été possible dans cette étude. Nous n'avons pas été en mesure de trouver des listes exhaustives et validées de ces paires lexicales. La création amnuelle de telles ressources aurait dépassé le cadre temporel de ce mémoire. Il serait intéressant que, dans le futur, des recherches se concentrent sur la compilation de ces listes et le développement de méthodes robustes pour leur extraction automatique.

3.5.3 Analyse de la transparence des mots composés

Comme l'ont démontré les travaux de Zwitserlood (1994), la compréhension des mots composés néerlandais dépend fortement de leur degré de transparence sémantique. Les mots composés transparents, tels que "*schoolboek*" (livre d'école), permettent aux apprenants d'inférer le sens global à partir des constituants. Cette notion de transparence pourrait constituer un prédicteur pertinent de la difficulté textuelle, complémentaire aux mesures de fréquence lexicale traditionnelles. De cette manière, les mots composés transparents ne seraient pas considéré systématiquement comme "difficile" par les modèles de lisibilité.

L'implémentation de telles variables nécessiterait cependant des ressources linguistiques spécialisées que nous n'avons pas pu mobiliser dans le cadre de cette étude. Il faudrait notamment disposer d'un analyseur morphologique capable de segmenter automatiquement les mots composés néerlandais et d'une base de données annotée pour la transparence sémantique. De plus, l'évaluation de la transparence implique souvent des jugements subjectifs qui nécessiteraient une validation empirique approfondie.

3.6 Conclusion

Dans ce chapitre, nous avons identifié et décrit un large éventail de variables linguistiques potentiellement pertinentes pour évaluer la lisibilité du néerlandais langue étrangère. Ces variables couvrent des aspects lexicaux et syntaxique du néerlandais. Pour chaque catégorie, nous avons sélectionné les variables les plus prometteuses en fonction de leur pertinence théorique, de leur faisabilité technique et de la disponibilité des ressources nécessaires.

Dans le chapitre suivant, nous présenterons le corpus utilisé pour entraîner notre modèle.

Chapitre 4

Le corpus d'entraînement

4.1 Introduction

Après avoir identifié et paramétré les variables linguistiques pertinentes pour évaluer la lisibilité du néerlandais langue étrangère, l'étape suivante de notre démarche consiste à constituer le corpus qui servira de fondement à nos analyses. Cette étape est fondamentale dans toute étude empirique : la qualité des données détermine directement la fiabilité des conclusions. Le corpus joue ici un rôle central pour deux raisons :

1. Il permettra, dans un premier temps, d'analyser la distribution des variables linguistiques que nous avons sélectionnées, afin de vérifier empiriquement leur capacité à discriminer les différents niveaux de difficulté du Cadre Européen Commun de Référence pour les Langues (CECR).
2. Il constitue, dans un second temps, la base sur laquelle nous testerons, via des analyses statistiques, le potentiel discriminant de ces variables dans un contexte contrôlé.

Parce qu'il servira de référence pour ces analyses, ce corpus doit être soigneusement construit : la représentativité des textes, la cohérence des niveaux attribués et la diversité des sources sont autant de facteurs essentiels pour garantir la pertinence des résultats.

Ce chapitre présente donc la constitution et la justification du corpus utilisé, en détaillant les choix de sélection des textes, les sources mobilisées, le critère de classification, ainsi que les difficultés rencontrées au cours du processus. L'objectif est de rendre transparente cette étape méthodologique afin de situer clairement les conditions et les limites des analyses menées dans le reste du travail.

4.2 Le choix du critère : comment mesurer la difficulté d'un texte ?

Avant de décrire la composition de notre corpus, il est essentiel de se pencher sur la question du critère : comment mesurer objectivement la difficulté d'un texte pour un groupe de lecteurs donné ? Le choix de ce critère est déterminant, car il conditionne la

qualité des annotations sur lesquelles notre modèle de lisibilité sera entraîné. La littérature scientifique a exploré plusieurs méthodes pour établir un tel "gold standard", chacune avec ses avantages et ses inconvénients.

Parmi les principaux critères envisagés, on peut citer :

- **Les tests de compréhension :** Il s'agit de soumettre des textes à un échantillon de lecteurs et de mesurer leur niveau de compréhension via des questionnaires (à choix multiples ou à réponses ouvertes). C'est une mesure directe de la compréhension, mais elle est très coûteuse à mettre en œuvre et la difficulté des questions peut elle-même biaiser l'évaluation de la difficulté du texte (François, 2011, p. 325)
- **Le test de closure :** Cette technique consiste à supprimer des mots à intervalle régulier dans un texte et à demander aux lecteurs de les restituer. Le score obtenu est considéré comme une mesure de la compréhensibilité globale du texte. Bien que plus simple à corriger qu'un test de compréhension, il reste lourd à déployer sur un grand nombre de textes et de sujets (François, 2011, p. 327).
- **La vitesse de lecture :** On mesure le temps que met un lecteur pour lire un texte. Une lecture plus lente est souvent associée à une plus grande difficulté. Cependant, cette méthode requiert du matériel spécifique (comme un oculomètre) et ne garantit pas que le lecteur a effectivement compris le texte (François, 2011, p. 332).
- **L'avis d'experts :** Cette approche consiste à faire appel au jugement de professionnels (enseignants, concepteurs de matériel pédagogique) pour évaluer et classer les textes. Ces experts par leur expérience, ont une connaissance approfondie des compétences des apprenants à un niveau donné (François, 2011, p. 323).

Dans le cadre de cette recherche, qui nécessite un corpus de grande taille pour entraîner efficacement un modèle d'apprentissage automatique, les trois premières options nous ont semblé impraticables en raison de leur coût élevé en temps et en ressources. C'est pourquoi nous avons opté pour un critère basé sur l'avis d'experts.

Ce choix s'est avéré à la fois pragmatique et pertinent. Pragmatiquement, il nous a peris de rassembler un volume conséquent de textes dont la difficulté était déjà annotée par des experts. Nous avons choisi uniquement des textes qui étaient annotés en fonction des compétences définies par les niveaux du Cadre euroéen commun de référence pour les langues (CECR). L'indication du niveau (A1, A2, B1, ...) sur un manuel représente ainsi un jugement expert sur la difficulté globale des textes qu'il contient. Il est important de noter que nous n'avons retenu que les textes clairement associés à un seul niveau CECR, excluant ainsi les manuels proposant des contenus couvrant plusieurs niveaux.

Ce choix présente également un intérêt dans la perspective de futures recherches visant à élaborer une formule de lisibilité. Le corpus que nous avons constitué pourra être mis à la disposition de la communauté scientifique, offrant ainsi une base de données directement exploitable pour développer un modèle capable d'aider enseignants et apprenants à sélectionner des textes adaptés à leur niveau. En s'appuyant sur des textes issus de manuels, notre approche garantit que ce travail constituera une ressource immédiatement pertinente et utile.

4.3 Constitution du corpus néerlandais

La construction de notre corpus de néerlandais s'appuie sur une approche privilégiant l'avis d'experts et se base donc sur l'utilisation de matériels dont le niveau de difficulté a déjà été évalué par des experts du domaine. Ce choix, comme nous l'avons justifié dans

la section précédente, est à la fois pragmatique et pertinent pour les objectifs de notre recherche. Il nous permet de rassembler un grand nombre de textes annotés tout en nous assurant que l'échelle de difficulté utilisée est directement liée au contexte de l'enseignement du néerlandais langue étrangère.

Pour garantir la cohérence et la pertinence de ce corpus, la sélection des sources a été guidée par un ensemble de critères stricts.

4.3.1 Critères de sélection des sources

Compatibilité avec le CECR

L'objectif principal de ce mémoire étant de faire progresser la recherche sur la lisibilité en néerlandais langue étrangère, en posant les bases scientifiques nécessaires au développement ultérieur d'outils opérationnels, le critère fondamental pour la sélection d'une source était son alignement explicite avec le Cadre Européen Commun de Référence pour les Langues (CECR). Adopter cette échelle garantit que les résultats obtenus seront directement exploitables dans de futures études visant à concevoir un outil pertinent et utilisable par les enseignants et les apprenants.

Authenticité et sources pour les niveaux avancés

La disponibilité des textes didactisés pour les niveaux très avancés (C1 et C2) est très limitée. A ce stade de l'apprentissage, les apprenants sont généralement capables de comprendre la majorité des écrits et sont donc souvent confrontés à des textes authentiques, c'est-à-dire non spécifiquement conçus pour un public d'apprenants. Pour refléter cette réalité, et pour enrichir notre corpus à ces niveaux, nous avons complété notre corpus en y intégrant des articles de presse provenant de journaux néerlandophones de qualité, tels que *De Tijd*.

Ces critères nous ont permis de sélectionner un ensemble de sources cohérentes, formant une base solide pour l'entraînement et la validation de notre modèle de lisibilité. La section suivante détaille les manuels et sources spécifiques qui ont été retenus.

4.3.2 Sources utilisées

Notre corpus a été principalement alimenté par les sources suivantes, comme détaillé dans le tableau fourni en annexe (voir pages 106 à 109).

- **Manuel "Néerlandais B2 - Vers une communication professionnelle"** (Bosmans et al., 2020) : Ce manuel, spécifiquement conçu pour le niveau B2, a servi de base pour les textes de ce niveau. La majorité des textes eux-mêmes proviennent d'articles de presse authentiques (De Standaard principalement, mais aussi De Morgen), sélectionnés pour correspondre aux objectifs de communication professionnelle du manuel. Cette combinaison en fait une source fiable et pertinente pour des apprenants avancés.
- **Lingua.com** (Lingua.com, 2025) : Cette plateforme en ligne propose des exercices et des textes pour l'apprentissage de diverses langues, dont le néerlandais. Une cinquantaine de textes sont répartis par niveau CECR sur la plateforme. Il s'agit principalement de textes descriptifs, ainsi que de quelques dialogues. Cette ressource nous a plutôt servi pour trouver des textes de niveau débutant à intermédiaire (de A1 à B1, avec quelques textes de niveau B2).

— **NT2 Taalmenu (Nederlands als Tweede Taal)** (NT2 Taalmenu, 2025) : Ce site est une platefore en ligne dédiée à l'apprentissage du néerlandais comme langue seconde. Il propose une large variété d'exercices, de textes et de ressources adaptés aux apprenants de niveaux A1 à B2 du CECR. Ces ressources sont conçues pour préparer aux examens officiels *Staatsexamen NT2*, en offrant différents textes accompagnés d'exercices de compréhension et de production. L'utilisation de ce site dans le cadre de ce mémoire est particulièrement pertinente, car il permet de disposer de textes variés et calibrés sur des niveaux précis, en phase avec les exigences officielles du NT2.
— **CNaVT (Certificaat Nederlands als Vreemde Taal)** (CNaVT, 2025) : Le CNavt est un certificat internationalement reconnu attestant de la maitrîse du néerlandais comme langue étrangère. Les exemples d'examens et les matériaux de préparation disponibles sur leur site fournissent des textes de niveaux variés (A2 à C1). L'utilisation de ces textes dans le cadre de ce mémoire est pertinente et sécurisée : d'une part parce qu'ils proviennent d'une institution officielle et suivent des critères d'évaluation standardidés et validés ; d'autre part, parce qu'ils offrent des données annotées en niveaux CECR, indispensables pour entraîner et tester notre modèle.
— **De Tijd** : Pour enrichir notre corpus aux niveaux C1 et C2, nous avons intégré des articles provenant du journal belge *De Tijd*. Ce journal de référence, reconnu pour la qualité de son néerlandais et la complexité de ses analyses, constitue une source authentique particulièrement adaptée aux apprenants de niveau avancé. Les articles sélectionnés couvrent des domaines variés (économie, politique, culture) et présentent un niveau de complexité lexicale et syntaxique correspondant aux compétences attendues aux niveaux C1-C2 du CECR.

4.3.3 Processus de collecte et de classification

Le processus de collecte a impliqué l'extraction des textes à partir de ces différentes sources. Pour les manuels, les textes ont été scannés ou transcrits. Pour les plateformes en ligne, les textes ont été directement récupérés. Tous les textes ont finalement été enregistrés au format ".txt". Il est donc important de noter que la mise en page ne pourra être analysée.

La classification des textes par niveau CECR (A1, A2, B1, B2, C1, C2) a été réalisée en se basant sur les indications fournies par les sources elles-mêmes. Excepté pour les niveaux C1, C2 pour lesquels nous avons utilisé un grand nombre de textes authentiques.

4.3.4 Difficultés rencontrées et solutions apportées

Les principales difficultés rencontrées lors de la constitution de ce corpus de néerlandais rejoignent celles identifiées par François pour le FLE (François, 2011) :
— **Hétérogénéité des manuels et de ressources en ligne** : Les critères de classification par niveau peuvent varier d'une source à l'autre. Certains manuels ou sites web peuvent regrouper plusieurs niveaux (ex : A1/A2) sans fournir de granularité suffisante pour une classification fine. Nous avons choisi d'écarter ces ressources pour nous concentrer sur celles indiquant clairement un seul niveau CECR.
— **Manque de textes pour certains niveaux** : Nous nous somme heurtés à un déséquilibre dans la disponibilité des textes pour certains niveaux, notamment les niveaux débutants (A1) ou très avancés (C1/C2). Pour les niveaux C1/C2, la rareté des textes didactisés rend la collecte plus complexe, nécessitant de se tourner vers des textes authentiques.

— **Accès aux ressources** : Nous n'avons trouvé qu'un nombre limité de ressources disponibles gratuitement en ligne, limitant ainsi l'accès à un corpus plus vaste. Le choix s'est donc porté sur des sources accessibles gratuitement (excepté pour Lingua.com (2025)).

4.3.5 Regroupement des niveaux C1 et C2 pour l'analyse

Dans le cadre de nos analyses, nous avons pris la décision de regrouper les niveaux C1 (utilisateurs expérimentés) et C2 (maîtrise) du CECR.

Tout d'abord, la distinction entre les niveaux C1 et C2 du cadre Européen Commun de Référence (CECR) est souvent plus une question de degré que de nature. Plusieurs études montrent que la principale différence entre les deux réside dans l'étendue du vocabulaire, notamment la maîtrise des idiomes et des expressions locales, ainsi que dans la capacité à lire et comprendre des textes complexes avec une aisance accrue (Milton et Alexiou, 2020). Pour le niveau C1, l'apprenant est capable de comprendre le contenu de textes longs et exigeants, y compris les significations implicites. Il peut s'exprimer spontanément et couramment sans avoir à chercher ses mots, et utiliser la langue de manière efficace et souple dans des contextes sociaux, professionnels ou académiques. Il peut également s'exprimer sur des sujets complexes de manière claire et structurée, en maîtrisant les outils d'organisation et de cohésion du discours. Le niveau C2, quant à lui, représente une maîtrise quasi bilingue. L'apprenant peut comprendre sans effort pratiquement tout ce qu'il lit ou entend, restituer des faits et des arguments de diverses sources écrites et orales en les résumant de manière cohérente. Il peut s'exprimer spontanément, très couramment et de manière précise, et rendre distinctes de fines nuances de sens en rapport avec des sujets complexes. La principale différence avec le C1 réside dans la capacité à puiser dans un vocabulaire plus large et à intégrer parfaitement les idiomes et les expressions dans la communication active (Council of Europe, 2001, pp. 61-62).

Ensuite, la disponibilité de textes didactisés spécifiquement étiquetés C1 ou C2 est très limitée. Les ressources pour ces niveaux tendent à être des textes authentiques (journaux, littérature, articles scientifiques) qui ne sont pas toujours conçus avec une progression pédagogique explicite. Regrouper C1 et C2 permet d'élargir le nombre de textes disponibles pour l'analyse, augmentant ainsi la robustesse statistique du corpus pour ces niveaux avancés. De plus, de cette manière, nous évitons de créer des catégories avec un nombre insuffisant d'échantillons, ce qui pourrait compromettre la validité des analyses ultérieures.

En somme, le regroupement des niveaux C1 et C2 dans ce corpus est une décision méthodologique justifiée par la disponibilité des ressources et les objectifs pratiques de l'étude de la lisibilité. Il permet de maintenir une granularité suffisante pour les analyses tout en assurant la viabilité de la collecte et de la classification des données.

Malgré ces défis, le corpus constitué offre une base solide pour l'analyse de la lisibilité en néerlandais. La transparence de sa constitution et la justification des choix méthodologiques, inspirées par le travail de François (2011), garantissent sa robustesse et sa pertinence pour les recherches futures.

4.4 Analyse critique du corpus : description et limitations

Dans cette section, nous analyserons de manière critique le corpus que nous avons constitué. L'objectif de cette analyse est double : d'une part fournir une description détaillée et objective de la composition du corpus ; d'autre part, identifier et discuter de manière critique les limitations méthodologiques qui pourraient affecter la validité des résultats obtenus.

4.4.1 Description quantitative du corpus

Dans cette section, nous présentons une vue d'ensemble de la composition du corus, en mettant en évidence sa répartition par niveaux CECR et la diversité des textes qui le constituent. l'objectif est de fournir un aperçu clair des données mobilisées avant leur exploitation statitique. Pour une description exhaustive, incluant la liste complète des textes et leurs caractéritiques détaillées (source, niveau, longueur, ...) nous renvoyons le lecteur au tableau récapitulatif fourni en annexe (pp. 106-109).

Composition générale

Notre corpus comprend 178 textes répartis selon les niveaux du CECR. Cette taille, bien que modeste comparée aux standards contemporains en traitement automatique du langage, s'inscrit dans la lignée des corpus de lisibilité traditionnels qui comptent généralement entre quelques dizaines et quelques centaines de textes (Crossley et al., 2023, p. 504).

La distribution des textes par niveau révèle un déséquilibre qui mérite une attention particulière.

Niveau CECR	Nombre de textes	Pourcentage
A1	22	12,4%
A2	39	21,9%
B1	40	22,5%
B2	41	23,0%
C1	6	3,4%
C2	30	16,9%

Tableau 4.1 – Répartition des textes par niveau CECR

Cette distribution révèle plusieurs problèmes méthodologiques importants. Premièrement, la sous-représentation du niveau C1 (seulement 6 textes, soit 3,4% du corpus) constitue une limitation majeure qui compromet la capacité du modèle à apprendre les caractéristiques linguistiques spécifiques à ce niveau. Cela a donc nécessité le regroupement des niveaux C1 et C2 (voir plus haut) pour les analyses, créant une catégorie "C1-C2" de 36 textes (20,2% du corpus). Bien que cette solution permette d'atteindre une taille d'échantillon plus acceptable, elle pourrait potentiellement masquer des différences linguistiques entre ces deux niveaux de maîtrise avancée et réduit la granularité de l'analyse. Deuxièmement, même après ce regroupement, un déséquilibre persiste entre les niveaux (notamment

au niveau A1), ce qui peut conduire à des biais dans l'apprentissage automatique, les modèles ayant tendance à favoriser les classes majoritaires (He et Garcia, 2009, p. 1263).

Diversité des sources et représentativité

L'analyse des sources révèle une dépendance importante vis-à-vis de ressources en ligne. La source principale, *Lingua.com*, contribue à 51 textes, soit 28,7% du corpus, suivie par diverses ressources du site *NT2 Taalmenu*. Cette concentration sur un nombre relativement restreint de sources soulève des questions concernant la représentativité du corpus.

Selon Biber (1993), la représentativité d'un corpus dépend de sa capacité à couvrir "la gamme complète des types de textes et des distributions linguistiques dans un domaine donné" (notre traduction). Or, nous retrouvons dans notre corpus une prédominance d'articles journalistiques (73 textes) et de textes didactiques. Bien que les textes journalistiques aient été choisis pour traiter de diverses thématiques pour le niveau B2 (directement dans le manuel "Vers une communication professionnelle"), ainsi que pour le niveau C2 avec des textes choisis dans les différentes rubriques du journal *De Tijd*, d'autres genres textuels que les apprenants pourraient rencontrer pourraient ne pas être représentés.

4.4.2 Analyse critique des choix méthodologiques

Problèmes de validation et de fiabilité

L'absence de validation empirique des annotations constitue une limitation méthodologique qui compromet la fiabilité du corpus comme "gold standard" pour l'entraînement de modèles de lisibilité. Contrairement aux meilleures pratiques en annotation de corpus, qui exigent une validation inter-annotateurs avec des mesures de fiabilité comme le coefficient kappa de Cohen (1960), le corpus repose entièrement sur des annotations uniques sans vérification ultérieure de leur cohérence.

De plus, l'absence de critères explicites pour l'attribution des niveaux de difficulté constitue une limitation importante. Les manuels et ressources utilisés pourraient appliquer des critères différents pour déterminer le niveau approprié d'un texte, créant une hétérogénéité dans les annotations qui peut compromettre la cohérence du corpus.

Enfin, il convient de souligner qu'aux niveaux les plus élevés (B2 et C1-C2), notre corpus est constitué majoritairement de textes journalistiques belges. Or, notre liste de fréquence lexicale de référence provient elle-même d'un sous-corpus journalistique belge. Cette proximité entre le type de textes analysés et la source des fréquences introduit un risque de circularité méthodologique : les textes avancés, étant évalués selon des fréquences issues du même registre, pourraient obtenir artificiellement des scores plus avorables que prévu, biaisant ainsi l'interprétation des résultats pour nos variables de fréquence lexicale.

4.5 Conclusion

La constitution de ce corpus de néerlandais, bien que présentant des défis inhérents à la nature des ressources didactiques et authentiques, a été menée avec une attention particulière à la méthodologie et à la justification des choix. En s'inspirant des travaux de François, nous avons pu structurer notre approche, notamment en ce qui concerne la classification des textes.

Ce corpus représente une ressource précieuse pour l'étude de la lisibilité en néerlandais et pour le développement d'outils pédagogiques adaptés. Les informations détaillées sur sa composition et les défis rencontrés permettent une meilleure compréhension des analyses qui en découleront. La prochaine étape consiste à exploiter ce corpus pour des analyses linguistiques approfondies et à valider sa pertinence par rapport aux objectifs de cette recherche.

Le corpus, ainsi qu'un excel décrivant celui-ci sera disponible sur le lien GitHub fourni en annexe (p.109).

Chapitre 5

Protocole expérimental et démarche d'analyse

5.1 Introduction

Dans ce chapitre, nous détaillerons le protocole technique et la démarche analytique adoptés pour évaluer la capacité prédictive des variables linguistiques décrites au chapitre 3, sur la base du corpus présenté au chapitre 4.

Nous décrirons d'abord l'environnement technique et les outils de Traitement Automatique du Langage (TAL) qui ont permis l'extraction automatisée de nos variables. Ensuite, nous exposerons la démarche d'analyse statistique en deux temps :

1. Les statistiques descriptives, qui offrent une première vue de la distribution des variables ;
2. Les analyses inférentielles, notamment l'analyse de la variance (ANOVA) et la régression linéaire, qui permettent de tester nos hypothèses et de quantifier les relations entre les variables et la difficulté textuelle.

5.2 Outils techniques et extraction des variables

L'extraction des 18 variables linguistiques a été automatisée au moyen de plusieurs scripts développés dans le langage de programmation Python. Ce choix a été motivé par la richesse de son écosystème de bibliothèques dédiées à l'analyse de données et au TAL.

L'ensemble des scripts développés pour cette recherche sera mis à disposition sur un dépôt GitHub, dont le lien figure en annexe. Ce dépôt contiendra non seulement les codes utilisés pour l'extraction et l'analyse des variables linguistiques, mais également l'ensemble des ressources nécessaires à la reproduction des analyses : liste de fréquence, liste de référence, corpus complet et fichiers intermédiaires. Cette mise à disposition vise à garantir la transparence méthodologique, à faciliter la reproductibilité des résultats et à offrir à la communauté scientifique un point de départ concret pour des recherches ultérieures sur la lisibilité en néerlandais langue étrangère.

Le traitement des textes a été réalisé à l'aide de la bibliothèque *spaCy* Honnibal et

Montani (2017) et de son modèle pré-entraîné pour le néerlandais *"nl_core_news_sm"*. Ce modèle a permis d'effectuer les tâches suivantes :

- **Tokenisation :** La segmentation des textes en unités (mots, ponctuation),
- **Étiquetage morpho-syntaxique (Part-of-Speech Tagging)** : L'assignation d'une catégorie grammaticale à chaque token (nom, verbe, adjectif, etc.),
- **Lemmatisation :** La réduction des formes fléchies d'un mot à sa forme canonique ou lemme,
- **Analyse syntaxique :** La construction d'un arbre de dépendances pour chaque phrase, essentiel pour l'analyse des relations grammaticales.

L'utilisation d'un modèle unique et intégré comme celui de *spaCy* garantit une cohérence dans le traitement initial des données. Les informations extraites ont ensuite été traitées par des scripts spécifiques pour calculer la valeur de chaque variable linguistique pour chaque texte. Les résultats de ces analyses ont été systématiquement recueillis et organisés dans des fichiers au format **.csv**, facilitant ainsi leur importation pour des analyses statistiques ultérieures. Ces fichiers seront également disponibles dans le dépôt GitHub fourni en annexe.

5.2.1 Extraction des variables spécifiques

Bien que la biliothèque *spaCy* soit extrêmement puissante pour l'analyse morpho-syntaxique, elle ne permet pas de calculer directement des indicateurs de complexité aussi spécifiques que ceux requis pour notre étude. En effet, des variables comme le nombre de subordonnées ou la présence de verbes à particule séparable ne sont pas des attributs directement fournis par le modèle. Il a donc été nécessaire de développer une logique applicative s'appuyant sur les analyses de bas niveau de *spaCy* pour extraire ces informations de plus haut niveau.

Nombre de subordonnées (NbSubord)

L'extraction de cette variable repose sur l'identification des tokens qui marquent explicitement une proposition subordonnée. Le script parcourt le texte analysé par spaCy et compte chaque token dont la relation de dépendance (`token.dep_`) correspond à un marqueur de subordination, tel que *mark* (marqueur de subordination comme "dat", "om"), *advcl* (proposition subordonnée adverbiale), *csubj* (sujet propositionnel) ou *ccomp* (complément propositionnel). Cette approche, bien qu'efficace, est une heuristique : elle se concentre sur les marqueurs les plus évidents et pourrait ne pas capturer toutes les formes de subordination, notamment les relatives sans marqueur explicite. Cependant, elle offre une approximation robuste et automatisable de la densité de subordination dans un texte.

Proportion de verbes à particule séparble

Cette variable, spécifique à la syntaxe du néerlandais, est particulièrement complexe à extraire. Les verbes à particule séparable (ex : *opbellen*) voient leur particule (*op*) détachée et souvent rejetée en fin de proposition principale. Pour quantifier ce phénomène, le script identifie d'abord tous les verbes du texte (`token.pos_` == "VERB"). Ensuite, pour chaque verbe, il examine ses "enfants" dans l'arbre de dépendances à la recherche d'une particule qui lui est syntaxiquement liée. En néerlandais, cette relation est typiquement marquée par la dépendance *compound :prt* (*compound* : l'élément fait partie d'une construction lexicale composée, *prt* : il s'agit d'une particule). La variable est ensuite calculée comme le ratio entre le nombre de verbes présentant une telle particule et le nombre total de verbes.

Cette méthode est une innovation de notre étude, car elle cible une difficulté grammaticale majeure pour les apprenants francophones qui n'est pas directement mesurable avec les outils TAL standards.

Proportion de phrases avec subordonnées (PropSOV)

Contrairement à *NbSubord* qui compte le nombre de subordonnée présent dans chaque texte, cette variable vise à mesurer la distribution de la complexité syntaxique à travers les phrases. Le script segmente d"abord le texte en phrases, il vérifie ensuite pour chaque phrase la présence d'au moins un verbe ou auxiliaire ayant une dépendance caractéristique d'une proposition subordonnée. Si une telle dépendance est trouvée, la phrase entière est comptabilisée comme "contenant une subordonnée". Le ratio final est le nombre de ces phrases complexes sur le nombre total de phrases. Cette approche permet de distinguer un texte où la complexité est concentrée dans quelques phrases très longues d'un texte où elle est répartie plus uniformément.

5.3 Démarche d'analyse statistique

Une fois les données extraites et structurées, nous avons procédé à l'analyse statistique en plusieurs étapes.

5.3.1 Statistiques descriptives

Pour chaque variable, des statistiques descriptives complètes ont été calculées (moyenne, médiane, écart-type, minimum, maximum, quartiles) via des scripts pythons. Les résultats sont présentés en tableaux et en graphes, réalisés grâce aux bibliothèques *seaborn* Waskom et al. (2017) ou *matplotlib* Hunter (2007) de python. Cette première étape est essentielle car elle permet de visualiser la distribution des variables à travers les différents niveaux du CECR et de repérer de premières tendances ou des anomalies potentielles.

5.3.2 Analyse de la variance (ANOVA)

Ensuite, pour déterminer si les différences observées entre les différents niveaux du CECR pour une variable donnée sont statistiquement significatives ou pas, nous avons eu recours à l'analyse de la variance (ANOVA). Nous avons choisi ce test statistique car il est fréquemment utilisé lorsque l'on souhaite comparer plus de deux moyennes entre elles. Cette méthode nous permet de répondre à la question : "La moyenne de cette variable linguistique est-elle significativement différente d'un niveau CECR à un autre ?". Si, suite au test, nous obtenons une p-value inférieur au seuil de signification de 0.05, nous pouvons conclure que cette variable a un pouvoir discriminant.

5.3.3 Régression linéaire multiple

Enfin, pour modéliser la relation entre plusieurs prédicteurs et la difficulté textuelle, nous avons utilisé une régression linéaire multiple. Cette technique vise à construire un modèle mathématique qui prédit la valeur d'une variable dépendante (ici, la difficulté) à partir de variables indépendantes (nos indices linguistiques).

La régression linéaire offre un cadre interprétable où les coefficients attribués à chaque variable permettent de quantifier son poids et la direction de son influence sur la prédiction

finale. L'objectif est de trouver la combinaison de variables qui explique le plus de variance posible dans les données de difficulté, tout en restant parcimonieux pour éviter le sur-ajustement (*overfitting*).

En combinant ces différentes approches, de l'extraction automatisée à la modélisation statistique, ce protocole assure une démarche à la fois rigoureuse, transparente et ancrée dans les pratiques actuelles du domaine de la lisibilité computationnelle.

Troisième partie

Expérimentations et résultats

Chapitre 6

Résultats et analyses

6.1 Introduction

Ce chapitre présente les résultats empiriques de notre étude sur les facteurs de lisibilité du néerlandais langue étrangère. Après avoir établi notre cadre théorique et notre méthodologie, nous nous concentrons ici sur l'analyse quantitative des variables linguistiques extraites de notre corpus.

L'objectif est double : premièrement, identifier les variables les plus discriminantes pour prédire la difficulté d'un texte pour un apprenant francophone, deuxièmement, évaluer la performance d'un modèle prédictif basé sur ces variables.

Dans une première partie, nous analysons la distribution de chaque variable linguistique, lexicale et syntaxique, à travers les différents niveaux de compétence. Nous y examinons les tendances, les variations et la significativité statistique (via des tests ANOVA) pour déterminer quelles caractéristiques du texte évoluent le plus significativement avec le niveau des apprenants. Une attention particulière est portée aux défis méthodologiques rencontrés, notamment un effet de circularité pour les variables de fréquence lexicale, qui nous a conduits à les exclure de la modélisation finale.

Dans une seconde partie, nous développons un modèle de régression linéaire multiple pour prédire le niveau de difficulté des textes. Nous y détaillons la performance du modèle, son pouvoir explicatif (R^2) et la contribution de chaque variable.

Enfin, ce chapitre se conclut par une discussion critique des résultats et des limites de notre modèle.

6.2 Analyse des variables lexicales

6.2.1 Analyse des variables de fréquence lexicale

Cette section est dédiée à l'analyse des variables lexicales, en examinant leur distribution à travers les différents niveaux de compétence linguistique. Nous discuterons des tendances observées, des variations et de la significativité statistique de ces variables.

Statistiques descriptives

L'ensemble des résultats obtenus pour les variables de fréquence lexicale révèle des anomalies théoriques majeures qui s'expliquent principalement par un effet de circularité méthodologique inhérent à notre approche. En effet, nos données de fréquence proviennent exclusivement des sous-collections "journaux" et "Belgique" du corpus SoNaR-500, tandis que notre corpus d'analyse est constitué quasi-exclusivement de textes journalistiques à partir du niveau B2. Cette convergence crée un biais de circularité fondamental : les textes journalistiques de niveaux avancés sont évalués selon des normes de fréquence établies à partir du même type de textes journalistiques, ce qui explique pourquoi le vocabulaire spécialisé de ces textes obtient artificiellement des scores de fréquence élevés. Contrairement aux prédictions théoriques classiques qui établissent une relation inverse entre fréquence lexicale et niveau de difficulté François (2011, p 219.), nos résultats montrent des progressions ascendantes ou des stabilisations inattendues.

Pour résoudre ces limitations méthodologiques, les recherches futures pourraient adopter plusieurs stratégies :

1. Utiliser une liste de fréquence généraliste couvrant l'ensemble des registres de la langue plutôt qu'un sous-domaine spécifique,
2. Constituer un corpus d'entraînement diversifié incluant des textes de genres variés à tous les niveaux CECR,
3. Implémenter une approche multi-référentielle combinant plusieurs dictionnaires de fréquence (général, journalistique, académique) pour capturer la complexité lexicale sous différents angles.

Ces améliorations permettraient d'obtenir des mesures de fréquence lexicale plus robustes et théoriquement cohérentes pour l'évaluation automatique de la lisibilité en néerlandais langue étrangère.

Cependant, après plusieurs tentatives nous nous sommes rendus à l'évidence que l'implémentation de ces améliorations méthodologiques dépasse largement le cadre et les ressources de la présente recherche. En effet, la constitution d'un corpus diversifié en genres textuels nécessiterait un travail de collecte et d'annotation considérable qui représenterait à lui seul un projet de recherche complet. Enfin,l'approche multi-référentielle implique des développements techniques complexes et des validations empiriques étendues qui excèdent les contraintes temporelles d'un mémoire de master. Dans le contexte de cette étude exploratoire, nous avons donc choisi de maintenir notre approche méthodologique initiale tout en documentant explicitement ses limitations, ce qui permet d'établir une base de référence pour les recherches futures et de contribuer à l'identification des défis méthodologiques spécifiques à l'évaluation automatique de la lisibilité en néerlandais langue étrangère.

Variable "mean_fl" (fréquence moyenne) : Nos résultats pour la variable **mean_fl** révèlent des tendances qui méritent d'être confrontées aux observations de François (2011) et à la littérature plus large sur les effets de fréquence en acquisition L2. En effet, nous observons une progression croissante claire du niveau A1 (9.47×10^5) au C1-C2 (1.23×10^6), soit une augmentation de 29%. Cette évolution, a priori inattendue au regard des prédictions classiques sur l'effet de fréquence, peut toutefois s'expliquer par le biais de circularité méthodologique évoqué précédemment, qui tend à renforcer artificiellement la corrélation entre fréquence et niveau de compétence.

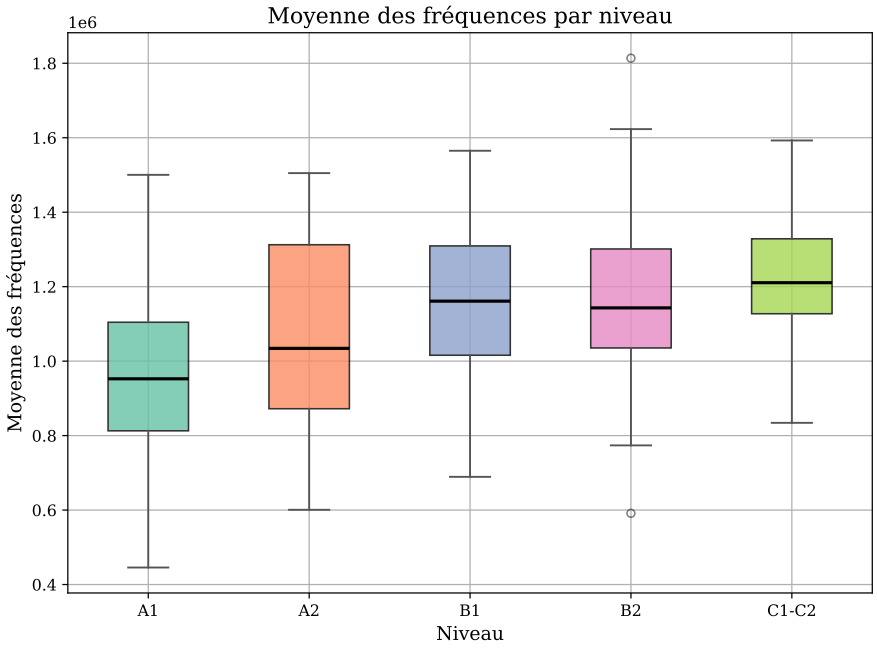

FIGURE 6.1 – Distribution de la fréquence lexicale moyenne (mean_fl) par niveau CECR

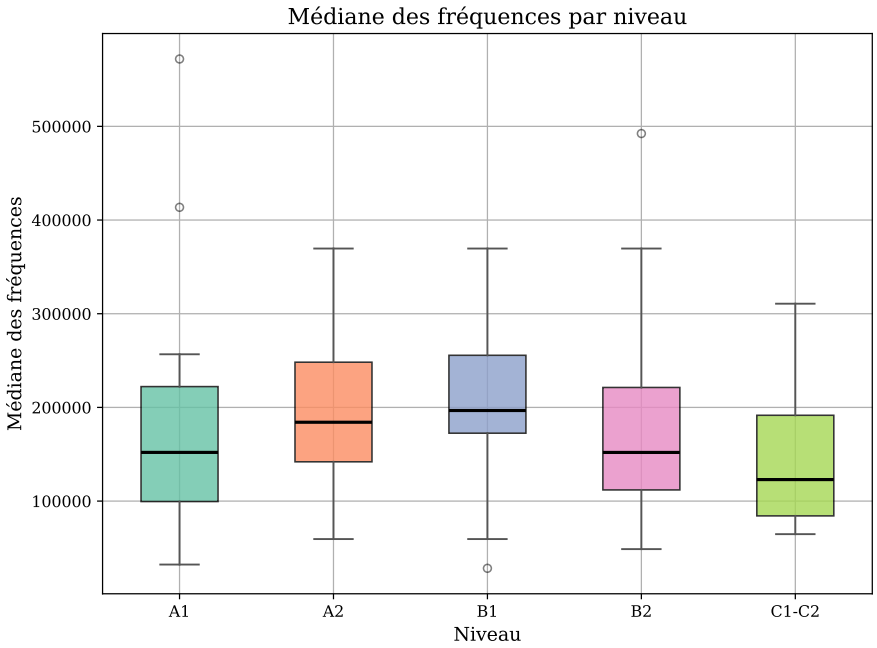

FIGURE 6.2 – Distribution de la fréquence lexicale médiane (median_fl) par niveau CECR

Tableau 6.1 – Statistiques descriptives — Mesure : **mean_fl**

Statistique	A1	A2	B1	B2	C1-C2
Moyenne	9.47×10^5	1.07×10^6	1.16×10^6	1.16×10^6	1.23×10^6
Écart-type	2.72×10^5	2.65×10^5	2.23×10^5	2.27×10^5	1.69×10^5
Min	4.46×10^5	6.01×10^5	6.89×10^5	5.91×10^5	8.34×10^5
Max	1.50×10^6	1.50×10^6	1.57×10^6	1.81×10^6	1.59×10^6
Médiane	9.53×10^5	1.03×10^6	1.16×10^6	1.14×10^6	1.21×10^6
1er quartile	8.13×10^5	8.72×10^5	1.02×10^6	1.04×10^6	1.13×10^6
3e quartile	1.10×10^6	1.31×10^6	1.31×10^6	1.30×10^6	1.33×10^6

Tableau 6.2 – Statistiques descriptives — Mesure : **median_fl**

Statistique	A1	A2	B1	B2	C1-C2
Moyenne	1.79×10^5	1.94×10^5	2.07×10^5	1.85×10^5	1.44×10^5
Écart-type	1.24×10^5	7.72×10^4	7.99×10^4	9.97×10^4	7.29×10^4
Min	3.23×10^4	5.95×10^4	2.83×10^4	4.87×10^4	6.47×10^4
Max	5.72×10^5	3.70×10^5	3.70×10^5	4.92×10^5	3.11×10^5
Médiane	1.52×10^5	1.84×10^5	1.97×10^5	1.52×10^5	1.23×10^5
1er quartile	9.95×10^4	1.42×10^5	1.72×10^5	1.12×10^5	8.42×10^4
3e quartile	2.22×10^5	2.48×10^5	2.56×10^5	2.21×10^5	1.92×10^5

Tableau 6.3 – Statistiques descriptives — Mesure : **p75_fl**

Statistique	A1	A2	B1	B2	C1-C2
Moyenne	1.08×10^6	1.44×10^6	1.47×10^6	1.40×10^6	1.59×10^6
Écart-type	6.07×10^5	7.35×10^5	6.78×10^5	6.27×10^5	6.32×10^5
Min	3.55×10^5	5.70×10^5	5.06×10^5	5.44×10^5	9.38×10^5
Max	2.85×10^6	2.85×10^6	2.85×10^6	3.38×10^6	2.85×10^6
Médiane	9.86×10^5	1.18×10^6	1.24×10^6	1.22×10^6	1.32×10^6
1er quartile	7.59×10^5	9.53×10^5	1.05×10^6	1.07×10^6	1.20×10^6
3e quartile	1.18×10^6	1.62×10^6	1.62×10^6	1.45×10^6	1.62×10^6

Variable "median_fl" (fréquence médiane) : La médiane des fréquences lexicales (median_fl) montre une tendance générale à la baisse. Elle passe d'une moyenne de 1.79×10^5 au niveau A1 avant de monter jusqu'à 2.07×10^5 au niveau B1, pour enfin redescendre à 1.44×10^5 au niveau C1-C2.

Contrairement à la moyenne qui montrait une progression paradoxale, la médiane suit une trajectoire plus conforme aux attentes théoriques : les textes destinés aux niveaux plus avancés ont tendance à utiliser un vocabulaire typiquement moins fréquent.

Quant à la dispersion, la hauteur des boîtes dans le boxplot (l'écart interquartile) indique que la variabilité est la plus forte au niveau A1, diminue en A2 et B1, puis augmente légèrement en B2, avant de se réduire de nouveau au niveau C1-C2. Cela suggère que les

Tableau 6.4 – Statistiques descriptives — Mesure : **p90_fl**

Statistique	A1	A2	B1	B2	C1-C2
Moyenne	2.96×10^6	3.34×10^6	3.63×10^6	3.63×10^6	3.53×10^6
Écart-type	7.04×10^5	8.92×10^5	1.07×10^6	1.10×10^6	1.74×10^5
Min	9.98×10^5	1.45×10^6	2.67×10^6	1.45×10^6	3.24×10^6
Max	3.55×10^6	8.17×10^6	8.17×10^6	8.17×10^6	4.47×10^6
Médiane	3.24×10^6	3.38×10^6	3.46×10^6	3.52×10^6	3.52×10^6
1er quartile	2.72×10^6	3.24×10^6	3.38×10^6	3.35×10^6	3.52×10^6
3e quartile	3.38×10^6	3.52×10^6	3.52×10^6	3.52×10^6	3.55×10^6

textes pour débutants ont une plus grande hétérogénéité dans la fréquence typique de leur vocabulaire.

En ce qui concerne les outliers (valeurs aberrantes), le graphique en montre quelques-unes notamment au niveau A1 et B2, indiquant la présence de textes dont le profil lexical s'écarte significativement de la tendance centrale de leur catégorie.

Ces résultats sont donc plus cohérents avec les fondements de la recherche sur la lisibilité et l'acquisition d'une langue seconde que ceux observés pour la moyenne (mean_fl). Ces analyses confirment la supériorité de la médiane sur la moyenne comme indicateur de difficulté lexicale. En effet, la moyenne est très sensible aux valeurs extrêmes. Dans un texte, la présence massive de mots grammaticaux très courts et extrêmement fréquents (articles, prépositions, etc.) peut gonfler artificiellement la fréquence moyenne, masquant ainsi la difficulté réelle du vocabulaire de contenu. La médiane, en revanche, n'est pas affectée par ces valeurs extrêmes. Elle représente le "point milieu" de la distribution des fréquences et reflète donc mieux la fréquence du mot typique du texte.

Variable "p75_fl" (75e percentile) Le 75e percentile de la fréquence lexicale (p75_fl) présente une progression ascendante de 47,2% entre les niveaux A1 et C1-C2, passant de 1.08×10^6 au niveau A1 à 1.59×10^6 au niveau C1-C2. Cette progression présente une augmentation rapide de A1 à A2, une stabilisation relative entre A2 et B1 suivi d'une légère régression au niveau B2, avant d'augmenter fortement au niveau C1-C2. Cette augmentation indique une progression marquée dans l'utilisation de mots plus fréquents aux niveaux avancés, ce qui constitue un pattern inattendu par rapport aux prédictions théoriques classiques. Ce pattern peut être expliqué par le biais méthodologique que présente notre étude et dont nous avons déjà discuté.

La variabilité maximale se trouve au niveau A2 et suggère une hétérogénéité importante dans l'usage lexical à ce niveau, tandis que la stabilisation aux niveaux supérieurs indique une plus grande homogénéité dans l'usage lexical aux niveaux avancés.

Les chevauchements entre niveaux adjacents confirment la continuité progressive de l'acquisition lexicale.

Variable "p90_fl" (90e percentile)

Le 90e percetile de la fréquence lexicale présente une progression ascendante modérée de 19,3% entre les niveaux A1 et C1-C2. Cette augmentation, bien que moins prononcée que celle observée pour le 75e percentile (47.2%), confirme néanmoins une tendance contre-intuitive par rapport aux prédictions théoriques classiques. Tendance qui, encore une fois,

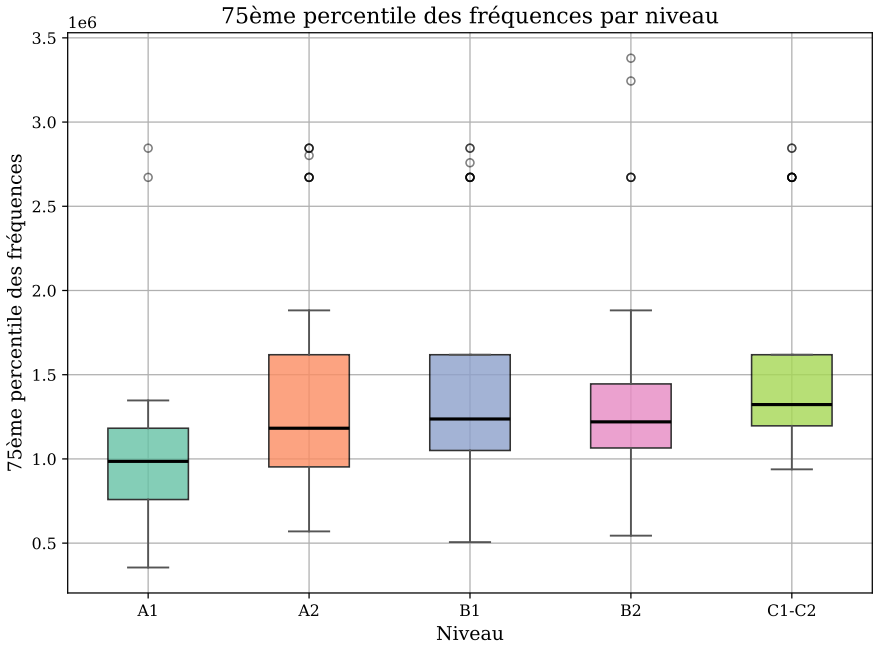

FIGURE 6.3 – Distribution du 75e percentile des fréquences lexicales (p75_fl) par niveau CECR

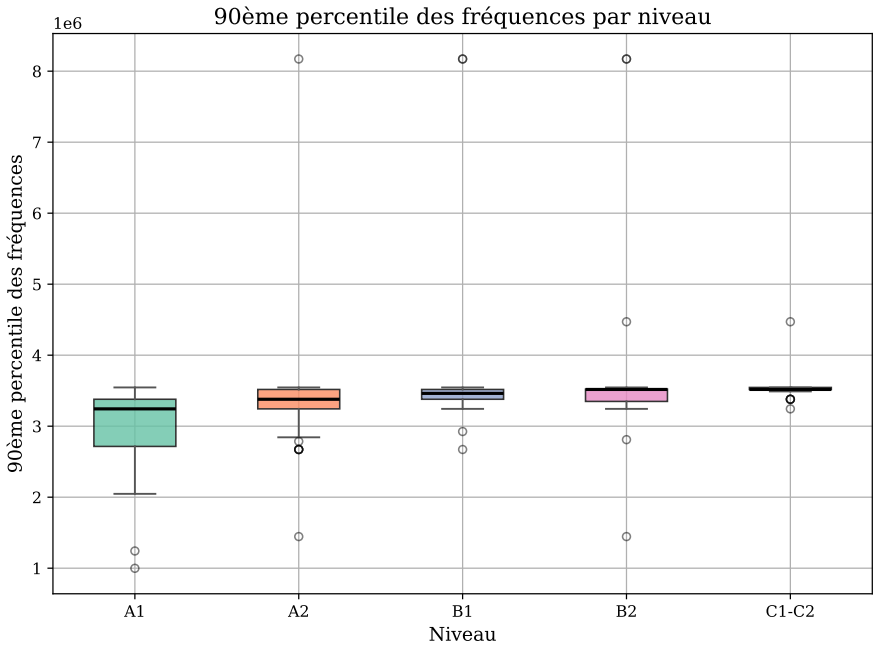

FIGURE 6.4 – Distribution du 90e percentile des fréquences lexicales (p75_fl) par niveau CECR

peut être expliquée par le biais méthodologique que présente notre étude et dont nous avons déjà discuté.

La variabilité suit un pattern en montagne avec une augmentation progressive de A1 à B2, atteignant son maximum au niveau B2, puis une légère chute au niveau C1-C2. Cette réduction suggère une homogénisation du vocabulaire de haute fréquence aux niveaux avancés.

Analyse de la variance

Nous avons également réalisé un test ANOVA unidirectionnel sur les différentes variables lexicales afin d'évaluer si leurs moyennes diffèrent de manière significative selon les niveaux CECR. Les résultats (voir tableau) indiquent que les variables **median_fl** ($F = 5.30$, $p < 0.05$) et **mean_fl** ($F = 3.50$, $p < 0.05$) présentent des différences significatives entre niveau.

Variable	F-statistique	p-value
mean_fl	3.50	0.009
median_fl	5.30	4.73×10^{-4}
p75_fl	1.51	0.202
p90_fl	2.48	0.046

Tableau 6.5 – Résultats du test ANOVA pour les variables lexicales selon les niveaux CECR.

La variable **P90_fl**, l'analyse de variance révèle une différence significative mais fragile entre les niveaux CECR ($F = 2.48$, $p = 0.046$). En revanche, la variable **p70_fl** ne présente pas de différences significative ($p = 0.202$).

Cependant, nous nous devons d'interpréter ces résultats avec une distance critique en raison de l'effet de circularité méthodologique identifié. Comme mentionné précédemment, le corpus d'analyse étant majoritairement composé de textes journalistiques de niveaux avancés, et les normes de fréquence étant établies à partir de données similaires, les significativités statistiques observées doivent être considérées avec prudence. Ces variables montrent des tendances contre-intuitives qui sont probablement des artefacts de ce biais plutôt que des reflets fidèles de la difficulté lexicale pour les apprenants.

En revanche, la variable **median_fl** présente une trajectoire plus conforme aux attentes théoriques (tendance à la baisse de la fréquence avec le niveau), suggérant une plus grande robustesse face à ce biais. La significativité de l'ANOVA pour **median_fl** est donc plus fiable et renforce son rôle potentiel comme indicateur pertinent de la complexité lexicale.

6.2.2 Analyse de la proportion d'absent d'une liste de référence

Statistiques descriptives

Selon les modèles théoriques de l'acquisition lexicale en langue seconde, nous nous attendions à observer une augmentation progressive des variables PA_SUBTLEX avec

les niveaux de compétence CECR.

Les résultats observés dans notre corpus révèlent des tendances conformes à ce qui est attendu par la littérature. Nous les détaillons ci-dessous.

Tableau 6.6 – Mesure : Proportion d'absents de la liste des 1500 mots

niveau	Moyenne	Écart-type	Min	Max	Médiane	Q1	Q3
A1	0.201700	0.082060	0.090910	0.383180	0.186620	0.150270	0.245560
A2	0.205730	0.059340	0.090230	0.310340	0.205530	0.165480	0.244040
B1	0.231680	0.062470	0.123290	0.380000	0.236120	0.180590	0.273880
B2	0.300920	0.050350	0.166670	0.402600	0.302220	0.275880	0.331770
C1-C2	0.341010	0.048250	0.191390	0.402140	0.346500	0.313840	0.382370

Tableau 6.7 – Mesure : Proportion d'absents de la liste des 3000 mots

niveau	Moyenne	Écart-type	Min	Max	Médiane	Q1	Q3
A1	0.158660	0.080700	0.062500	0.364490	0.142720	0.100740	0.173230
A2	0.152050	0.054210	0.045200	0.282760	0.150770	0.113170	0.183730
B1	0.180330	0.051670	0.094240	0.293330	0.182910	0.138060	0.224580
B2	0.246740	0.044450	0.155560	0.343280	0.254570	0.216920	0.274040
C1-C2	0.288380	0.049870	0.157890	0.366170	0.292350	0.258300	0.330880

Variable "PA_SUBTLEX_1500" (Proportion d'absent de la sous-liste de 1500 mots) : L'analyse du boxplot de **PA_SUBTLEX_1500** révèle une progression claire et régulière avec les niveaux CECR. Au niveau A1, la distribution présente une médiane située autour de 0,19 avec une dispersion relativement faible, les valeurs s'étendant approximativement de 0,09 à 0,38. Le premier quartile se situe vers 0,15 et le troisième quartile vers 0,25, indiquant une distribution relativement concentrée.

Au niveau A2, nous observons une stabilité remarquable par rapport à A1, avec une médiane légèrement supérieure (environ 0,20) et une dispersion similaire. Cette stabilité suggère que les textes de niveaux A1 et A2 présentent une charge lexicale comparable en termes de vocabulaire absent de la liste des 1500 mots les plus fréquents.

Tableau 6.8 – Mesure : Proportion d'absents de la liste complète

niveau	Moyenne	Écart-type	Min	Max	Médiane	Q1	Q3
A1	0.130590	0.087880	0.034480	0.364490	0.104930	0.073100	0.133400
A2	0.121530	0.055750	0.036720	0.275860	0.100840	0.085760	0.156380
B1	0.152430	0.045650	0.065420	0.246670	0.152860	0.107860	0.181080
B2	0.214360	0.043000	0.141670	0.313430	0.210610	0.180930	0.248660
C1-C2	0.250800	0.045900	0.133970	0.326530	0.260430	0.224830	0.274830

Tableau 6.9 – Mesure : Proportion d'absents uniques de la liste des 1500 mots

niveau	Moyenne	Écart-type	Min	Max	Médiane	Q1	Q3
A1	0.231460	0.061860	0.125000	0.318180	0.244010	0.188950	0.279910
A2	0.272590	0.070720	0.119050	0.428570	0.270490	0.222790	0.315220
B1	0.318090	0.071470	0.168540	0.454550	0.313350	0.272200	0.363250
B2	0.412700	0.094020	0.213480	0.534980	0.445510	0.374540	0.487850
C1-C2	0.515740	0.074200	0.253730	0.606770	0.527820	0.491180	0.561680

Tableau 6.10 – Mesure : Proportion d'absents uniques de la liste des 3000 mots

niveau	Moyenne	Écart-type	Min	Max	Médiane	Q1	Q3
A1	0.173640	0.048480	0.086960	0.250000	0.178410	0.142690	0.204340
A2	0.201110	0.054710	0.095240	0.329670	0.200000	0.163040	0.231130
B1	0.242140	0.056880	0.126210	0.363640	0.239620	0.206330	0.276160
B2	0.340060	0.084110	0.157140	0.448600	0.359430	0.293530	0.408960
C1-C2	0.435640	0.069960	0.208960	0.526040	0.448700	0.411300	0.486210

Le niveau B1 marque une transition importante avec une augmentation notable de la médiane (environ 0,24) et un élargissement de la distribution. L'écart interquartile s'étend davantage, suggérant une plus grande hétérogénéité dans la complexité lexicale des textes de ce niveau.

Au niveau B2, l'augmentation se poursuit de manière marquée avec une médiane atteignant environ 0,30. La distribution présente une asymétrie vers les valeurs élevées, avec quelques valeurs aberrantes inférieures (autour de 0,17 - 0,19), suggérant la présence de textes exceptionnellement simples pour ce niveau.

Les niveaux C1-C2 présentent la plus forte proportion d'absents avec une médiane autour de 0,35. La distribution est large et présente plusieurs valeurs aberrantes, tant inférieures (autour de 0,19) que supérieures (dépassant 0,40), reflétant une grande diversité dans les types de textes proposés à ces niveaux avancés.

Variable PA_SUBTLEX_3000 (Proportion d'absents de la sous-liste de

Tableau 6.11 – Mesure : Proportion d'absents uniques de la liste complète

niveau	Moyenne	Écart-type	Min	Max	Médiane	Q1	Q3
A1	0.135030	0.055420	0.031750	0.250000	0.135020	0.095740	0.156660
A2	0.155260	0.048820	0.071430	0.302630	0.152170	0.124520	0.183020
B1	0.202070	0.053900	0.100000	0.353540	0.196450	0.172680	0.228950
B2	0.293270	0.079240	0.142860	0.401130	0.313970	0.233540	0.359340
C1-C2	0.380340	0.065840	0.171640	0.487290	0.392420	0.357750	0.413550

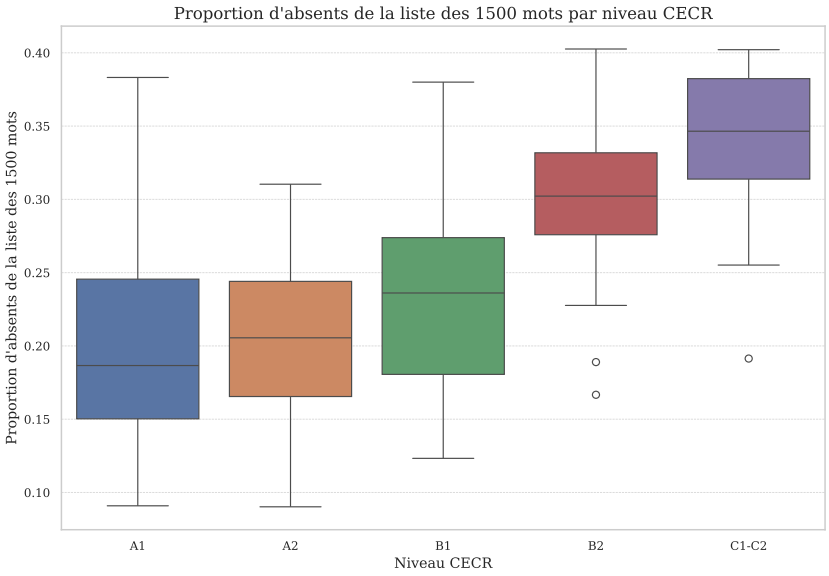

FIGURE 6.5 – Proportion d'absent de la liste des 1500 mots les plus fréquents

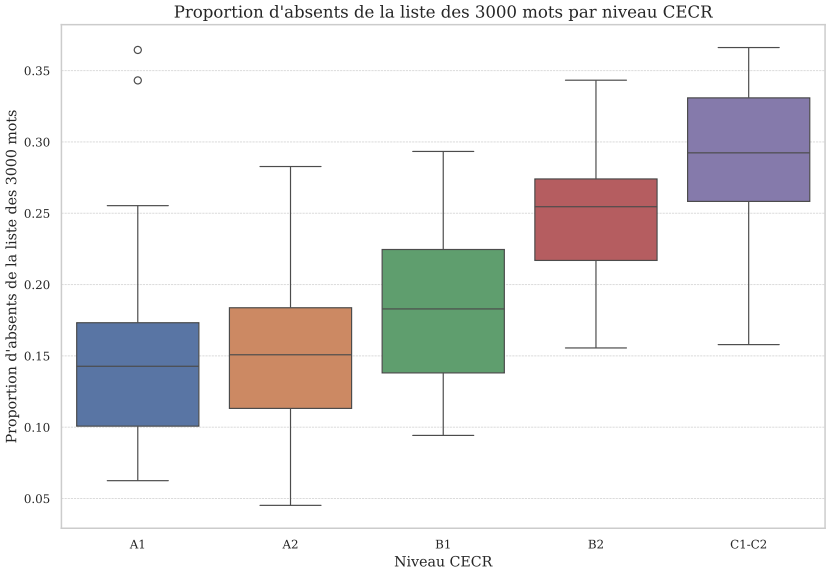

FIGURE 6.6 – Proportion d'absent de la liste des 3000 mots les plus fréquents

3000 mots) : L'analyse de **PA_SUBTLEX_3000** révèle des tendances similaires aux variables précédentes mais avec des valeurs absolues plus faibles, ce qui est attendu compte tenu de l'élargissement de la liste de référence.

Au niveau A1, la médiane se situe autour de 0,16 avec une distribution relativement concentrée. Les valeurs s'étendent approximativement de 0,06 à 0,26, indiquant une charge lexicale modérée même pour les mots au-delà des 1500 plus fréquents.

Le niveau A2 présente une stabilité remarquable avec une médiane similaire (environ 0,15) et une dispersion comparable. Cette stabilité renforce l'observation d'une progression lexicale graduelle entre les premiers niveaux.

Au niveau B1, nous observons une augmentation notable avec une médiane atteignant environ 0,18. La distribution s'élargit, particulièrement vers les valeurs supérieures, suggérant l'introduction d'un vocabulaire de fréquence intermédiaire.

Le niveau B2 marque une progression substantielle avec une médiane autour de 0,25. La distribution présente une asymétrie vers les valeurs élevées, avec quelques valeurs aberrantes inférieures, indiquant une hétérogénéité croissante.

Les niveaux C1-C2 présentent les valeurs les plus élevées avec une médiane approchant 0,29. La distribution est étalée avec plusieurs valeurs aberrantes, reflétant la diversité lexicale des textes avancés.

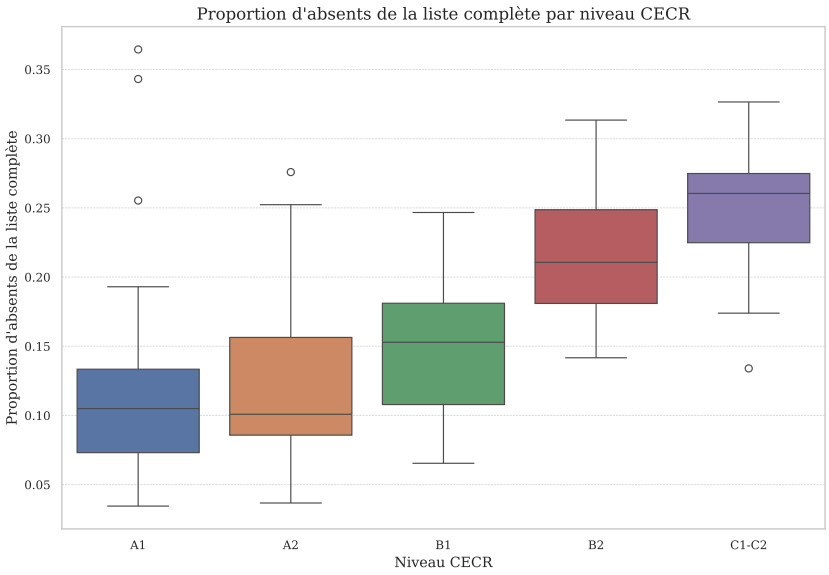

FIGURE 6.7 – Proportion d'absent de la liste des 4621 mots les plus fréquents

Variable "PA_SUBTLEX_MAX" (Proportion d'absents de la sous-liste de 4621 mots) : **PA_SUBTLEX_MAX**, utilisant l'ensemble de la base SUBTLEX, présente les valeurs les plus faibles mais des tendances proportionnellement similaires.

Au niveau A1, la médiane se situe autour de 0,13 avec une distribution concentrée. Ces valeurs relativement faibles indiquent que même les textes débutants contiennent principalement du vocabulaire présent dans SUBTLEX.

Le niveau A2 maintient une médiane similaire (environ 0,12) avec une légère réduction, suggérant une possible optimisation lexicale à ce niveau.

Au niveau B1, nous observons une augmentation vers 0,15, marquant l'introduction de vocabulaire de très basse fréquence.

Le niveau B2 présente une progression notable avec une médiane atteignant environ 0,21, indiquant l'introduction de vocabulaire spécialisé.

Les niveaux C1-C2 atteignent une médiane d'environ 0,25 avec une distribution étalée, reflétant l'utilisation de vocabulaire très spécialisé ou technique.

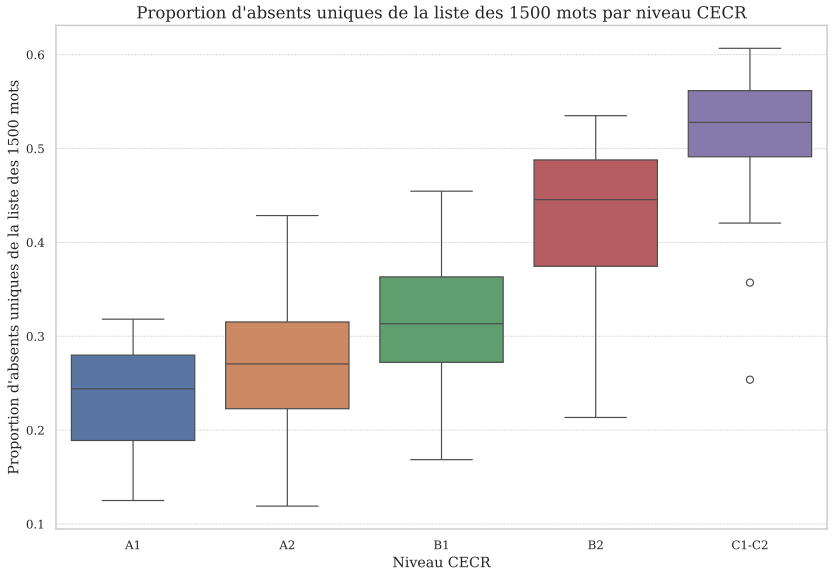

FIGURE 6.8 – Proportion d'absents uniques de la liste des 1500 mots les plus fréquents

Variable "PA_SUBTLEX_1500_U" (Proportion d'absents uniques de la sous-liste de 1500 mots) : Le graphique de **PA_SUBTLEX_1500_U** révèle des tendances similaires mais avec des valeurs systématiquement plus élevées, ce qui est cohérent avec la nature de cette mesure qui comptabilise les types lexicaux distincts.

Au niveau A1, la médiane se situe autour de 0,24 avec une distribution relativement symétrique s'étendant d'environ 0,12 à 0,32. Cette valeur plus élevée par rapport à PA_SUBTLEX_1500 standard indique que même aux niveaux débutants, la diversité lexicale des mots absents est substantielle.

Le niveau A2 présente une médiane légèrement supérieure (environ 0,27) avec une dispersion accrue, particulièrement vers les valeurs élevées. Cette augmentation suggère une diversification du vocabulaire introduit dès le niveau A2.

Au niveau B1, nous observons une progression notable avec une médiane atteignant environ 0,31. La distribution s'élargit considérablement, avec un troisième quartile s'approchant de 0,36, indiquant que 25% des textes de ce niveau présentent plus de 36% de types lexicaux absents de la liste des 1500 mots.

Le niveau B2 marque une augmentation substantielle avec une médiane autour de 0,45. La distribution présente une asymétrie marquée vers les valeurs élevées, avec un troisième quartile dépassant 0,49. Quelques valeurs aberrantes inférieures (autour de 0,21-0,22) suggèrent la présence de textes exceptionnellement simples.

Les niveaux C1-C2 présentent les valeurs les plus élevées avec une médiane atteignant environ 0,53. La distribution est très étalée, s'étendant de 0,25 à plus de 0,61, avec plusieurs

valeurs aberrantes. Cette grande variabilité reflète la diversité des domaines de spécialisation et des genres textuels aux niveaux avancés.

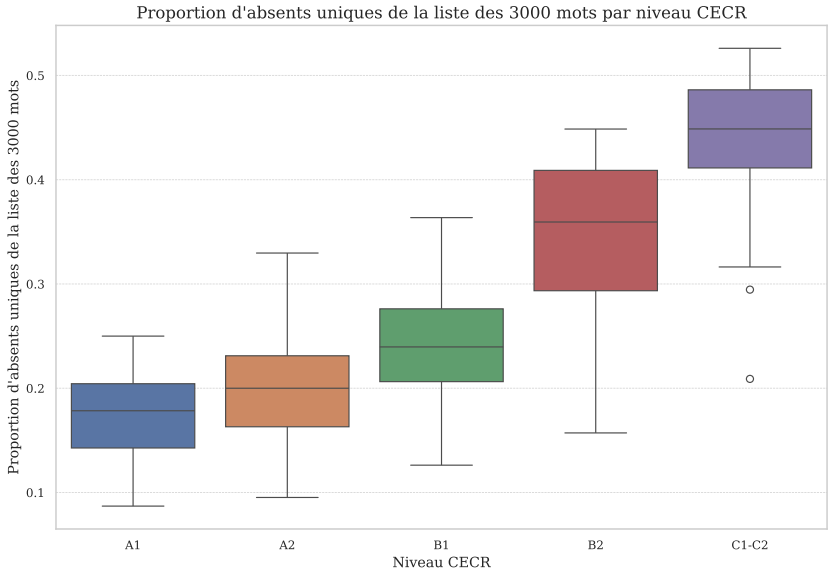

FIGURE 6.9 – Proportion d'absents uniques de la liste des 3000 mots les plus fréquents

Variable "PA_SUBTLEX_3000_U" (Proportion d'absents uniques de la sous-liste de 3000 mots) : **PA_SUBTLEX_3000_U** suit des tendances similaires aux autres variables uniques, avec des valeurs systématiquement supérieures à la version standard.

Au niveau A1, la médiane se situe autour de 0,17 avec une distribution relativement symétrique. Cette valeur, bien que plus faible que PA_SUBTLEX_1500_U, reste substantielle.

Le niveau A2 présente une médiane légèrement supérieure (environ 0,20) avec une dispersion accrue. L'augmentation est plus marquée que pour la version standard, soulignant l'importance de la diversité lexicale.

Au niveau B1, la médiane atteint environ 0,24 avec un élargissement notable de la distribution. Le troisième quartile s'approche de 0,27, indiquant une diversification lexicale importante.

Le niveau B2 marque une progression substantielle avec une médiane autour de 0,34. La distribution présente une asymétrie vers les valeurs élevées, avec un troisième quartile dépassant 0,37.

Les niveaux C1-C2 présentent une médiane atteignant environ 0,44 avec une distribution très étalée, reflétant la complexité lexicale des textes avancés.

Variable "PA_SUBTLEX_MAX_U" (Proportion d'absents de la sous-liste de 4621 mots) : PA_SUBTLEX_MAX_U présente des tendances cohérentes avec les autres variables uniques, avec des valeurs supérieures à la version standard.

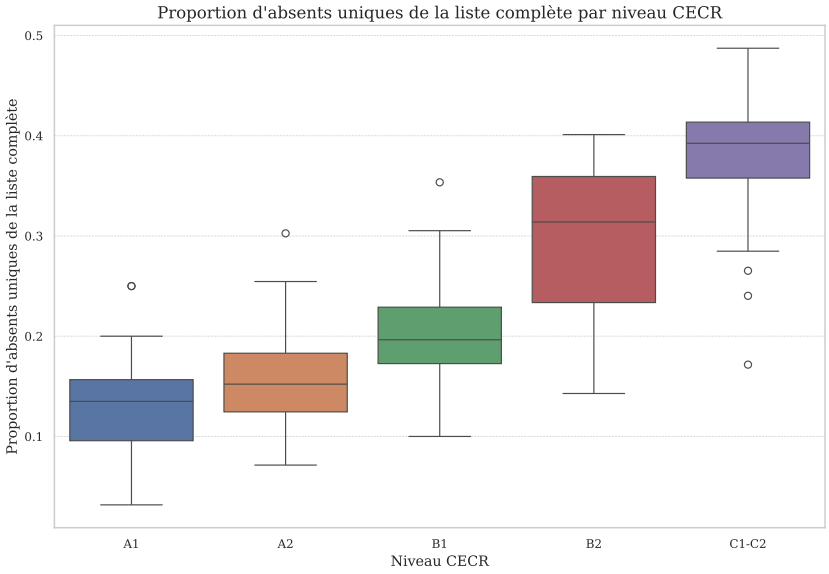

FIGURE 6.10 – Proportion d'absents uniques de la liste des 4621 mots les plus fréquents

Au niveau A1, la médiane se situe autour de 0,14 avec une distribution relativement symétrique.

Le niveau A2 présente une médiane légèrement supérieure (environ 0,15) avec une dispersion accrue.

Au niveau B1, la médiane atteint environ 0,20 avec un élargissement de la distribution.

Le niveau B2 marque une progression substantielle avec une médiane autour de 0,29.

Les niveaux C1-C2 présentent une médiane atteignant environ 0,38 avec une grande variabilité, soulignant la diversité lexicale des textes avancés.

Analyse de la variance

Variable	**F-statistique**	**p-value**
PA_SUBTLEX_1500	37.49	1.35×10^{-22}
PA_SUBTLEX_3000	41.93	1.38×10^{-24}
PA_SUBTLEX_MAX	37.98	8.06×10^{-23}
PA_SUBTLEX_1500_U	73.08	3.14×10^{-36}
PA_SUBTLEX_3000_U	90.67	1.66×10^{-41}
PA_SUBTLEX_MAX_U	90.41	1.96×10^{-41}

TABLEAU 6.12 – Résultats de l'ANOVA pour les variables d'absents d'une liste de référence

Nous avons également réalisé un test ANOVA unidirectionnel sur les différentes variables lexicales liées à la proportion d'absents d'une liste de référence afin d'évaluer si leurs moyennes diffèrent de manière significative selon les niveaux CECR. Les résultats (voir tableau) montrent que les moyennes des variables lexicales diffèrent de manière significative selon les niveux CECR, ce qui suggère un effet clair du niveu de compétence sur la complexité lexicale des textes.

6.2.3 Analyse de la diversité lexicale

Statistiques descriptives

L'analyse de la diversité lexicale mesurée par le MTLD (Measure of Textual Lexical Diversity) révèle une progression remarquablement systématique et statistiquement robuste à travers les niveaux du CECR. Les données présentées dans le tableau 6.13 illustrent une augmentation continue de la diversité lexicale, avec une moyenne passant de 31,79 au niveau A1 à 197,41 aux niveaux C1-C2, soit une progression de 521% sur l'ensemble de l'échelle de compétence.

Tableau 6.13 – Statistiques descriptives du MTLD (lemmatisé) par niveau CECR

Niveau CECR	Moyenne	Écart-type	Min	Q1	Médiane	Q3	Max
A1	31.79	18.33	11.50	20.58	27.30	38.63	82.90
A2	60.92	54.24	0.00	32.49	44.28	68.82	276.50
B1	81.18	58.22	31.80	54.22	63.58	81.83	356.19
B2	139.22	87.68	28.23	71.45	107.09	199.53	354.37
C1-C2	197.41	103.41	66.40	120.92	182.87	232.30	491.48

Au niveau A1, la distribution présente une médiane de 27,30 avec un écart interquartile relativement restreint (Q1 = 20,58 ; Q3 = 38,63), suggérant une homogénéité relative dans la diversité lexicale des textes débutants. La valeur maximale de 82,90 indique néanmoins la présence de quelques textes présentant une diversité lexicale exceptionnellement élevée pour ce niveau, possiblement dus à l'introduction précoce de vocabulaire thématique spécialisé.

Au niveau A2, une augmentation notable de 91,7% par rapport au niveau A1 (moyenne de 60,92) s'accompagne d'une expansion considérable de la variabilité (écart-type de 54,24). Cette augmentation de la dispersion suggère une hétérogénéisation des approches pédagogiques à ce niveau.

Au niveau B1, la progression se poursuit avec une moyenne de 81,18, représentant une augmentation de 33,3% par rapport au niveau A2. L'écart-type (58,22) reste élevé, confirmant la variabilité observée au niveau précédent. La médiane de 63,58 indique que la moitié des textes de ce niveau présentent une diversité lexicale supérieure à celle de la plupart des textes A2.

Au niveau B2, une accélération marquée de la diversité lexicale se manifeste avec une moyenne de 139,22 (+71,5% par rapport à B1). Cette progression substantielle coïncide avec l'introduction de textes journalistiques dans le corpus. L'écart-type de 87,68 atteint son maximum, reflétant la coexistence de textes pédagogiques traditionnels et de documents authentiques.

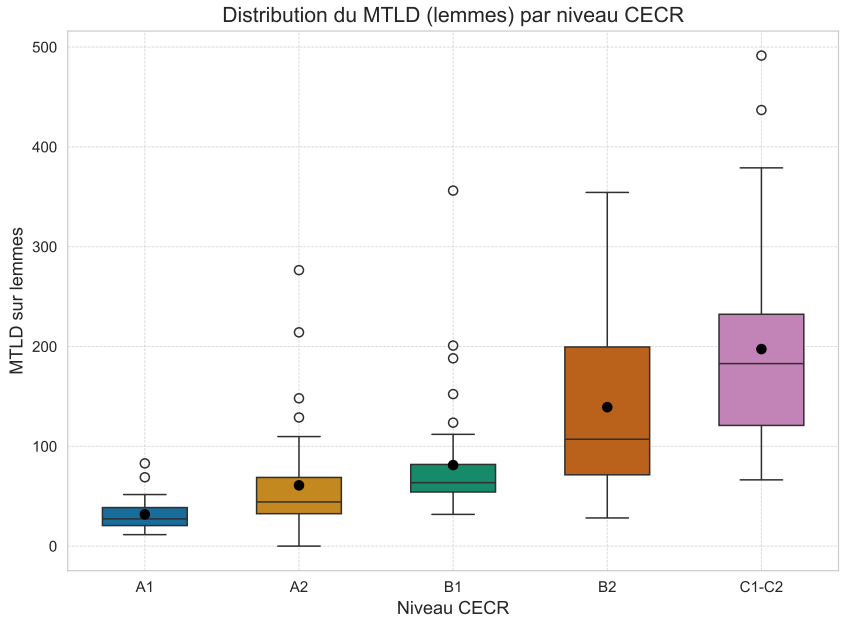

FIGURE 6.11 – Distribution du MTLD en lemmes par niveau CECR

Enfin, le niveau C1-C2 présente une moyenne de 197,41 (+41,8% par rapport à B2), avec une médiane de 182,87. Malgré l'augmentation continue de la diversité lexicale, l'écart-type (103,41) suggère une stabilisation relative de la variabilité, indiquant une convergence vers des standards lexicaux élevés caractéristiques des textes authentiques de niveau avancé.

Ces résultats convergent avec les observations de FRANÇOIS (2011, pp. 399-401) concernant l'évolution de la diversité lexicale en français langue étrangère.

Analyse de la variance

TABLEAU 6.14 – ANOVA sur le MTLD (lemmatisé) selon le niveau CECR

Variable	F	p-value
MTLD	26.44	3.87×10^{-17}

L'analyse de variance unidirectionnelle révèle un effet hautement significatif du niveau CECR sur la diversité lexicale mesurée par le MTLD ($F = 26{,}44$, $p = 3.87 \times 10^{-17}$). Cette valeur F exceptionnellement élevée, dépassant largement les seuils conventionnels de significativité, confirme que la diversité lexicale constitue un des prédicteurs robustes du niveau de compétence linguistique dans notre corpus.

6.2.4 Analyse des scores moyens de concrétude

Statistiques descriptives

L'analyse des scores moyens de concrétude (Meanconc) révèle une progression systématique et théoriquement cohérente à travers les niveaux du CECR. Les données présentées

TABLEAU 6.15 – Statistiques descriptives du score moyen de concrétude (Meanconc) par niveau CECR

Niveau	Moyenne	Écart-Type	Min	Q25	Méd.	Q75	Max
A1	2.32	0.24	1.85	2.15	2.34	2.45	2.76
A2	2.30	0.21	1.95	2.13	2.27	2.43	2.78
B1	2.21	0.19	1.88	2.10	2.15	2.33	2.77
B2	2.03	0.14	1.83	1.94	2.01	2.07	2.44
C1-C2	1.99	0.10	1.87	1.92	1.99	2.05	2.24

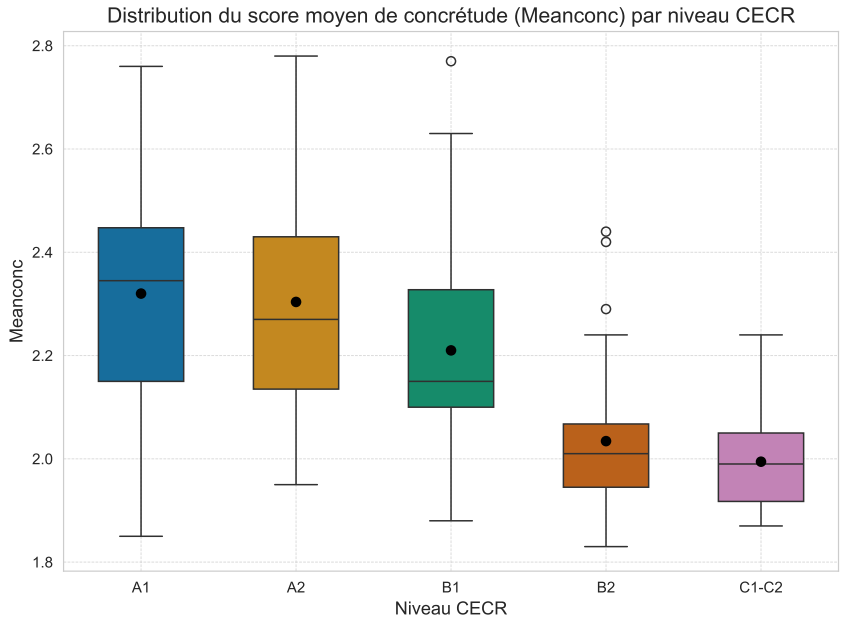

FIGURE 6.12 – Distribution des scores de concrétude par niveau CECR

dans le tableau 6.15 illustrent une diminution progressive et statistiquement robuste du score de concrétude moyen, passant de 2,32 au niveau A1 à 1,99 aux niveaux C1-C2, soit une réduction de 14,2% sur l'ensemble de l'échelle de compétence.

Les niveaux débutants (A1 et A2) présentent les scores de concrétude les plus élevés, avec des moyennes respectives de 2,32 (A1) et 2,30 (A2). Cette stabilité relative entre A1 et A2 (différence de seulement 0,02 point) suggère une continuité pédagogique dans l'utilisation d'un vocabulaire concret aux premiers stades d'apprentissage. L'écart-type au niveau A1 (0,24) indique une variabilité modérée, reflétant probablement l'introduction progressive de concepts légèrement plus abstraits même aux niveaux débutants.

La médiane au niveau A1 (2,34) est légèrement supérieure à la moyenne (2,32), suggérant une distribution légèrement asymétrique vers les valeurs inférieures, avec quelques textes présentant un vocabulaire exceptionnellement concret (maximum de 2,76). L'écart interquartile (Q1 = 2,15 ; Q3 = 2,45) reste relativement restreint, montrant une certaine

homogénéité du vocabulaire concret à ce niveau.

Le niveau B1 marque une transition notable avec une moyenne de 2,21, représentant une diminution de 3,9% par rapport au niveau A2. Cette réduction s'accompagne d'une diminution de l'écart-type (0,19), suggérant une plus grande homogénéité dans le degré d'abstraction des textes. La médiane (2,15) devient inférieure à la moyenne, indiquant un déplacement de la distribution vers des valeurs de concrétude plus faibles.

Une accélération marquée de l'abstraction lexicale se manifeste au niveau B2 avec une moyenne de 2,03, soit une diminution de 8,1% par rapport au niveau B1. Cette progression substantielle coïncide avec l'introduction de textes journalistiques dans le corpus. L'écart-type continue de diminuer (0,14), reflétant une convergence vers des standards lexicaux plus homogènes.

Les niveaux les plus avancés (C1-C2) présentent la moyenne la plus faible (1,99) avec l'écart-type le plus réduit (0,10), indiquant une stabilisation vers un vocabulaire majoritairement abstrait. L'écart interquartile très restreint (Q1 = 1,92 ; Q3 = 2,05) confirme l'homogénéité lexicale caractéristique des textes authentiques de niveau avancé.

Le boxplot révèle une progression visuelle claire de la diminution de la concrétude à travers les niveaux CECR. Les boîtes présentent une réduction progressive de leur hauteur, illustrant la diminution de la variabilité avec l'augmentation du niveau. Les médianes (lignes centrales des boîtes) suivent une trajectoire descendante régulière, sans chevauchement significatif entre les niveaux adjacents, témoignant de la robustesse de cette variable comme discriminateur de niveau.

On observe également une amplitude décroissante des boites, particulièrement marquée aux niveaux avancés. L'absence de valeurs aberrantes au niveau C1-C2 suggère une convergence vers des normes lexicales standardisées. Cette homogénéisation progressive reflète probablement l'influence croissante de textes authentiques aux niveaux supérieurs, caractérisés par des conventions stylistiques établies.

Ces résultats s'inscrivent parfaitement dans le cadre théorique établi par la psycholinguistique cognitive concernant l'effet de concrétude. Les travaux fondateurs de Paivio (1971) sur la théorie du double codage avaient déjà établi que les mots concrets, bénéficiant d'un codage à la fois verbal et imagé, sont traités plus facilement que les mots abstraits. Cette facilité de traitement explique leur prédominance aux niveaux débutants d'apprentissage des langues étrangères.

Analyse de la variance

Tableau 6.16 – Analyse de variance (ANOVA) du score moyen de concrétude par niveau CECR

Source de variation	F	p-value
Niveau CECR	25,17	1.85×10^{-16}

L'analyse de variance unidirectionnelle révèle un effet significatif du niveau CECR sur les scores moyens de concrétude ($F = 25{,}17$, $p = 1.85 \times 10^{-16}$). Cette valeur F substantielle, dépassant largement les seuils conventionnels de significativité, confirme que la concrétude

lexicale constitue un prédicteur robuste et fiable du niveau de compétence linguistique dans notre corpus.

6.3 Analyse des variables syntaxiques

6.3.1 Analyse de la longueur moyenne des phrases (en mots)

Statistiques descriptives

L'analyse de la longueur moyenne des phrases révèle une progression claire et statistiquement robuste à travers les niveaux du CECR. Les données présentées dans le tableau 6.17 illustrent une augmentation constante et régulière de la longueur moyenne des phrases du niveau A1 au niveau C1-C2, témoignant d'une complexification syntaxique progressive des textes destinés aux apprenants de néerlandais langue étrangère.

Niveau	Moyenne	Écart-type	Médiane	Q25	Q75	Min	Max
A1	8,79	2,47	8,78	7,27	10,20	4,07	15,18
A2	9,91	2,32	9,86	8,72	10,93	3,76	17,00
B1	11,29	2,98	10,99	9,73	13,64	5,33	17,31
B2	13,53	3,94	13,32	10,81	15,66	4,78	23,57
C1–C2	16,04	3,36	16,47	14,43	17,80	9,00	26,83

Tableau 6.17 – Statistiques descriptives de la longueur moyenne des phrases selon les niveaux CECR

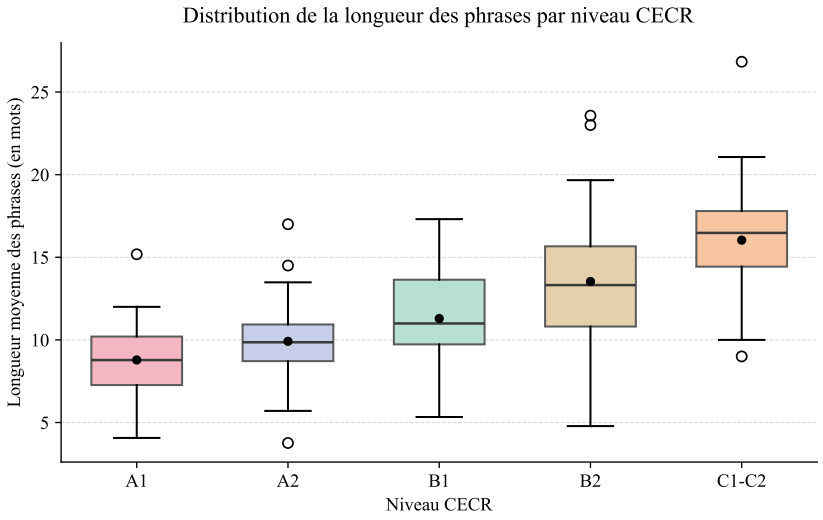

Figure 6.13 – Distribution de la longueur moyenne des phrases par niveau CECR

Au niveau A1, les textes présentent une longueur moyenne de phrases de 8,79 mots, avec un écart-type de 2,47, reflétant une relative homogénéité dans la simplicité syntaxique. La médiane (8,78) est pratiquement identique à la moyenne, suggérant une distribution

symétrique des valeurs. L'écart interquartile relativement restreint (Q1 = 7,27 ; Q3 = 10,20) confirme cette homogénéité, tandis que la valeur maximale de 15,18 mots indique que même aux niveaux débutants, certains textes peuvent présenter des phrases légèrement plus complexes.

La transition vers le niveau A2 marque une première augmentation modérée, avec une moyenne de 9,91 mots par phrase (+1,12 mots par rapport à A1, soit une augmentation de 12,7%). Cette progression s'accompagne d'une légère diminution de l'écart-type (2,32), suggérant une stabilisation relative de la variabilité syntaxique à ce niveau. La médiane (9,86) reste très proche de la moyenne, maintenant une distribution équilibrée. L'augmentation de la valeur maximale à 17,00 mots témoigne de l'introduction progressive de phrases plus complexes dans les textes de niveau A2.

Le niveau B1 constitue un palier intermédiaire notable avec une moyenne de 11,29 mots par phrase, représentant une augmentation de 1,38 mots par rapport au niveau A2 (+13,9%). Cette progression s'accompagne d'une augmentation de l'écart-type (2,98), indiquant une plus grande variabilité dans la longueur des phrases. Cette variabilité accrue reflète probablement la diversification des types textuels et des structures syntaxiques introduites à ce niveau de compétence. La médiane (10,99) demeure proche de la moyenne, mais l'écart interquartile s'élargit (Q1 = 9,73 ; Q3 = 13,64), confirmant cette diversification syntaxique.

Une accélération marquée de la complexification syntaxique se manifeste au niveau B2 avec une moyenne de 13,53 mots par phrase, soit une augmentation substantielle de 2,24 mots par rapport au niveau B1 (+19,8%). Cette progression importante coïncide avec l'introduction de textes journalistiques. L'écart-type atteint son maximum à ce niveau (3,94), reflétant la plus grande hétérogénéité syntaxique observée dans notre corpus. Cette variabilité maximale suggère une période de transition où coexistent des textes pédagogiques traditionnels et des documents authentiques présentant une complexité syntaxique variable.

Les niveaux les plus avancés (C1-C2) présentent la moyenne la plus élevée avec 16,04 mots par phrase, représentant une augmentation de 2,51 mots par rapport au niveau B2 (+18,5%). Cette progression massive confirme que la maîtrise avancée de structures syntaxiques complexes constitue un marqueur distinctif des niveaux de compétence supérieurs. Paradoxalement, l'écart-type diminue légèrement (3,36), suggérant une convergence vers des standards syntaxiques élevés mais relativement homogènes, caractéristiques des textes authentiques de niveau avancé. La médiane (16,47) dépasse légèrement la moyenne, indiquant une distribution légèrement asymétrique vers les valeurs supérieures.

L'analyse des valeurs extrêmes révèle des patterns particulièrement instructifs. Si les valeurs minimales restent relativement stables à travers les niveaux (entre 3,76 et 9,00 mots), témoignant de la persistance de phrases courtes même dans les textes avancés, les valeurs maximales montrent une progression dramatique, culminant à 26,83 mots au niveau C1-C2. Cette asymétrie dans la distribution suggère que les niveaux avancés se caractérisent par la présence de phrases particulièrement longues et complexes, tout en conservant également des phrases plus courtes pour maintenir la lisibilité globale du texte.

La distribution visualisée dans la figure 6.13 confirme ces observations quantitatives. Le boxplot révèle une progression visuelle claire et monotone de l'augmentation de la longueur des phrases à travers les niveaux CECR. Les boîtes présentent une expansion progressive de leur position sur l'axe vertical, illustrant l'augmentation constante de la longueur moyenne. Les médianes (lignes centrales des boîtes) suivent une trajectoire ascendante régulière, sans

chevauchement significatif entre les niveaux adjacents, témoignant de la robustesse de cette variable comme discriminateur de niveau.

L'amplitude des boîtes, représentant l'écart interquartile, suit un pattern intéressant : relativement stable aux niveaux débutants (A1-A2), elle s'élargit progressivement aux niveaux intermédiaires (B1-B2) avant de se stabiliser aux niveaux avancés (C1-C2). Cette évolution reflète la diversification syntaxique aux niveaux intermédiaires, suivie d'une convergence vers des normes stylistiques établies aux niveaux supérieurs.

Ces résultats s'inscrivent dans la continuité des observations de François (2011, pp. 406-407) concernant l'évolution de la longueur moyenne des phrases en français langue étrangère. François avait identifié la longueur moyenne des phrases (NMP) comme l'une des variables syntaxiques les plus discriminantes pour différencier les niveaux CECR.

Analyse de la variance

L'analyse de la variance unidirectionnelle confirme de manière statistiquement robuste l'existence d'un effet hautement significatif du niveau CECR sur la longueur moyenne des phrases. Les résultats présentés dans le tableau 18 révèlent une valeur F de 27,932 avec une p-value égale à 6.37×10^{-18}, dépassant largement tous les seuils conventionnels de significativité statistique. Cette analyse confirme de manière incontestable que la longueur moyenne des phrases constitue un prédicteur robuste et statistiquement significatif du niveau de difficulté textuelle en néerlandais langue étrangère.

Test	**F**	**p-value**
ANOVA (niveau CECR)	27,932	6.37×10^{-18}

Tableau 6.18 – Résultat de l'ANOVA sur la longueur moyenne des phrases selon le niveau CECR

6.3.2 Analyse du nombre de subordonnées moyen par texte

Statistiques descriptives

L'analyse du nombre de subordonnées révèle une progression significative et théoriquement cohérente à travers les niveaux CECR. Les données présentées dans le tableau 6.19 illustrent une augmentation générale mais non linéaire du nombre de subordonnées du niveau A1 au niveau C1-C2, témoignant d'une complexification syntaxique progressive mais différenciée selon les paliers de compétence.

Niveau CECR	**Moyenne**	**Écart-type**	**Médiane**	**Q25**	**Q75**	**Min**	**Max**
A1	8,41	6,80	7,00	3,25	12,75	0,00	25,00
A2	13,31	11,43	8,00	4,00	18,50	2,00	47,00
B1	18,80	14,71	16,50	6,00	25,00	0,00	52,00
B2	19,50	15,87	16,00	9,00	25,75	0,00	70,00
C1-C2	45,75	61,90	27,50	18,75	44,75	10,00	374,00

Tableau 6.19 – Statistiques descriptives pour la variable *NbSubord* par niveau CECR

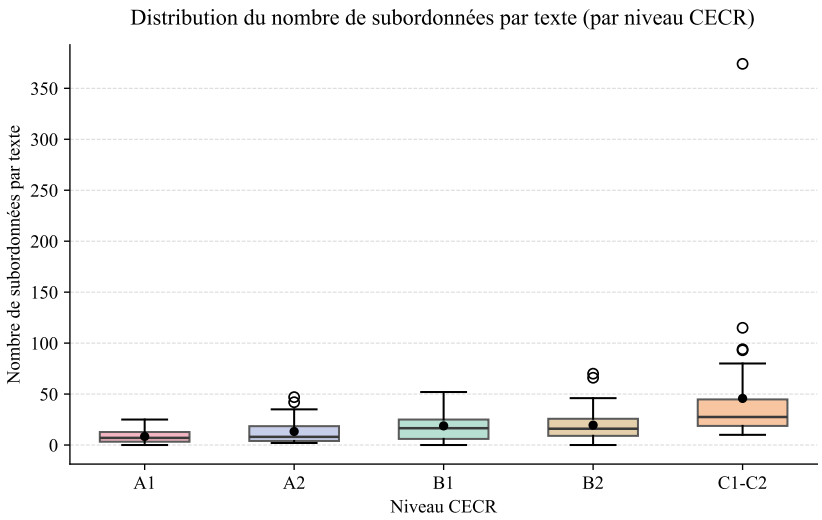

FIGURE 6.14 – Distribution du nombre de subordonnées moyen par niveau CECR

Au niveau A1, les textes présentent une moyenne de 8,41 subordonnées par texte, avec un écart-type de 6,80, reflétant une relative simplicité syntaxique caractéristique des textes destinés aux débutants. La médiane (7,00) est légèrement inférieure à la moyenne, suggérant une distribution légèrement asymétrique vers les valeurs supérieures, avec quelques textes présentant un nombre de subordonnées plus élevé. L'écart interquartile (Q1 = 3,25 ; Q3 = 12,75) révèle une variabilité modérée, tandis que la valeur maximale de 25,00 subordonnées indique que même aux niveaux débutants, certains textes peuvent introduire des structures syntaxiques plus complexes.

La transition vers le niveau A2 marque une première augmentation substantielle, avec une moyenne de 13,31 subordonnées par texte (+4,90 subordonnées par rapport à A1, soit une augmentation de 58,3%). Cette progression importante suggère que l'introduction des structures subordonnées constitue un marqueur crucial de la transition entre les niveaux élémentaires. L'augmentation de l'écart-type (11,43) témoigne d'une diversification des profils syntaxiques à ce niveau, reflétant probablement l'introduction progressive de différents types de subordonnées. La médiane (8,00) demeure inférieure à la moyenne, confirmant une distribution asymétrique avec une queue de distribution étendue vers les valeurs élevées.

Le niveau B1 constitue un palier intermédiaire significatif avec une moyenne de 18,80 subordonnées par texte, représentant une augmentation de 5,49 subordonnées par rapport au niveau A2 (+41,2%). Cette progression soutenue s'accompagne d'une augmentation continue de l'écart-type (14,71), indiquant une variabilité croissante dans l'utilisation des subordonnées. La médiane (16,50) se rapproche davantage de la moyenne, suggérant une distribution plus équilibrée. L'écart interquartile s'élargit considérablement (Q1 = 6,00 ; Q3 = 25,00), confirmant la diversification des profils syntaxiques à ce niveau de compétence.

Le niveau B2 présente une augmentation plus modeste avec une moyenne de 19,50 subordonnées par texte (+0,70 par rapport à B1, soit +3,7%). Cette stabilisation relative suggère un palier dans l'évolution de la subordination aux niveaux intermédiaires supérieurs. L'écart-type continue d'augmenter (15,87), témoignant d'une hétérogénéité persistante. La médiane (16,00) reste proche de celle du niveau B1, mais l'écart interquartile

s'élargit légèrement (Q1 = 9,00 ; Q3 = 25,75), indiquant une diversification continue des profils syntaxiques.

L'augmentation la plus spectaculaire s'observe entre B2 et C1-C2, avec un bond considérable de 26,25 subordonnées pour atteindre 45,75 subordonnées en moyenne (+134,6%). Cette progression massive confirme que la maîtrise avancée de la subordination constitue un marqueur distinctif des niveaux de compétence supérieurs.

L'écart-type atteint des proportions exceptionnelles au niveau C1-C2 (61,90), dépassant même la moyenne, indiquant une très grande hétérogénéité dans l'utilisation des subordonnées à ce niveau. Cette variabilité extrême suggère que les textes de niveau avancé peuvent présenter des profils syntaxiques très différents, allant de textes relativement simples à des textes extrêmement complexes sur le plan syntaxique. La médiane (27,50) demeure nettement inférieure à la moyenne, confirmant une distribution fortement asymétrique avec une queue de distribution très étendue.

L'analyse des valeurs extrêmes révèle des patterns particulièrement instructifs. Les valeurs minimales restent relativement basses à tous les niveaux, incluant des valeurs nulles aux niveaux B1 et B2, témoignant de la persistance de textes syntaxiquement simples même aux niveaux intermédiaires. Cette observation suggère que la progression syntaxique n'est pas uniforme et que certains types textuels ou certains domaines thématiques peuvent privilégier des structures syntaxiques plus simples, même aux niveaux avancés.

En revanche, les valeurs maximales montrent une progression dramatique, culminant à 374 subordonnées au niveau C1-C2. Cette valeur exceptionnellement élevée est relevée dans le texte *C2_ 7_ TIJD.txt* et peut s'expliquer par le fait que le texte est sensiblement plus long que les autres, avec 4074 mots pour 296 phrases (voir tableau en annexe).

Cette variable brute correspondant au nombre total de subordonnées dépend de la longueur totale du texte. C'est pourquoi la variable *PropSOV* apparaît plus pertinente pour capter cette dimension. En rapportant la fréquence d'une structure syntaxique spécifique à l'ensemble des unités textuelles, cette variable permet de distinguer les effets dus à une densité syntaxique propre de ceux qui ne sont qu'une conséquence de la longueur du document.

Analyse de la variance

L'analyse de variance confirme l'existence d'un effet significatif du niveau CECR sur le nombre de subordonnées ($F = 7{,}667$, $p = 1.03 \times 10^{-5}$). La valeur F de 7,667, bien que significative, est considérablement plus faible que celle observée pour la longueur moyenne des phrases ($F = 27{,}932$). Cette différence suggère que le nombre de subordonnées, tout en étant un indicateur valide de la complexité syntaxique, présente une plus grande variabilité intra-niveau et constitue un prédicteur moins stable du niveau de difficulté textuelle que la longueur des phrases.

Malgré cette variabilité plus importante, la significativité statistique de l'effet confirme que le nombre de subordonnées constitue néanmoins un indicateur valide de la complexité syntaxique en néerlandais langue étrangère. La robustesse de cet effet, maintenue malgré la variabilité intra-niveau, témoigne de l'importance fondamentale de la subordination dans la différenciation des niveaux de compétence linguistique. Cette variable doit toutefois être interprétée avec prudence, celle-ci ne normalisant pas le nombre de subordonnée par la longueur du texte.

Test	F	p-value
ANOVA (niveau CECR)	7,667	1.03×10^{-5}

Tableau 6.20 – Résultat de l'ANOVA sur le nombre de subordonnées par texte selon le niveau CECR

6.3.3 Analyse de la proportion de verbes à particules séparables

Statistiques descriptives

L'analyse de la proportion de verbes à particules séparables révèle un pattern de progression modeste mais théoriquement cohérent à travers les niveaux CECR. Les données présentées dans le tableau 6.21 illustrent une tendance générale à l'augmentation de ces constructions spécialisées du niveau A1 au niveau C1-C2, bien que cette progression demeure relativement limitée en amplitude comparativement à d'autres variables syntaxiques plus discriminantes.

Niveau CECR	Moyenne	Écart-type	Médiane	Q25	Q75	Min	Max
A1	0,061	0,066	0,052	0,008	0,089	0,000	0,198
A2	0,076	0,064	0,071	0,029	0,117	0,000	0,175
B1	0,079	0,065	0,066	0,034	0,114	0,000	0,252
B2	0,076	0,059	0,067	0,032	0,117	0,000	0,253
C1-C2	0,091	0,056	0,083	0,045	0,123	0,000	0,225

Tableau 6.21 – Statistiques descriptives pour la variable *PropSepVerb* par niveau CECR

Au niveau A1, les textes présentent une proportion moyenne de verbes à particules séparables de 0,061 (soit 6,1% des verbes), avec un écart-type de 0,066, reflétant une utilisation limitée mais variable de ces constructions complexes dans les textes destinés aux débutants. La médiane (0,052) est légèrement inférieure à la moyenne, suggérant une distribution asymétrique avec quelques textes présentant des proportions plus élevées. L'écart interquartile relativement large (Q1 = 0,008 ; Q3 = 0,089) témoigne d'une variabilité considérable dans l'usage de ces constructions, même aux niveaux débutants. La présence de valeurs nulles indique que certains textes évitent complètement ces structures complexes, privilégiant des constructions verbales plus simples et transparentes.

La transition vers le niveau A2 marque une première augmentation modérée, avec une moyenne de 0,076 (+0,015 par rapport à A1, soit une augmentation de 24,6%). Cette progression, bien que limitée en amplitude absolue, représente une augmentation relative qui suggère une introduction progressive de ces constructions spécialisées dans les textes de niveau élémentaire supérieur. L'écart-type demeure stable (0,064), indiquant une variabilité comparable au niveau précédent. La médiane (0,071) se rapproche de la moyenne, suggérant une distribution plus équilibrée. L'augmentation du premier quartile (Q1 = 0,029) témoigne d'une présence plus systématique de ces constructions, même dans les textes les moins complexes de ce niveau.

Le niveau B1 constitue un palier de stabilisation relative avec une moyenne de 0,079 verbes à particules séparables, représentant une augmentation minime de 0,003 par rapport au niveau A2 (+3,9%). Cette stagnation relative suggère que l'introduction de ces constructions complexes ne suit pas une progression linéaire simple, mais plutôt un développement par paliers. L'écart-type reste stable (0,065), confirmant une variabilité constante dans l'usage de ces structures. La médiane (0,066) demeure légèrement inférieure à la moyenne, maintenant une distribution asymétrique. L'écart interquartile (Q1 = 0,034 ; Q3 = 0,114) révèle une variabilité persistante, avec des textes présentant des profils d'usage très différents au sein du même niveau.

Le niveau B2 présente une légère diminution avec une moyenne de 0,076, soit un retour au niveau observé en A2. Cette régression apparente peut s'expliquer par plusieurs facteurs méthodologiques et pédagogiques. D'une part, l'introduction de textes journalistiques et académiques à ce niveau peut privilégier des registres de langue plus formels où ces constructions familières sont moins fréquentes. D'autre part, la diversification des genres textuels peut diluer la présence de ces structures spécialisées. L'écart-type diminue légèrement (0,059), suggérant une homogénéisation relative dans l'usage de ces constructions. La stabilité de la médiane (0,067) et de l'écart interquartile confirme cette tendance à l'homogénéisation.

Les niveaux les plus avancés (C1-C2) présentent la moyenne la plus élevée avec 0,091 verbes à particules séparables, représentant une augmentation de 0,015 par rapport au niveau B2 (+19,7%). Cette progression finale suggère que la maîtrise complète de ces constructions spécialisées constitue effectivement un marqueur des niveaux de compétence supérieurs, mais que cette maîtrise se manifeste de manière subtile et progressive. L'écart-type atteint son minimum à ce niveau (0,056), indiquant une convergence vers des normes d'usage plus homogènes. La médiane (0,083) se rapproche davantage de la moyenne, suggérant une distribution plus équilibrée.

L'analyse des valeurs extrêmes révèle des patterns particulièrement instructifs concernant la variabilité de l'usage de ces constructions. La persistance de valeurs nulles à tous les niveaux, y compris aux niveaux avancés, confirme que l'usage de verbes à particules séparables demeure optionnel et dépend fortement du genre textuel, du registre de langue et des choix stylistiques. Cette optionalité explique en partie la variabilité observée au sein de chaque niveau et la progression relativement modeste de cette variable.

Les valeurs maximales montrent une progression irrégulière, culminant à 0,253 au niveau B2 avant de redescendre à 0,225 au niveau C1-C2. Cette irrégularité suggère que les textes présentant les proportions les plus élevées de verbes à particules séparables ne se concentrent pas nécessairement aux niveaux les plus avancés, mais peuvent apparaître à différents niveaux selon les spécificités du corpus et les choix éditoriaux.

Analyse de la variance

L'analyse de la variance unidirectionnelle appliquée à la proportion de verbes à particules séparables révèle l'absence d'un effet statistiquement significatif du niveau CECR sur cette variable. Les résultats présentés dans le tableau 6.22 montrent une valeur F de 1,0022 avec une p-value de 0,408, dépassant largement le seuil conventionnel de significativité statistique de 0,05.

L'analyse de la variance nous indique que la proportion de verbes à particules séparables par rapport au nombre total de verbe ne constitue pas un prédicteur statistiquement signi-

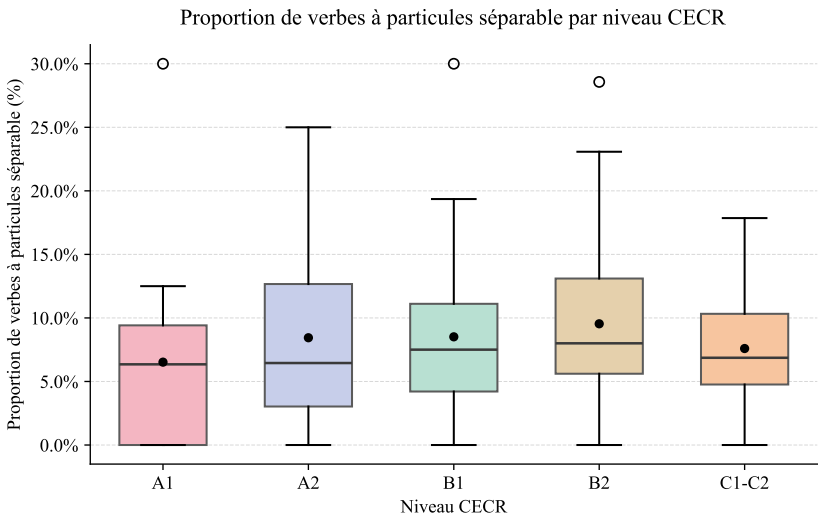

FIGURE 6.15 – Proportion de verbes à particules séparables par niveau CECR

ficatif du niveau de difficulté textuelle en néerlandais langue étrangère, du moins dans le contexte de notre corpus et de notre méthodologie. Ces résultats pourraient également être dus à un manque de performance de *spaCy* dans la reconnaissance de verbes à particule séparable.

Néanmoins, cette variable demeure théoriquement pertinente pour capturer des aspects spécifiques de la compétence en néerlandais et pourrait s'avérer utile dans des contextes d'analyse plus spécialisés ou en combinaison avec d'autres variables dans un modèle multivariable. L'originalité de cette analyse dans le contexte de la recherche en lisibilité du néerlandais contribue à élargir notre compréhension des dimensions linguistiques pertinentes pour l'évaluation automatique de la complexité textuelle dans le contexte du néerlandais L2.

Test	**F**	**p-value**
ANOVA (niveau CECR)	1,0022	0,408

TABLEAU 6.22 – Résultat de l'ANOVA sur la proportion de verbes à particule séparable par texte selon le niveau CECR

6.3.4 Analyse de la proportion de structures SOV par texte

Statistiques descriptives

L'analyse de la proportion de structures SOV par texte révèle une progression remarquablement claire et monotone à travers les niveaux CECR, constituant l'un des patterns de développement les plus cohérents observés dans notre étude des variables syntaxiques. Les données présentées dans le tableau 6.23 illustrent une augmentation systématique et substantielle qui reflète fidèlement la progression théoriquement attendue pour cette dimension spécifiquement néerlandaise de la complexité syntaxique.

Niveau CECR	Moyenne	Écart-type	Médiane	Q25	Q75	Min	Max
A1	0.136	0.124	0.114	0.063	0.187	0.000	0.545
A2	0.217	0.167	0.170	0.117	0.276	0.000	0.714
B1	0.229	0.157	0.192	0.118	0.300	0.056	0.769
B2	0.301	0.135	0.308	0.219	0.396	0.000	0.636
C1-C2	0.322	0.126	0.319	0.226	0.409	0.042	0.591

TABLEAU 6.23 – Statistiques descriptives pour la variable *PropSOV* par niveau CECR

Au niveau A1, les textes présentent une proportion moyenne de structures SOV de 0,136, soit 13,6% des structures analysées, avec un écart-type de 0,124, reflétant une utilisation limitée mais déjà présente de ces constructions complexes dans les textes destinés aux débutants. La médiane (0,114) est légèrement inférieure à la moyenne, suggérant une distribution asymétrique positive avec quelques textes présentant des proportions plus élevées que la tendance centrale. L'écart interquartile (Q1 = 0,063 ; Q3 = 0,187) révèle une variabilité modérée, indiquant que 50% des textes présentent des proportions comprises entre 6,3% et 18,7%, une fourchette relativement étroite qui témoigne d'une certaine homogénéité dans les approches pédagogiques au niveau débutant. La présence de valeurs nulles (Min = 0,000) indique que certains textes évitent complètement les structures SOV, privilégiant des constructions syntaxiques plus simples et plus transparentes pour les apprenants débutants. À l'inverse, la valeur maximale de 0,545 témoigne de l'existence de textes qui utilisent intensivement ces constructions.

La transition vers le niveau A2 marque une augmentation substantielle et statistiquement significative avec une moyenne de 0,217 (+0,081 par rapport à A1, soit une augmentation de 59,6%). Cette progression remarquable suggère une introduction plus systématique et intensive des structures SOV dans les textes de niveau élémentaire supérieur. L'écart-type augmente proportionnellement (0,167), indiquant une diversification des approches pédagogiques et une variabilité accrue dans l'usage de ces structures. La médiane (0,170) demeure légèrement inférieure à la moyenne, maintenant la distribution asymétrique positive observée au niveau précédent. L'écart interquartile s'élargit considérablement (Q1 = 0,117 ; Q3 = 0,276), témoignant d'une diversification des stratégies pédagogiques au niveau A2. La persistance de valeurs nulles indique que certains textes maintiennent une approche d'évitement, tandis que la valeur maximale atteint 0,714, suggérant l'existence de textes particulièrement riches en structures SOV.

Le niveau B1 présente une progression plus modérée avec une moyenne de 0,229 (+0,012 par rapport à A2, soit une augmentation de 5,5%). L'écart-type diminue légèrement (0,157), suggérant une homogénéisation relative dans l'usage de ces structures. La médiane (0,192) se rapproche de la moyenne, indiquant une distribution plus équilibrée et moins asymétrique. L'écart interquartile (Q1 = 0,118 ; Q3 = 0,300) révèle une variabilité maintenue mais mieux structurée, avec une concentration plus marquée autour de la tendance centrale. L'augmentation du minimum (0,056) indique que l'évitement complet des structures SOV devient moins fréquent à ce niveau, suggérant une reconnaissance croissante de leur importance pour la progression linguistique. La valeur maximale atteint 0,769, témoignant de l'existence de textes particulièrement sophistiqués du point de vue des structures SOV.

Le niveau B2 marque une accélération remarquable de la progression avec une moyenne de 0,301 (+0,072 par rapport à B1, soit une augmentation de 31,4%). Cette augmentation substantielle suggère une transition qualitative importante où les structures SOV deviennent un élément central de la complexité syntaxique des textes. Cette transition peut refléter l'introduction de textes plus authentiques et diversifiés qui présentent naturellement des fréquences plus élevées de ces constructions caractéristiques du néerlandais. L'écart-type diminue significativement (0,135), témoignant d'une convergence vers des normes d'usage plus homogènes. La médiane (0,308) devient très proche de la moyenne, indiquant une distribution remarquablement équilibrée et symétrique. L'écart interquartile (Q1 = 0,219 ; Q3 = 0,396) révèle une concentration accrue autour de la tendance centrale, avec 50% des textes présentant des proportions comprises entre 21,9% et 39,6%.

Les niveaux les plus avancés (C1-C2) présentent la progression la plus élevée avec une moyenne de 0,322 (+0,021 par rapport à B2, soit une augmentation de 7,0%). Cette progression finale, bien que plus modérée que la transition B1-B2, confirme la tendance générale à l'augmentation et positionne les structures SOV comme un marqueur robuste des niveaux de compétence supérieurs. Cette proportion élevée (32,2%) témoigne de l'importance cruciale de ces constructions dans les textes authentiques de niveau avancé. L'écart-type atteint son minimum à ce niveau (0,126), indiquant une homogénéisation maximale dans l'usage de ces structures. Cette convergence peut refléter l'émergence de normes d'usage stables dans les textes authentiques de niveau avancé, où les structures SOV apparaissent selon des fréquences naturelles déterminées par les caractéristiques intrinsèques de la langue néerlandaise plutôt que par des considérations pédagogiques artificielles. La médiane (0,319) est remarquablement proche de la moyenne, confirmant une distribution parfaitement équilibrée qui témoigne de la stabilité des patterns d'usage à ce niveau. L'écart interquartile (Q1 = 0,226 ; Q3 = 0,409) révèle une variabilité modérée et bien structurée, avec une concentration marquée autour de la tendance centrale. L'augmentation du minimum (0,042) confirme que l'évitement complet des structures SOV devient exceptionnel aux niveaux avancés, où ces constructions constituent un élément incontournable de la compétence syntaxique.

Analyse de la variance

L'analyse de la variance unidirectionnelle appliquée à la proportion de structures SOV par texte révèle un effet significatif du niveau CECR sur cette variable. Les résultats présentés dans le tableau 6.24 montrent une valeur F de 7,5987 avec une p-value de 1.15×10^{-5} , dépassant largement les seuils conventionnels de significativité statistique et témoignant d'un effet puissant et fiable.

Test	**F**	**p-value**
ANOVA (niveau CECR)	7,5987	1.15×10^{-5}

Tableau 6.24 – Résultat de l'ANOVA sur la proportion de structures SOV par phrase selon le niveau CECR

Ces résultats révèlent que l'analyse spécifique de la proportion de structures SOV constitue un prédicteur plus fin et plus discriminant que le simple comptage du nombre de subordonnées, capturant une dimension syntaxique spécifiquement néerlandaise qui reflète directement les défis cognitifs rencontrés par les apprenants francophones.

Néanmoins, il convient de maintenir une certaine prudence quant à l'interprétation de ces résultats, car des améliorations dans les méthodes d'extraction automatique des

structures SOV pourraient encore affiner la précision de cette variable et potentiellement modifier les patterns observés. Cette réserve méthodologique n'invalide pas la contribution significative de cette analyse, mais souligne l'importance de poursuivre le développement d'outils d'analyse syntaxique spécialisés pour optimiser l'évaluation automatique de la lisibilité en néerlandais langue étrangère.

6.3.5 Analyse de la proportion de connecteurs causaux par texte

Statistiques descriptives

L'analyse de la proportion de connecteurs causaux par texte révèle des valeurs remarquablement faibles à travers tous les niveaux CECR, avec des moyennes oscillant entre 0,000235 et 0,000564, soit entre 0,0235% et 0,0564% du contenu textuel total. Ces proportions extrêmement réduites témoignent de la rareté relative des connecteurs causaux explicites dans le corpus analysé.

Niveau CECR	Moyenne	Écart-type	Médiane	Q25	Q75	Min	Max
A1	0,000367	0,001719	0,000	0,000	0,000	0,000	0,008065
A2	0,000477	0,001671	0,000	0,000	0,000	0,000	0,009174
B1	0,000564	0,001731	0,000	0,000	0,000	0,000	0,008000
B2	0,000278	0,001277	0,000	0,000	0,000	0,000	0,006849
C1-C2	0,000235	0,000788	0,000	0,000	0,000	0,000	0,003492

Tableau 6.25 – Statistiques descriptives pour la variable *RatioConnCaus* par niveau CECR

Au niveau A1, les textes présentent une proportion moyenne de connecteurs causaux de 0,000367, soit environ 0,037% du contenu textuel, avec un écart-type considérable de 0,001719 qui témoigne d'une variabilité extrême dans l'usage de ces marqueurs discursifs. La médiane nulle (0,000) révèle que la majorité des textes de niveau A1 n'utilisent aucun connecteur causal explicite, stratégie cohérente avec les approches pédagogiques qui privilégient l'acquisition des structures de base avant l'introduction de marqueurs discursifs complexes. L'écart interquartile entièrement nul ($Q1 = 0{,}000$; $Q3 = 0{,}000$) confirme que au moins 75% des textes de niveau A1 ne contiennent aucun connecteur causal, soulignant le caractère exceptionnel de leur usage à ce niveau. La valeur maximale de 0,008065 (environ 0,81%) indique néanmoins l'existence de textes qui utilisent ces connecteurs de manière plus intensive, probablement dans des contextes narratifs ou explicatifs spécifiques où l'expression des relations causales devient nécessaire. Cette valeur maximale, bien que faible en termes absolus, représente une proportion relativement élevée par rapport à la moyenne du niveau,

La progression vers le niveau A2 révèle une augmentation modeste avec une moyenne de 0,000477 (+0,000110 par rapport à A1, soit une augmentation de 30,0%). Cette progression, bien que statistiquement modeste, peut refléter une introduction plus systématique des connecteurs causaux dans les textes de niveau élémentaire supérieur. L'écart-type

demeure élevé (0,001671), maintenant la variabilité importante observée au niveau précédent. La médiane reste nulle, indiquant que la majorité des textes continuent d'éviter ces connecteurs. L'écart interquartile demeure entièrement nul, confirmant que l'usage des connecteurs causaux reste exceptionnel même au niveau A2.

Le niveau B1 présente le pic de cette progression avec une moyenne de 0,000564 (+0,000087 par rapport à A2, soit une augmentation de 18,2%). L'écart-type atteint son maximum à ce niveau (0,001731), témoignant d'une diversification maximale des approches pédagogiques et des genres textuels. La persistance de la médiane nulle et de l'écart interquartile nul confirme que l'usage des connecteurs causaux demeure minoritaire même au niveau B1, suggérant que ces marqueurs ne constituent pas encore un élément central de la compétence discursive à ce niveau. La valeur maximale (0,008000) demeure dans la même fourchette que les niveaux précédents, suggérant une stabilisation des usages plus intensifs de ces connecteurs dans les textes spécialisés.

Le niveau B2 présente une diminution surprenante avec une moyenne de 0,000278 (-0,000286 par rapport à B1, soit une diminution de 50,7%). L'écart-type diminue proportionnellement (0,001277), suggérant une homogénéisation relative dans l'usage de ces connecteurs. Cette réduction de la variabilité peut refléter l'influence des conventions stylistiques des textes authentiques qui tendent à privilégier des stratégies cohésives plus subtiles et moins explicites. La médiane et l'écart interquartile demeurent nuls, confirmant que l'évitement des connecteurs causaux reste la norme même aux niveaux intermédiaires supérieurs. La valeur maximale diminue également (0,006849), suggérant que même les textes les plus intensifs dans leur usage des connecteurs causaux présentent des proportions plus modérées à ce niveau, probablement en raison de l'influence des normes stylistiques des textes authentiques.

Les niveaux les plus avancés (C1-C2) présentent la proportion la plus faible avec une moyenne de 0,000235 (-0,000043 par rapport à B2, soit une diminution supplémentaire de 15,5%). L'écart-type atteint son minimum à ce niveau (0,000788), témoignant d'une homogénéisation maximale dans l'usage de ces connecteurs. La médiane et l'écart interquartile demeurent nuls. La valeur maximale diminue considérablement (0,003492), suggérant que même les textes les plus riches en connecteurs causaux aux niveaux avancés présentent des proportions modérées.

Analyse de la variance

L'analyse de la variance unidirectionnelle appliquée à la proportion de connecteurs causaux par texte révèle l'absence d'un effet statistiquement significatif du niveau CECR sur cette variable. Les résultats présentés dans le tableau 6.26 montrent une valeur F de 0,3389 avec une p-value de 0,85144, dépassant très largement le seuil conventionnel de significativité statistique de 0,05 et indiquant l'absence d'effet systématique du niveau de compétence sur l'usage des connecteurs causaux.

Test	**F**	**p-value**
ANOVA (niveau CECR)	0,3389	0,85144

Tableau 6.26 – Résultat de l'ANOVA sur la proportion de connecteurs causaux selon le niveau CECR

L'absence de significativité statistique ne doit cependant pas être interprétée comme

une invalidation complète de l'importance des connecteurs causaux dans l'apprentissage du néerlandais langue étrangère. Cette observation souligne plutôt la complexité des relations entre compétence linguistique et usage des marqueurs discursifs, suggérant que l'acquisition de ces outils cohésifs suit des parcours développementaux différents de ceux observés pour les variables syntaxiques ou lexicales.

6.3.6 Analyse de la proportion de connecteurs contrastifs par texte

Statistiques descriptives

L'analyse de la proportion de connecteurs contrastifs par texte révèle une distribution particulière caractérisée par une absence totale de ces marqueurs aux niveaux débutants et une émergence progressive aux niveaux intermédiaires et avancés.

Niveau CECR	Moyenne	Écart-type	Médiane	Q25	Q75	Min	Max
A1	0.000000	0.000000	0.000	0.000	0.000000	0.000	0.000000
A2	0.000000	0.000000	0.000	0.000	0.000000	0.000	0.000000
B1	0.000305	0.001177	0.000	0.000	0.000000	0.000	0.005263
B2	0.000373	0.001269	0.000	0.000	0.000000	0.000	0.007264
C1-C2	0.001060	0.001698	0.000	0.000	0.001619	0.000	0.005263

Tableau 6.27 – Statistiques descriptives pour la variable *RatioConnContr* par niveau CECR

Les niveaux A1 et A2 présentent des valeurs strictement nulles pour tous les indicateurs statistiques (moyenne, écart-type, médiane, quartiles, minimum et maximum), indiquant une absence complète de connecteurs contrastifs dans les textes de ces niveaux débutants.

Le niveau B1 marque l'apparition des connecteurs contrastifs avec une moyenne de 0,000305 (0,0305%). L'écart-type de 0,001177 révèle une variabilité importante, tandis que la médiane nulle indique que la majorité des textes n'utilisent toujours pas ces connecteurs. Les quartiles Q25 et Q75 restent nuls, confirmant que leur usage demeure exceptionnel. La valeur maximale de 0,005263 (0,53%) témoigne de l'existence de quelques textes utilisant ces marqueurs de manière plus intensive.

Au niveau B2, la moyenne augmente légèrement à 0,000373 (+0,000068 par rapport à B1, soit +22,3%). L'écart-type augmente proportionnellement (0,001269), maintenant une variabilité élevée. La médiane et les quartiles demeurent nuls, mais la valeur maximale atteint 0,007264 (0,73%), indiquant une intensification de l'usage dans certains textes spécialisés.

Le niveau C1-C2 présente la progression la plus marquée avec une moyenne de 0,001060 (+0,000687 par rapport à B2, soit +184,2%). Cette augmentation substantielle s'accompagne d'un écart-type élevé (0,001698) et, pour la première fois, d'un troisième quartile non nul (Q75 = 0,001619), indiquant qu'au moins 25% des textes de niveau avancé utilisent des connecteurs contrastifs. La valeur maximale (0,005263) demeure dans la même fourchette que les niveaux précédents.

Analyse de la variance

L'analyse de la variance unidirectionnelle révèle un effet statistiquement significatif du niveau CECR sur la proportion de connecteurs contrastifs par texte. Les résultats présentent une valeur F de 5,0484 avec une p-value de 7.1342×10^{-4}, dépassant largement le seuil de significativité de 0,05.

La progression observée, caractérisée par une absence totale aux niveaux débutants (A1-A2) suivie d'une émergence progressive aux niveaux intermédiaires et avancés (B1-C2), suggère que l'usage des connecteurs contrastifs constitue un marqueur de sophistication discursive qui se développe avec l'augmentation de la compétence linguistique. Cette variable se distingue ainsi des autres connecteurs analysés par sa capacité à discriminer efficacement les niveaux de compétence CECR.

Test	**F**	**p-value**
ANOVA (niveau CECR)	5,0484	7.1342×10^{-4}

Tableau 6.28 – Résultat de l'ANOVA sur la proportion de connecteurs contrastifs selon le niveau CECR

6.4 Analyse de la régression linéaire multiple

Cette section constitue l'aboutissement de notre démarche analytique, visant à synthétiser les résultats obtenus lors des analyses univariées précédentes dans un modèle prédictif intégré. L'objectif principal est de quantifier la contribution relative de chaque variable linguistique à la prédiction du niveau de difficulté d'un texte en néerlandais langue étrangère, tout en évaluant la performance globale du modèle résultant. Cette approche multivariée nous permet de dépasser les limitations des analyses individuelles en tenant compte des interactions potentielles entre les différentes dimensions linguistiques et en identifiant les prédicteurs les plus discriminants dans un contexte d'analyse simultanée.

La régression linéaire multiple représente une méthode statistique particulièrement adaptée à notre problématique, car elle permet de modéliser la relation entre une variable dépendante continue (le niveau CECR numérisé) et un ensemble de variables explicatives (les caractéristiques linguistiques extraites). Cette approche nous offre la possibilité d'évaluer non seulement la significativité statistique de chaque prédicteur, mais également sa contribution unique à la variance expliquée, tout en contrôlant l'effet des autres variables incluses dans le modèle. De plus, la nature interprétable des coefficients de régression facilite la compréhension des mécanismes sous-jacents à la difficulté textuelle et permet d'établir des liens directs avec les théories linguistiques et psycholinguistiques qui sous-tendent notre cadre conceptuel.

6.4.1 Préparation des données et justification des inclusions/exclusions

La constitution du jeu de données pour la régression linéaire multiple a nécessité une série de décisions méthodologiques cruciales concernant l'inclusion ou l'exclusion de certaines variables. Ces choix, loin d'être arbitraires, reposent sur une analyse critique approfondie des résultats obtenus lors des analyses univariées et sur des considérations théoriques et méthodologiques solides. Cette section détaille les critères de sélection appliqués et justifie

chaque décision d'exclusion ou d'inclusion dans une perspective de rigueur scientifique et de validité méthodologique.

Exclusion des variables de fréquence lexicale

Quatre variables de fréquence lexicale ont été exclues de l'analyse de régression : *mean_fl, median_fl, p75_fl, et p90_fl*. Cette exclusion se fonde sur l'identification d'un effet de circularité méthodologique fondamental qui compromet la validité de ces mesures dans notre contexte d'analyse. La problématique réside dans le fait que les données de fréquence proviennent exclusivement des sous-collections "journaux" et "Belgique" du corpus SoNaR-500, tandis que notre corpus d'analyse présente une surreprésentation de textes de genre journalistique aux niveaux avancés (B2, C1-C2).

Cette circularité méthodologique génère un biais systématique qui se manifeste par des tendances contre-intuitives dans l'évolution de ces variables en fonction du niveau CECR. Ainsi, contrairement aux attentes théoriques qui prédisent une diminution de la fréquence lexicale moyenne avec l'augmentation du niveau de difficulté, nos analyses révèlent des patterns d'augmentation ou de stagnation qui reflètent davantage l'adéquation entre le genre textuel de la liste de référence et celui du corpus analysé que la véritable complexité lexicale des textes. Bien que la variable *median_fl* ait montré une certaine résistance à ce biais comparativement aux autres mesures de fréquence, nous avons néanmoins choisi de l'exclure par précaution du modèle de régression afin de maintenir une cohérence méthodologique stricte.

L'exclusion de ces variables s'avère donc nécessaire pour préserver la validité interne du modèle et éviter l'introduction de prédicteurs biaisés qui pourraient compromettre la généralisation des résultats.

Inclusion des variables lexicales basées sur les listes de référence

Les variables de proportion d'absents (PA) basées sur les listes SUBTLEX-NL ont été maintenues dans le modèle en raison de leur robustesse méthodologique et de leur pertinence théorique pour l'évaluation de la difficulté lexicale en néerlandais L2.

Les analyses univariées ont confirmé la pertinence de ces variables, montrant des progressions statistiquement significatives et théoriquement cohérentes avec l'augmentation du niveau CECR.

Inclusion des variables de diversité lexicale

La variable de diversité lexicale *MTLD* (Measure of Textual Lexical Diversity) a été maintenue dans le modèle en raison de sa capacité démontrée à discriminer les niveaux de compétence linguistique et de sa robustesse méthodologique face aux variations de longueur textuelle.

L'analyse univariée de cette variable révèle une progression statistiquement significative avec une tendance à l'augmentation de la diversité lexicale en fonction du niveau CECR. Cette tendance s'inscrit dans la logique développementale de l'apprentissage des langues, où l'augmentation de la compétence linguistique s'accompagne d'une capacité accrue à mobiliser un vocabulaire diversifié et nuancé.

Inclusion des variables de concrétude

La variable *MeanConc* (score moyen de concrétude) a été intégrée au modèle malgré des résultats univariés moins tranchés, en raison de son importance théorique dans les modèles de traitement lexical et de sa pertinence spécifique pour la lisibilité en L2. Cette variable, basée sur les normes de concrétude établies pour le néerlandais par Brysbaert et al. (2014), mesure le degré de tangibilité et d'accessibilité perceptuelle du vocabulaire utilisé dans les textes.

L'analyse univariée révèle une tendance à la diminution de la concrétude moyenne avec l'augmentation du niveau CECR, passant de 3,2891 au niveau A1 à 3,1466 au niveau C1-C2. Cette progression, bien que modeste en amplitude, présente une significativité statistique ($p < 0{,}05$) et s'inscrit dans la logique théorique qui prédit une abstraction croissante du vocabulaire avec l'augmentation du niveau de compétence linguistique.

Inclusion et exclusion des variables syntaxiques

Les variables syntaxiques retenues dans le modèle reflètent les spécificités structurelles du néerlandais et leur impact sur la difficulté de traitement pour les apprenants francophones. Quatre variables de cette catégorie ont été incluses : *MeanSentL* (longueur moyenne des phrases), *NbSubord* (nombre moyen de subordonnées par texte), *PropSepVerb* (proportion de verbes à particules séparables), et *PropSOV* (proportion de structures SOV).

La variable *MeanSentL* constitue un prédicteur classique et robuste de la complexité syntaxique. L'analyse univariée révèle une progression statistiquement significative avec une augmentation de la longueur moyenne des phrases de 11,58 mots au niveau A1 à 18,67 mots au niveau C1-C2. Cette progression reflète l'introduction progressive de structures syntaxiques plus complexes et d'enchâssements propositionnels plus sophistiqués dans les textes de niveaux supérieurs.

La variable *NbSubord* capture la complexité syntaxique liée à l'enchâssement propositionnel, phénomène particulièrement exigeant pour les apprenants de L2. Bien qu'elle doive être interprétée avec prudence étant donné qu'elle ne normalise pas le nombre de subordonnées par la longeur du texte, nous la maintenons ici, accompagnée de *PropSOV*, qui quantifie la proportion de structures à ordre SOV (Sujet-Objet-Verbe), caractéristique fondamentale du néerlandais dans les propositions subordonnées. Cette mesure présente un intérêt particulier pour les apprenants francophones, habitués à l'ordre SVO du français, et constitue un indicateur direct de la complexité syntaxique spécifiquement néerlandaise. L'analyse univariée confirme la pertinence de cette variable avec une progression statistiquement significative et une augmentation marquée du nombre de SOV avec le niveau CECR.

La variable *PropSepVerb* quant à elle, a été écartée du modèle en raison de sa faible capacité discriminante et de sa progression limitée à travers les niveaux CECR. Bien qu'elle cible une difficulté syntaxique réelle du néerlandais L2, son évolution non linéaire et sa variabilité élevée au sein des niveaux réduisent sa pertinence prédictive par rapport à d'autres indicateurs plus robustes. Son maintien risquait d'introduire du bruit dans la modélisation sans gain substantiel en pouvoir explicatif.

Les variables de connecteurs

Les variables de connecteurs causaux et contrastifs (*RatioConnCaus* et *RatioConnContr*) présentent des profils d'analyse contrastés qui ont conduit à des décisions d'inclusion différenciées.

La variable *RatioConnCaus* a été exclue du modèle en raison de l'absence de significativité statistique dans l'analyse univariée, témoignant de l'absence d'effet systématique du niveau CECR sur l'usage des connecteurs causaux. Cette absence de discrimination peut s'expliquer par la rareté relative de ces connecteurs dans le corpus analysé, avec des proportions extrêmement faibles à tous les niveaux. L'exclusion de cette variable se justifie également par des considérations méthodologiques liées à la distribution des données. La prédominance de valeurs nulles (médiane = 0 à tous les niveaux) et la forte variabilité des écarts-types suggèrent une distribution non normale qui pourrait compromettre la validité des analyses paramétriques. De plus, l'absence de progression systématique avec le niveau CECR indique que cette variable ne constitue pas un prédicteur fiable de la difficulté textuelle dans notre contexte d'analyse.

En revanche, la variable *RatioConnContr* a été maintenue dans le modèle en raison de sa capacité démontrée à discriminer les niveaux de compétence linguistique. L'analyse univariée révèle un effet statistiquement significatif, avec une progression caractérisée par une absence totale aux niveaux débutants (A1-A2) suivie d'une émergence progressive aux niveaux intermédiaires et avancés (B1-C2). Il faut tout de même l'interpréter avec prudence compte tenu de la faible présence de connecteurs contrastifs dans notre corpus.

Synthèse des variables retenues

Au terme de cette analyse critique, le modèle de régression linéaire multiple intègre un ensemble de douze variables linguistiques soigneusement sélectionnées pour leur pertinence théorique, leur robustese méthoddologique et leur capacité discriminante démontrée.

Tableau 6.29 – Synthèse des variables retenues pour la régression linéaire multiple

Type	**Sous-type**	**Variable**	**Mesure**
Lexicale	Proportion d'absents	PA_SUBTLEX_1500	Proportion d'absents de la sous-liste de fréquence de 1500 mots
Lexicale	Proportion d'absents	PA_SUBTLEX_3000	Proportion d'absents de la sous-liste de fréquence de 3000 mots
Lexicale	Proportion d'absents	PA_SUBTLEX_MAX	Proportion d'absents de la liste de fréquence complète
Lexicale	Proportion d'absents	PA_SUBLTEX_1500_U	Proportion d'absents uniques de la sous-liste de fréquence de 1500 mots
Lexicale	Proportion d'absents	PA_SUBLTEX_3000_U	Proportion d'absents uniques de la sous-liste de fréquence de 3000 mots

Tableau 6.29 – Synthèse des variables retenues (suite)

Type	Sous-type	Variable	Mesure
Lexicale	Proportion d'absents	PA_SUBLTEX_MAX_U	Proportion d'absents uniques de la liste de fréquence complète
Lexicale	Diversité lexicale	MTLD	Mesure de la diversité lexicale (alternative au TTR)
Lexicale	Concrétude	MeanConc	Moyenne des scores de concrétude des mots
Syntaxique	Longueur de phrases	MeanSentL	Longueur moyenne des phrases (en mots)
Syntaxique	Complexité hiérarchique	NbSubord	Nombre de subordonnées par texte
Syntaxique	Spécificités NL	PropSOV	Proportion de phrase contenant au moins une proposition subordonnée (ordre SOV)
Syntaxique	Connecteurs logiques	RatioConnContr	Proportion de connecteurs contrastifs

Encodage de la variable dépendante

La variable dépendante, correspondant ici aux niveaux CECR, a été encodée numériquement selon l'ordre croissant de difficulté : A1 = 0, A2 = 1, B1 = 2, B2 = 3, C1-C2 = 4. Cette approche ordinale respecte la hiérarchie naturelle des niveaux de compétence tout en permettant l'application de techniques de régression linéaire standard.

Division des données

Les données ont été divisées selon une approche de validation croisée standard, avec 80% des textes (143 textes) alloués à l'ensemble d'entraînement et 20% (36 textes) à l'ensemble de test. Cette division a été effectuée de manière stratifiée pour maintenir la représentativité de chaque niveau CECR dans les deux ensembles, garantissant ainsi la validité de l'évaluation de la performance du modèle.

6.4.2 Résultats de la régression linéaire multiple

Performance globale du modèle

L'analyse de régression linéaire multiple révèle des résultats encourageants qui confirment le potentiel prédictif des variables linguistiques sélectionnées pour la modélisation de la lisibilité du néerlandais langue étrangère. Le modèle obtient un coefficient de détermination R^2 de 0,7167 sur l'ensemble d'entraînement et de 0,6477 sur l'ensemble de test, indiquant que les variables retenues expliquent respectivement 71,67 % et 64,77 % de la variance dans la difficulté des textes.

Tableau 6.30 – Performances du modèle sur les ensembles d'entraînement et de test

Métrique	**Entraînement**	**Test**
R^2	0,7167	0,6477
RMSE	0,7120	0,7103
MAE	0,5434	0,4863

A titre de comparaison François (2011, p. 454) obtient un R^2 de 0,77 pour son modèle Expert1 en français langue étrangère, tandis que les études sur l'application des formules classiques en contexte L2 révèlent des performances variables : Brown (1998) cité par Crossley et al. (2008a)obtient des corrélations relativement faibles (0,48-0,55) pour les formules traditionnelles appliquées à l'anglais L2, confirmant la nécessité de développer des approches spécifiquement adaptées aux langues étrangères. Notre résultat de 64,77% sur l'ensemble de test témoigne donc d'une performance relativement solide et comparable aux standards internationaux, en particulier pour une première modélisation systématique de la lisibilité du néerlandais L2

L'erreur quadratique moyenne (RMSE) de 0,7103 sur l'ensemble de test indique que les prédictions du modèle s'écartent en moyenne de 0,71 niveau CECR par rapport aux valeurs réelles. Cette précision suggère que le modèle peut distinguer efficacement entre les niveaux adjacents, avec une marge d'erreur acceptable pour des applications pratiques d'évaluation de la difficulté textuelle.

Le MAE (0,4863) complète cette analyse en montrant que, sur la majorité des textes, l'erreur absolue moyenne est inférieure à un demi-niveau CECR. Cela signifie que les écarts les plus fréquents entre prédiction et réalité sont modestes, et que les erreurs importantes restent rares. Combiné au RMSE, ce résultat illustre une performance équilibrée, le modèle maintenant une précision satisfaisante tout en limitant les erreurs extrêmes.

Analyse des coefficients de régression

Tableau 6.31 – Coefficients de régression ordonnés par importance absolue

Variable	Coefficient	Coefficient_abs
RatioConnContr	94.8544	94.8544
PA_SUBTLEX_MAX_U	4.8827	4.8827
PA_SUBTLEX_3000_U	-2.5517	2.5517
PA_SUBTLEX_MAX	2.5211	2.5211
PA_SUBTLEX_3000	2.4138	2.4138
Meanconc	-1.0592	1.0592
PA_SUBTLEX_1500	0.8211	0.8211
PA_SUBTLEX_1500_U	0.6143	0.6143
PropSOV	0.5080	0.5080
MeanSentL	0.0317	0.0317
NbSubord	0.0027	0.0027
MTLD	0.0016	0.0016

Les variables lexicales dominent clairement le modèle, confirmant l'importance cruciale du facteur lexical dans la détermination de la difficulté textuelle pour les apprenants de néerlandais L2. Cette prédominance est cohérente avec les observations de François (2011) pour le français langue étrangère et reflète les défis particuliers que représente l'acquisition du vocabulaire en contexte L2.

La variable **PA_SUBTLEX_MAX_U** (coefficient : +4,8827) émerge comme le prédicteur lexical le plus puissant, indiquant que la proportion de mots absents uniques de la liste de fréquence maximale constitue un facteur déterminant de la difficulté. Cette variable capture la diversité des éléments lexicaux inconnus rencontrés par l'apprenant, au-delà de leur simple densité. L'effet positif confirme que l'exposition à un vocabulaire varié et peu fréquent augmente significativement la charge cognitive de traitement textuel.

L'effet négatif surprenant de **PA_SUBTLEX_3000_U** (coefficient : -2,5517) mérite une attention particulière. Cette observation contre-intuitive peut s'expliquer par plusieurs facteurs méthodologiques et linguistiques. D'une part, la nature hiérarchique des listes SUBTLEX peut créer des effets de suppression statistique où certaines variables présentent des coefficients inversés en raison de leur corrélation avec d'autres prédicteurs plus puissants. D'autre part, cette variable peut capturer des phénomènes linguistiques spécifiques au néerlandais, tels que l'impact des mots composés transparents qui, bien qu'absents des listes de fréquence en tant qu'unités lexicales distinctes, demeurent accessibles aux apprenants grâce à la transparence de leurs composants.

Les variables **PA_SUBTLEX_MAX** et **PA_SUBTLEX_3000** présentent des coefficients positifs modérés (+2,5211 et +2,4138 respectivement), confirmant l'effet attendu de la fréquence lexicale sur la difficulté. Ces résultats valident l'approche par listes de référence comme méthode robuste d'évaluation de la complexité lexicale en néerlandais L2.

La variable **Meanconc** (coefficient : -1,0592) présente un effet négatif significatif, confirmant que l'augmentation de la concrétude moyenne diminue la difficulté textuelle. Ce résultat est parfaitement cohérent avec la théorie du double codage de Paivio (1971). Les mots concrets, bénéficiant d'une représentation à la fois verbale et imagée, facilitent le traitement cognitif et réduisent la charge de la mémoire de travail.

Les variables syntaxiques, bien que présentant des coefficients plus modestes, apportent une contribution significative et complémentaire au modèle. La variable **PropSOV** (coefficient : +0,5080) confirme l'impact de la subordination sur la difficulté textuelle, validant les observations théoriques sur la complexité cognitive des structures hiérarchiques en néerlandais.

L'effet remarquablement faible de **MeanSentL**(coefficient : +0,0317) peut surprendre au regard de l'importance traditionnellement accordée à cette variable dans les formules de lisibilité classiques. Cette observation suggère que, dans le contexte spécifique du néerlandais L2, la longueur des phrases constitue un prédicteur moins discriminant que les facteurs lexicaux ou que d'autres dimensions de la complexité syntaxique. Cette relativisation de l'importance de la longueur des phrases peut refléter les spécificités structurelles du néerlandais, où la complexité syntaxique ne se résume pas à la longueur mais implique des phénomènes plus subtils liés à l'ordre des mots et aux dépendances à distance.

Le coefficient exceptionnellement élevé de **RatioConnContr** (+94,8544) mérite une analyse approfondie. Cette valeur, largement supérieure à tous les autres coefficients, peut s'expliquer par plusieurs facteurs méthodologiques et linguistiques.

D'un point de vue statistique, cette amplification peut résulter de la faible variance de cette variable dans le corpus, où la plupart des textes présentent des valeurs nulles ou très faibles. Dans de telles conditions, les rares occurrences de connecteurs contrastifs peuvent exercer un effet disproportionné sur la prédiction, conduisant à des coefficients artificiellement élevés.

D'un point de vue linguistique, ce résultat peut refléter l'importance qualitative des connecteurs contrastifs comme marqueurs de sophistication discursive. L'usage de ces connecteurs témoigne d'une capacité à structurer des arguments complexes et à exprimer des relations logiques nuancées, compétences caractéristiques des niveaux avancés. Leur rareté dans le corpus peut paradoxalement renforcer leur valeur prédictive, chaque occurrence constituant un signal fort de complexité textuelle.

Néanmoins, la prudence s'impose dans l'interprétation de ce coefficient, car sa magnitude peut également résulter d'un sur-ajustement aux particularités du corpus d'entraînement. Des études complémentaires sur des corpus plus larges et diversifiés seraient nécessaires pour confirmer la robustesse de cet effet.

6.4.3 Implications pour la modélisation de la lisibilité

Les résultats de cette analyse de régression multiple apportent plusieurs contributions importantes à la compréhension de la lisibilité laangue étrangère et ouvrent des perspectives prometteuses pour le développement futur d'ouils d'évaluation automatique.

Validation de l'approche multidimensionnelle

Les performances obtenues confirment la pertinence d'une approche multidimensionnelle intégrant des variables lexicales et syntaxiques pour la modélisation de la lsibilité. Cette

validation empirique soutient les critiques adressées aux formules traditionnelles, souvent limitées à quelques variables superficielles, et plaide pour le développement d'outils plus sophistiqués exploitant les avancées du traitement automatique du langage.

Spécificités du néerlandais langue étrangère

L'analyse révèle certaines spécificités du néerlandais L2 qui distinguent cette langue d'autres contextes étudiés dans la littérature. La prédominance des variables lexicales, bien que cohérente avec les observations générales sur l'apprentissage des L2, présente des nuances particulières liées aux caractéristiques morphologiques du néerlandais (mots composés, transparence morphologique).

La relativisation de l'importance de la longueur des phrases, traditionnellement considérée comme un prédicteur majeur, suggère que la complexité syntaxique du néerlandais ne se résume pas à des mesures quantitatives simples mais implique des phénomènes plus subtils liés à l'ordre des mots et aux structures spécifiques de cette langue.

Chapitre 7

Conclusion générale

Ce mémoire avait pour objectif principal de faire avancer la recherche sur la lisibilité en néerlandais langue étrangère, en explorant et testant le potentiel discriimnatoire de différentes variables linguistiques. A travers une approche méthodologique rigoureuse combinant les techniques du Traitement Automatique du Langage et l'analyse statistique, cette étude s'est attachée à identifier les facteurs linguistiques les plus pertinents pour évaluer le niveau de difficulté d'un texte destiné aux apprenants de néerlandais L2.

L'ambition de ce travail était double : d'une part, approfondir notre compréhension des mécanismes qui influencent la compréhension écrite en néerlandais langue étrangère ; d'autre part, jeter les bases méthodologiques nécessaires au développement futur d'outils d'évaluation automatique de la complexité textuelle. Cette recherche s'inscrit dans une démarche exploratoire qui vise à combler un vide dans la littérature scientifique, le néerlandais demeurant relativement peu exploré dans le domaine de la lisibilité en contexte L2, contrairement à l'anglais ou au français langue étrangère.

Les résultats obtenus témoignent de la complexité des phénomènes linguistiques en jeu dans l'évaluation de la difficulté textuelle et ouvrent des perspectives prometteuses pour la recherche future. Ce chapitre conclusif propose une synthèse des contributions apportées par cette étude, ainsi qu'une analyse critique de ses limites et des pistes d'amélioration qui en découlent.

7.1 Contribution de la recherche

Cette étude apporte plusieurs contributions significatives au domaine de la lisibilité en langue étrangère, tant sur le plan théorique que méthodologique. Ces contributions s'articulent autour de trois axes principaux qui enrichissent notre compréhension des facteurs de complexité textuelle en néerlandais L2 et ouvrent de nouvelles perspectives pour la recherche future.

7.1.1 Première modélisation systématique de la lisibilité du néerlandais langue étrangère

La contribution la plus fondamentale de ce travail réside dans le fait qu'il constitue la première modélisation systématique de la lisibilité du néerlandais en tant que langue étrangère. Alors que des études approfondies ont été menées sur la lisibilité de l'anglais et du français L2, notamment grâce aux travaux pionniers de François (2011), le néerlandais langue étrangère demeurait un territoire largement inexploré dans ce domaine spécifique. Cette lacune était d'autant plus regrettable que le néerlandais, avec plus de 24 millions de locuteurs natifs et son statut de langue officielle de l'Union européenne, représente un enjeu linguistique et pédagogique considérable, particulièrement dans le contexte belge.

Les résultats obtenus, avec un R^2 de 0,6477 sur l'ensemble de test, témoignent d'une performance relativement solide et comparable aux standards internationaux. Cette performance est d'autant plus remarquable qu'elle s'inscrit dans le cadre d'une première exploration systématique du domaine. Cette première modélisation établit donc un point de référence méthodologique solide pour les recherches futures et confirme la faisabilité d'une approche computationnelle pour l'évaluation de la lisibilité du néerlandais L2, ouvrant ainsi la voie à des développements ultérieurs plus sophistiqués.

7.1.2 Identification des spécificités du néerlandais langue étrangère

L'analyse révèle certaines spécificités du néerlandais L2 qui distinguent cette langue d'autres contextes étudiés dans la littérature, apportant ainsi une contribution originale à la compréhension des mécanismes de la lisibilité en contexte multilingue. Ces spécificités se manifestent à plusieurs niveaux et enrichissent notre compréhension des défis particuliers que représente l'apprentissage du néerlandais pour les locuteurs d'autres langues.

La prédominance des variables lexicales dans notre modèle confirme leur rôle central dans l'évaluation de la lisibilité du néerlandais L2, en cohérence avec les observations générales sur l'apprentissage des langues secondes. Cependant, certaines observations concernant l'impact des mots composés et de la transparence morphologique, spécificités structurelles du néerlandais, suggèrent des dynamiques complexes dans la perception de la difficulté textuelle. Il convient toutefois de noter que ces phénomènes pourraient résulter de corrélations avec d'autres variables linguistiques plutôt que d'effets directs de la composition morphologique. Cette observation indique néanmoins que les stratégies d'évaluation de la complexité lexicale gagneraient à intégrer davantage les particularités morphologiques du néerlandais, notamment la capacité des apprenants à décomposer et comprendre les mots composés transparents, tout en maintenant la robustesse des prédicteurs lexicaux traditionnels qui demeurent les plus discriminants.

La relativisation de l'importance de la longueur des phrases, traditionnellement considérée comme un prédicteur majeur dans les formules de lisibilité classiques, constitue une découverte particulièrement intéressante. Cette observation suggère que la complexité syntaxique du néerlandais ne se résume pas à des mesures quantitatives simples mais implique des phénomènes plus subtils liés à l'ordre des mots et aux structures spécifiques de cette langue. L'effet modéré mais significatif de la variable PropSOV (proportion de structures SOV) confirme l'importance des spécificités syntaxiques du néerlandais, où l'ordre des mots dans les subordonnées peut créer des dépendances à distance particulièrement challengeantes pour les apprenants.

Ces spécificités linguistiques soulignent la nécessité de développer des approches adaptées à chaque langue cible, plutôt que d'appliquer aveuglément des modèles conçus pour d'autres contextes linguistiques. Elles contribuent à une meilleure compréhension des défis cognitifs spécifiques que représente l'apprentissage du néerlandais et offrent des pistes précieuses pour l'adaptation des méthodes pédagogiques.

7.1.3 Développement de variables spécifiques au contexte néerlandais-français

Une contribution originale de cette recherche réside dans le développement et la validation de variables linguistiques spécifiquement adaptées au contexte d'apprentissage du néerlandais par des francophones. Cette adaptation contextuelle représente une avancée significative par rapport aux approches génériques qui ne tiennent pas compte des spécificités de l'interférence linguistique entre langues particulières.

L'introduction de variables syntaxiques spécifiques au néerlandais, telles que PropSOV (proportion de structures SOV) et PropSepVerb (proportion de verbes à particules séparables), témoigne d'une approche théoriquement informée qui prend en compte les particularités structurelles de la langue cible. Ces variables capturent des phénomènes linguistiques qui peuvent poser des défis particuliers aux apprenants francophones, habitués à des structures syntaxiques différentes.

La sélection théoriquement motivée de connecteurs causaux et contrastifs spécifiques aux difficultés des apprenants francophones constitue une autre innovation méthodologique. Plutôt que d'inclure tous les connecteurs de ces catégories, notre approche se fonde sur des critères théoriques précis : la complexité conceptuelle, l'absence d'équivalents directs en français, la fréquence d'usage discriminante, et la transparence sémantique réduite. Cette sélection ciblée, basée sur les travaux de Perrez (2006) et Degand et Pander Maat (2003), permet de capturer des nuances pragmatiques et sémantiques particulièrement challengeantes pour les apprenants francophones.

Cette contribution méthodologique illustre l'importance de développer des outils d'évaluation sensibles aux spécificités de l'interférence linguistique entre langues particulières. Elle ouvre la voie à des approches plus personnalisées de l'évaluation de la lisibilité, tenant compte du profil linguistique spécifique des apprenants.

7.2 Limites et pistes d'amélioration

Bien que cette étude apporte des contributions significatives au domaine de la lisibilité du néerlandais langue étrangère, il convient de reconnaître ses limites et d'identifier les pistes d'amélioration qui pourraient enrichir les recherches futures. Cette analyse critique permet de contextualiser les résultats obtenus et d'orienter les développements ultérieurs vers une compréhension encore plus fine des mécanismes de la lisibilité en contexte L2.

7.2.1 Limitations liées au corpus et à la méthodologie

Contraintes de taille et de diversité du corpus

Une première limitation de cette étude concerne la taille et la diversité du corpus utilisé pour l'entraînement du modèle. En effet, son extension et sa diversification pourraient

améliorer la robustesse et la généralisation des résultats obtenus. La constitution d'un corpus plus large permettrait notamment de mieux capturer la variabilité intra-niveau et d'affiner la discrimination entre les différents niveaux de compétence.

La diversité des sources et des genres textuels représente également un enjeu méthodologique important. Une plus grande diversité pourrait enrichir la modélisation en capturant des variations stylistiques et discursives plus étendues.

Regroupement des niveaux C1 et C2

Le regroupement des niveaux C1 et C2 pour l'analyse, bien que méthodologiquement justifié par la faible différenciation de ces niveaux en compréhension écrite, constitue une limitation qui mériterait d'être revisitée dans des études futures. Des recherches complémentaires avec des corpus plus larges pourraient permettre de maintenir la distinction entre ces niveaux avancés et d'identifier des variables discriminantes spécifiques aux compétences de très haut niveau.

Cette limitation soulève des questions importantes sur l'opérationnalisation de la difficulté aux niveaux les plus élevés du CECR. Les recherches futures pourraient explorer des variables plus sophistiquées, capables de capturer les nuances subtiles qui distinguent les textes de niveau C1 de ceux de niveau C2, notamment en termes de complexité conceptuelle, de densité informationnelle ou de sophistication stylistique.

Biais méthodologique

Cette étude présente certains biais méthodologiques qui peuvent influencer l'interprétation des résultats obtenus. Le principal biais identifié concerne la circularité méthodologique liée à l'utilisation des listes de fréquence, principalement basées sur des corpus journalistiques belges, alors que notre corpus d'analyse contient une forte proportion de textes de ce même genre, particulièrement aux niveaux avancés (B2, C1-C2). Cette circularité masque la véritable complexité lexicale des textes de niveaux supérieurs, qui peuvent véhiculer des informations complexes avec un vocabulaire de haute fréquence, conduisant ainsi à des résultats contre-intuitifs dans l'évaluation de la difficulté lexicale, ce qui nous a mené à exclure les variables de fréquence lexicale de notre modèle.

Un second biais concerne le risque de sur-ajustement aux particularités du corpus d'entraînement, comme l'illustre le coefficient exceptionnellement élevé observé pour la variable **RatioConnContr**. Ce phénomène peut conduire à une surestimation de l'importance de certains facteurs linguistiques et compromettre la généralisation des résultats à d'autres contextes d'apprentissage.

Pour remédier à ces biais, les recherches futures devraient diversifier les listes de fréquence en incluant des corpus didactiques ou de langue apprise, utiliser des techniques de validation croisée plus robustes, et tester les modèles sur des corpus indépendants et diversifiés afin de confirmer la stabilité et la généralisation des résultats obtenus.

7.2.2 Variables linguistiques non explorées

Potentiel des N-grammes pour l'analyse contextuelle

Une autre limitation de cette étude concerne la non-intégration de variables basées sur les N-grammes, séquences contiguës de N éléments qui constituent des outils puissants en

traitement automatique du langage pour capturer des informations contextuelles et des dépendances linguistiques dépassant le niveau du mot isolé. L'inclusion de telles variables a été limitée par des contraintes computationnelles, bien que l'interface N-grams d'OpenSonar permette d'extraire des modèles n-grammes de taille 2 à 5.

Cette limitation technique a empêché une exploration exhaustive de ces variables dans le cadre de cette recherche, mais représente une avenue prometteuse pour des études futures disposant de ressources informatiques plus importantes. Les recherches futures devraient prioritairement explorer cette dimension, car les N-grammes permettent de capturer des patterns linguistiques récurrents qui peuvent significativement influencer la perception de la difficulté par les apprenants.

Absence de variables d'interférence linguistique

Malgré leur pertinence théorique, l'intégration de variables spécifiques aux faux-amis et cognats néerlandais-français n'a pas été possible dans cette étude. Cette limitation représente une lacune importante, car ces phénomènes d'interférence linguistique jouent un rôle crucial dans la compréhension et la production en contexte L2. La proportion de faux-amis ou de cognats dans un texte pourrait constituer un indicateur pertinent de sa difficulté ou de sa facilité pour un apprenant francophone du néerlandais.

L'absence de ces variables s'explique par l'indisponibilité de listes exhaustives et validées de ces paires lexicales néerlandais-français. La création manuelle de telles ressources aurait dépassé le cadre temporel de ce mémoire, mais représente une piste de développement prioritaire pour les recherches futures. Il serait particulièrement intéressant que des études ultérieures se concentrent sur la compilation de ces listes et le développement de méthodes robustes pour leur extraction automatique.

Cette lacune souligne l'importance de développer des ressources linguistiques spécialisées pour l'analyse de l'interférence entre langues spécifiques. De telles ressources permettraient de créer des modèles de lisibilité encore plus précis et adaptés aux profils linguistiques particuliers des apprenants.

Absence de variables mesurant la transparence des mots composés

Une limitation méthodologique de cette étude concerne l'absence de variables spécifiquement conçues pour mesurer la transparence des mots composés néerlandais, phénomène linguistique pourtant central dans cette langue. Cette lacune représente une opportunité manquée d'explorer un mécanisme fondamental de la compréhension lexicale en néerlandais L2. La capacité des apprenants à décomposer et comprendre les mots composés transparents constitue une stratégie cognitive importante qui peut significativement influencer la perception de la difficulté textuelle. Des variables mesurant le degré de transparence morphologique, la fréquence des composants individuels, ou la prévisibilité sémantique des compositions auraient pu enrichir notre modèle et offrir une compréhension plus fine des mécanismes lexicaux en jeu.

Cette limitation souligne la nécessité pour les recherches futures de développer des outils computationnels spécialisés dans l'analyse de la morphologie compositionnelle du néerlandais, permettant ainsi une évaluation plus précise de l'impact des mots composés sur la lisibilité des textes destinés aux apprenants francophones.

7.2.3 Questions de robustesse statistique

Coefficient exceptionnel des connecteurs contrastifs

L'analyse des résultats révèle certaines anomalies statistiques qui méritent une attention particulière et soulèvent des questions sur la robustesse de certains coefficients. Le coefficient exceptionnellement élevé de *RatioConnContr* (+94,8544), largement supérieur à tous les autres coefficients, illustre cette problématique. Cette valeur peut s'expliquer par plusieurs facteurs méthodologiques et linguistiques qui nécessitent une interprétation prudente.

D'un point de vue statistique, cette amplification peut résulter de la faible variance de cette variable dans le corpus, où la plupart des textes présentent des valeurs nulles ou très faibles. Dans de telles conditions, les rares occurrences de connecteurs contrastifs peuvent exercer un effet disproportionné sur la prédiction, conduisant à des coefficients artificiellement élevés. Cette situation soulève des questions sur la représentativité du corpus et la nécessité de techniques de régularisation plus sophistiquées.

La prudence s'impose dans l'interprétation de ce coefficient, car sa magnitude peut également résulter d'un sur-ajustement aux particularités du corpus d'entraînement. Des études complémentaires sur des corpus plus larges et diversifiés seraient nécessaires pour confirmer la robustesse de cet effet et distinguer les phénomènes linguistiques authentiques des artefacts statistiques.

Effets contre-intuitifs

L'effet négatif surprenant de PA_SUBTLEX_3000_U (coefficient : -2,5517) mérite également une attention particulière, car cette observation contre-intuitive peut s'expliquer par plusieurs facteurs méthodologiques et linguistiques qui révèlent la complexité des interactions entre variables dans le modèle. D'une part, la nature hiérarchique des listes SUBTLEX peut créer des effets de suppression statistique où certaines variables présentent des coefficients inversés en raison de leur corrélation avec d'autres prédicteurs plus puissants.

D'autre part, cette variable peut capturer des phénomènes linguistiques spécifiques au néerlandais, tels que l'impact des mots composés transparents qui, bien qu'absents des listes de fréquence en tant qu'unités lexicales distinctes, demeurent accessibles aux apprenants grâce à la transparence de leurs composants. Cette observation souligne la nécessité d'approfondir l'analyse des interactions entre variables et de développer des méthodes d'interprétation plus sophistiquées pour les modèles multivariés complexes.

7.3 Perspectives

Cette recherche s'inscrit dans une dynamique scientifique plus large qui vise à améliorer notre compréhension des mécanismes de la compréhension écrite en langue étrangère et à développer des outils d'aide à l'apprentissage plus efficaces. Les résultats obtenus, bien qu'ils constituent une première exploration du domaine, ouvrent des perspectives prometteuses pour l'avenir de la recherche sur la lisibilité du néerlandais L2.

L'évolution rapide des technologies du Traitement Automatique du Langage, notamment avec l'émergence des modèles de langue de grande taille et des techniques d'apprentissage profond, offre des opportunités inédites pour l'amélioration des modèles de lisibilité.

Ces avancées technologiques pourraient permettre de capturer des dimensions encore plus subtiles de la complexité textuelle et d'adapter les évaluations aux profils spécifiques des apprenants.

La dimension applicative de cette recherche ne doit pas être négligée. Si ce travail demeure fondamentalement exploratoire, il jette les bases nécessaires au développement futur d'outils opérationnels qui pourraient transformer les pratiques pédagogiques. L'automatisation de l'évaluation de la lisibilité pourrait permettre aux enseignants de sélectionner plus efficacement des textes adaptés au niveau de leurs apprenants, contribuant ainsi à optimiser les parcours d'apprentissage et à personnaliser l'enseignement.

Cette recherche contribue également à une réflexion plus large sur l'apport des technologies numériques à l'enseignement et à l'apprentissage des langues. Elle illustre comment la recherche fondamentale en linguistique computationnelle peut nourrir des applications pratiques susceptibles d'améliorer l'efficacité pédagogique et de répondre aux défis contemporains de l'enseignement des langues étrangères.

En définitive, cette étude constitue une première pierre dans l'édification d'une compréhension scientifique approfondie de la lisibilité du néerlandais langue étrangère. Elle ouvre un champ de recherche prometteur qui, nous l'espérons, suscitera l'intérêt de la communauté scientifique et contribuera à l'amélioration des pratiques d'enseignement et d'apprentissage du néerlandais. Les défis identifiés et les pistes d'amélioration proposées offrent un programme de recherche riche et stimulant pour les années à venir, au service d'une meilleure compréhension des mécanismes de l'acquisition linguistique en contexte multilingue.

Chapitre 8

Annexes

Annexe 1

Tableau 8.1 – Tableau récapitulatif du corpus

Fichier	Niveau	Source	Nb. mots	Nb. phrases
A1_1.txt	A1	lingua.com	156	21
A1_2.txt	A1	lingua.com	180	27
A1_3.txt	A1	lingua.com	160	19
A1_4.txt	A1	lingua.com	175	25
A1_5_B.txt	A1	lingua.com	47	6
A1_6_B.txt	A1	lingua.com	107	18
A1_7_B.txt	A1	lingua.com	103	23
A1_8.txt	A1	lingua.com	164	11
A1_8.txt	A1	lingua.com	164	11
A1_9.txt	A1	lingua.com	114	14
A1_11_NT2.txt	A1	nt2taalmenu	152	17
A1_12_NT2.txt	A1	nt2taalmenu	95	10
A1_13_NT2.txt	A1	nt2taalmenu	86	8
A1_14_NT2.txt	A1	nt2taalmenu	125	13
A1_15_NT2.txt	A1	nt2taalmenu	181	17
A1_16_NT2.txt	A1	nt2taalmenu	158	17
A1_17_NT2.txt	A1	nt2taalmenu	134	11
A1_18_NT2.txt	A1	nt2taalmenu	245	20
A1_19_NT2.txt	A1	nt2taalmenu	262	31
A1_20_NT2.txt	A1	nt2taalmenu	185	21
A1_21_NT2.txt	A1	nt2taalmenu	90	10
A1_22_NT2.txt	A1	nt2taalmenu	196	22
A2_1.txt	A2	lingua.com	204	25
A2_2.txt	A2	lingua.com	146	18
A2_3.txt	A2	lingua.com	193	20
A2_4.txt	A2	lingua.com	154	22
A2_5.txt	A2	lingua.com	151	23
A2_6.txt	A2	lingua.com	167	10
A2_7.txt	A2	lingua.com	172	10
A2_8.txt	A2	lingua.com	229	27
A2_9.txt	A2	lingua.com	184	14
A2_10.txt	A2	lingua.com	221	24
A2_11.txt	A2	lingua.com	104	12
A2_12.txt	A2	lingua.com	120	11
A2_13.txt	A2	lingua.com	130	14
A2_14.txt	A2	lingua.com	128	13
A2_15.txt	A2	lingua.com	125	12

Suite en page suivante

Tableau 8.1 – Corpus des textes analysés (suite)

Fichier	Niveau	Source	Nb. mots	Nb. phrases
A2_16.txt	A2	lingua.com	125	11
A2_17.txt	A2	lingua.com	125	18
A2_18.txt	A2	lingua.com	123	9
A2_19.txt	A2	lingua.com	132	11
A2_20_NT2.txt	A2	nt2taalmenu	268	29
A2_21_NT2.txt	A2	nt2taalmenu	220	24
A2_22_NT2.txt	A2	nt2taalmenu	301	29
A2_23_NT2.txt	A2	nt2taalmenu	311	37
A2_24_NT2.txt	A2	nt2taalmenu	418	40
A2_25_NT2.txt	A2	nt2taalmenu	295	27
A2_26_NT2.txt	A2	nt2taalmenu	294	25
A2_27_NT2.txt	A2	nt2taalmenu	203	25
A2_28_NT2.txt	A2	nt2taalmenu	325	42
A2_29_NT2.txt	A2	nt2taalmenu	218	25
A2_30_NT2.txt	A2	nt2taalmenu	391	41
A2_31_NT2.txt	A2	nt2taalmenu	253	28
A2_32_CNAVT.txt	A2	cnavt	135	16
A2_33_CNAVT.txt	A2	cnavt	612	48
A2_34_CNAVT.txt	A2	cnavt	58	8
A2_35_CNAVT.txt	A2	cnavt	52	7
A2_36_CNAVT.txt	A2	cnavt	49	6
A2_37_CNAVT.txt	A2	cnavt	68	7
A2_38_CNAVT.txt	A2	cnavt	311	32
A2_39_CNAVT.txt	A2	cnavt	362	30
B1_1.txt	B1	lingua.com	205	22
B1_2.txt	B1	lingua.com	171	27
B1_3.txt	B1	lingua.com	218	27
B1_4.txt	B1	lingua.com	259	36
B1_5.txt	B1	lingua.com	204	25
B1_6.txt	B1	lingua.com	179	27
B1_7.txt	B1	lingua.com	204	27
B1_8.txt	B1	lingua.com	147	18
B1_9.txt	B1	lingua.com	221	12
B1_10.txt	B1	lingua.com	111	9
B1_11.txt	B1	lingua.com	191	12
B1_12.txt	B1	lingua.com	120	10
B1_13.txt	B1	lingua.com	121	11
B1_14.txt	B1	lingua.com	186	20
B1_15.txt	B1	lingua.com	251	15
B1_16.txt	B1	lingua.com	146	7
B1_17.txt	B1	lingua.com	262	17
B1_18_CNAVT.txt	B1	cnavt	515	36
B1_19_CNAVT.txt	B1	cnavt	285	26
B1_20_CNAVT.txt	B1	cnavt	590	46
B1_21_CNAVT.txt	B1	cnavt	160	11
B1_22_NT2.txt	B1	nt2taalmenu	515	46
B1_23_NT2.txt	B1	nt2taalmenu	287	25
B1_24_NT2.txt	B1	nt2taalmenu	332	33
B1_25_NT2.txt	B1	nt2taalmenu	383	39
B1_26_NT2.txt	B1	nt2taalmenu	368	37
B1_27_NT2.txt	B1	nt2taalmenu	334	34
B1_28_NT2.txt	B1	nt2taalmenu	261	27
B1_29_NT2.txt	B1	nt2taalmenu	495	41
B1_30_NT2.txt	B1	nt2taalmenu	293	24
B1_31_NT2.txt	B1	nt2taalmenu	563	39
B1_32_NT2.txt	B1	nt2taalmenu	628	49
B1_33_NT2.txt	B1	nt2taalmenu	294	38
B1_34_NT2.txt	B1	nt2taalmenu	382	27
B1_35_NT2.txt	B1	nt2taalmenu	511	37
B1_36_NT2.txt	B1	nt2taalmenu	658	69
B1_37_NT2.txt	B1	nt2taalmenu	837	62
B1_38_NT2.txt	B1	nt2taalmenu	815	55

Suite en page suivante

TABLEAU 8.1 – Corpus des textes analysés (suite)

Fichier	Niveau	Source	Nb. mots	Nb. phrases
B1_39_NT2.txt	B1	nt2taalmenu	498	34
B1_40_NT2.txt	B1	nt2taalmenu	1009	86
B2_1.txt	B2	lingua.com	160	26
B2_2.txt	B2	lingua.com	180	17
B2_3_COMPROF.txt	B2	Vers une communication professionnelle	514	29
B2_4_COMPROF.txt	B2	Vers une communication professionnelle	524	25
B2_5_COMPROF.txt	B2	Vers une communication professionnelle	327	21
B2_6_COMPROF.txt	B2	Vers une communication professionnelle	303	19
B2_7_COMPROF.txt	B2	Vers une communication professionnelle	276	12
B2_8_COMPROF.txt	B2	Vers une communication professionnelle	414	18
B2_9_COMPROF.txt	B2	Vers une communication professionnelle	351	18
B2_10_COMPROF.txt	B2	Vers une communication professionnelle	415	30
B2_11_COMPROF.txt	B2	Vers une communication professionnelle	494	30
B2_12_COMPROF.txt	B2	Vers une communication professionnelle1	488	29
B2_13_COMPROF.txt	B2	Vers une communication professionnelle	350	17
B2_14_COMPROF.txt	B2	Vers une communication professionnelle	414	26
B2_15_COMPROF.txt	B2	Vers une communication professionnelle	478	23
B2_16_COMPROF.txt	B2	Vers une communication professionnelle	206	16
B2_17_COMPROF.txt	B2	Vers une communication professionnelle	158	6
B2_18_COMPROF.txt	B2	Vers une communication professionnelle	257	14
B2_19_COMPROF.txt	B2	Vers une communication professionnelle	362	18
B2_20_COMPROF.txt	B2	Vers une communication professionnelle	309	15
B2_21_COMPROF.txt	B2	Vers une communication professionnelle	399	20
B2_22_COMPROF.txt	B2	Vers une communication professionnelle	146	8

Suite en page suivante

Tableau 8.1 – Corpus des textes analysés (suite)

Fichier	Niveau	Source	Nb. mots	Nb. phrases
B2_23_COMPROF.txt	B2	Vers une communication professionnelle	146	7
B2_24.txt	B2	lingua.com	126	10
B2_25.txt	B2	lingua.com	130	15
B2_26.txt	B2	lingua.com	123	11
B2_27.txt	B2	lingua.com	131	13
B2_28.txt	B2	lingua.com	127	11
B2_29.txT	B2	lingua.com	109	8
B2_30_CNAVT.txt	B2	cnavt	278	17
B2_31_CNAVT.txt	B2	cnavt	92	6
B2_32_CNAVT.txt	B2	cnavt	627	48
B2_33_CNAVT.txt	B2	cnavt	214	14
B2_34_CNAVT.txt	B2	cnavt	416	25
B2_35_CNAVT.txt	B2	cnavt	200	14
B2_36_CNAVT.txt	B2	cnavt	153	12
B2_37_CNAVT.txt	B2	cnavt	130	5
B2_38_CNAVT.txt	B2	cnavt	118	10
B2_39_CNAVT.txt	B2	cnavt	107	8
B2_41_CNAVT.txt	B2	cnavt	833	53
B2_42_CNAVT.txt	B2	cnavt	961	52
C1_CNAVT1.txt	C1	cnavt	312	17
C1_CNAVT2.txt	C1	cnavt	327	20
C1_CNAVT3.txt	C1	cnavt	625	38
C1_CNAVT4.txt	C1	cnavt	230	12
C1_CNAVT5.txt	C1	cnavt	436	23
C1_CNAVT6.txt	C1	cnavt	379	21
C2_1_TIJD.txt	C2	tijd.be	491	25
C2_2_TIJD.txt	C2	tijd.be	441	36
C2_3_TIJD.txt	C2	tijd.be	1044	56
C2_4_TIJD.txt	C2	tijd.be	301	16
C2_5_TIJD.txt	C2	tijd.be	312	15
C2_6_TIJD.txt	C2	tijd.be	533	33
C2_7_TIJD.txt	C2	tijd.be	4074	296
C2_8_TIJD.txt	C2	tijd.be	188	10
C2_9_TIJD.txt	C2	tijd.be	319	21
C2_10_TIJD.txt	C2	tijd.be	368	20
C2_11_TIJD.txt	C2	tijd.be	210	22
C2_12_TIJD.txt	C2	tijd.be	502	29
C2_13_TIJD.txt	C2	tijd.be	195	10
C2_14_TIJD.txt	C2	tijd.be	655	45
C2_15_TIJD.txt	C2	tijd.be	453	28
C2_16_TIJD.txt	C2	tijd.be	587	27
C2_17_TIJD.txt	C2	tijd.be	879	48
C2_18_TIJD.txt	C2	tijd.be	324	13
C2_19_TIJD.txt	C2	tijd.be	414	23
C2_20_TIJD.txt	C2	tijd.be	869	47
C2_21_TIJD.txt	C2	tijd.be	1220	64
C2_22_TIJD.txt	C2	tijd.be	318	19
C2_23_TIJD.txt	C2	tijd.be	664	33
C2_24_TIJD.txt	C2	tijd.be	2369	117
C2_25_TIJD.txt	C2	tijd.be	416	27
C2_26_TIJD.txt	C2	tijd.be	827	43
C2_27_TIJD.txt	C2	tijd.be	1291	101
C2_28_TIJD.txt	C2	tijd.be	402	19
C2_29_TIJD.txt	C2	tijd.be	872	71
C2_30_TIJD.txt	C2	tijd.be	760	66

Annexe 2

Code source du mémoire (GitHub) : `https://github.com/VictorineColin2002/MEMOIRE_FINAL_2025`

Bibliographie

Bailin, A. et Grafstein, A. (2001). "The linguistic assumptions underlying readability formulae : a critique". In : *Language Communication* 21.3, p. 285-301.

Barberán, S. X. et al. (2024). "Reading Comprehension and Cognitive Processes in Ecuadorian Middle School Students". In : *Journal of Ecohumanism* 3.8, p. 6305-6314. issn : 2752-6798. doi : 10.62754/joe.v3i8.5232. url : https://ecohumanism.co.uk/joe/ecohumanism.

Bernhardt, E. B. (2011). *Understanding Advanced Second-Language Reading*. New York : Routledge. isbn : 978-0-203-85240-8.

Bernhardt, E. B. (2005). "Progress and procrastination in second language reading". In : *Annual Review of Applied Linguistics* 25, p. 133-150. doi : 10.1017/S0267190505000073.

Bestgen, Y. (2004). *Analyse sémantique latente et segmentation automatique des textes*. Technical Report. FNRS – UCL/PSOR.

Biber, D. (1993). "Representativeness in corpus design". In : *Literary and Linguistic Computing* 8.4, p. 243-257. doi : 10.1093/llc/8.4.243. url : https://pure.strath.ac.uk/ws/portalfiles/portal/64389598/strathprints002381.pdf.

Biber, D. et Conrad, S. (2009). *Register, Genre, and Style*. Cambridge : Cambridge University Press.

Bormuth, J. (1969). *Development of Readability Analysis*. Technical Report Project No. 70052. Washington, DC : U.S. Office of Education, Bureau of Research, Department of Health, Education et Welfare.

Bormuth, J. R. (1966). "Readability : A New Approach". In : *Reading Research Quarterly* 1.3, p. 79-132. doi : 10.2307/747021. url : https://doi.org/10.2307/747021.

Bosmans, H. et al. (2020). *Néerlandais B2 - Vers une communication professionnelle : Économie - gestion - commerce - communication*. De Boeck Supérieur.

Brenders, P., Van Hell, J. G. et Dijkstra, T. (2011). "Word recognition in child second language learners : Evidence from cognates and false friends". In : *Journal of Experimental Child Psychology* 109.3, p. 334-350. doi : 10.1016/j.jecp.2010.12.006.

Brouwer, R. H. M. (1963). "Onderzoek naar de leesmoeilijkheden van Nederlands proza". In : *Pedagogische Studiën* 40, p. 454-464.

Brysbaert, M. et al. (2014). "Norms of age of acquisition and concreteness for 30,000 Dutch words". In : *Acta Psychologica* 150, p. 80-84. doi : 10.1016/j.actpsy.2014.04.010.

Brysbaert, M. et New, B. (2009). "Moving beyond Kucera and Francis : A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English". In : *Behavior Research Methods* 41.4, p. 977-990. doi : 10.3758/BRM.41.4.977.

Brysbaert, M., Van Wijnendaele, I. et Speybroeck, S. V. (2000). "The effects of word frequency and age of acquisition on word recognition : Evidence from Dutch". In : *Journal of Experimental Psychology : Learning, Memory, and Cognition* 26.1, p. 61-75.

Carrell, P. (1987). "Readability in ESL". In : *Reading in a Foreign Language* 4.1, p. 21-40.

Chall, J. et Dale, E. (1995). *Readability Revisited : The New Dale-Chall Readability Formula*. Cambridge, MA : Brookline Books.

Chuquet, H. et Paillard, M. (1987). *Approche linguistique des problèmes de traduction, anglais-français*. Paris : Ophrys.

Cito, Centraal Instituut voor Toets Ontwikkeling (2023). *Cito website*. https://www.cito.nl/. Accessed : 2023-04-29.

CNaVT (2025). *Certificaat Nederlands als Vreemde Taal*. Consulté en juin 2025. url : https://cnavt.org/.

Cohen, J. (1960). "A coefficient of agreement for nominal scales". In : *Educational and Psychological Measurement* 20.1, p. 37-46. doi : 10.1177/001316446002000104.

Collard, C., Przybyl, H. et Defrancq, B. (juin 2019). "Interpreting into an SOV Language : Memory and the Position of the Verb. A Corpus-Based Comparative Study of Interpreted and Non-mediated Speech". In : *Meta : Journal des traducteurs* 63.3. doi : 10.7202/1060169ar. url : https://doi.org/10.7202/1060169ar.

Council of Europe (2001). *Common European Framework of Reference for Languages : Learning, teaching, assessment (CEFR)*. Cambridge University Press.

Crossley, S. et al. (2023). "A large-scaled corpus for assessing text readability". In : *Behavior Research Methods* 55.2, p. 491-507. url : https://link.springer.com/article/10.3758/s13428-022-01802-x.

Crossley, S. A., Greenfield, J. et McNamara, D. S. (2008a). "Assessing text readability using cognitively based indices". In : *TESOL Quarterly* 42.3, p. 475-493. url : https://www.jstor.org/stable/40264479.

Crossley, S. A., Kyle, K. et Dascalu, M. (2019). "The Tool for the Automatic Analysis of Cohesion 2.0 : Integrating semantic similarity and text overlap". In : *Behavior Research Methods* 51.1, p. 14-27. doi : 10.3758/s13428-018-1145-z.

Crossley, S. A. et al. (2008b). "A linguistic analysis of the Common European Framework of Reference for Languages". In : *Language Testing* 25.2, p. 217-236.

Dale, E. et Chall, J. S. (1948). "A formula for predicting readability". In : *Educational Research Bulletin* 27.1, p. 11-20.

— (1949). "The concept of readability". In : *Elementary English* 26.1, p. 19-26.

de Jong, N. H. et al. (2002). "The Processing and Representation of Dutch and English Compounds : Peripheral Morphological and Central Orthographic Effects". In : *Brain and Language* 81.3. Published online December 6, 2001, p. 555-567. doi : 10.1006/brln.2001.2547.

Degand, L. et Pander Maat, H. (2003). "A contrastive study of Dutch and French causal connectives on the speaker involvement scale". In : *Usage-based approaches*

to Dutch. Sous la dir. de VERHAGEN, A. et WEIJER, J. van de. Utrecht : LOT, p. 175-199.

DEGAND, L. et SANDERS, T. (2002). "The impact of relational markers on expository text comprehension in L1 and L2". In : *Reading and Writing* 15.7-8, p. 739-757. DOI : 10.1023/A:1020987922877. URL : https://doi.org/10.1023/A:1020987922877.

DIJKSTRA, T. et VAN HEUVEN, W. J. B. (2002). "The architecture of the bilingual word recognition system : From identification to decision". In : *Bilingualism : Language and Cognition* 5.3, p. 175-197.

DOUMA, W. H. (1960). *De leesbaarheid van landbouwbladen : een onderzoek naar en een toepassing van leesbaarheidsformules*. 17. Wageningen : Afdeling Sociologie en Sociografie van de Landbouwhogeschool. URL : https://edepot.wur.nl/276323.

FRANÇOIS, T. (2011). "Les apports du traitement automatique du langage à la lisibilité du français langue étrangère". Thèse de doctorat. Louvain-la-Neuve : Université catholique de Louvain.

FRANÇOIS, T. et FAIRON, C. (2013). "Les apports du TAL à la lisibilité du français langue étrangère". In : *Traitement Automatique des Langues* 54.1, p. 171-202. URL : https://aclanthology.org/2013.tal-1.6.pdf.

GEERAERTS, D. (2001). "Een zondagspak ? Het Nederlands in Vlaanderen : gedrag, beleid, attitudes". In : *Waar gaat het Nederlands naartoe ? Panorama van een taal*. Sous la dir. de STROOP, J. Amsterdam : Bert Bakker, p. 231-253.

GIBSON, E. (1998). "Linguistic complexity : Locality of syntactic dependencies". In : *Cognition* 68.1, p. 1-76. DOI : 10.1016/S0010-0277(98)00034-1.

GOODMAN, K. S. (1967). "Reading : A psycholinguistic guessing game". In : *Journal of the Reading Specialist* 6.4, p. 126-135.

GOUGH, P. B. (1972). "One second of reading". In : *Language by ear and by eye*. Sous la dir. de KAVANAGH, J. F. et MATTINGLY, I. G. Cambridge, MA : MIT Press, p. 291-319.

GRAESSER, A. C. et al. (2004). "Coh-Metrix : Analysis of text on cohesion and language". In : *Behavior Research Methods, Instruments, & Computers* 36.2, p. 193-202. DOI : 10.3758/BF03195564.

GUNNING, R. (1952). *The Technique of Clear Writing*. New York : McGraw-Hill.

HARRIS, A. et JACOBSON, M. (1974). "Revised Harris-Jacobson readability formulas". In : *Proceedings of the annual meeting of the College Reading Association*. Bethesda, Maryland.

HE, H. et GARCIA, E. A. (2009). "Learning from imbalanced data". In : *IEEE Transactions on Knowledge and Data Engineering* 21.9, p. 1263-1284. URL : https://doi.org/10.1109/TKDE.2008.239.

HELL, J. G. van et GROOT, A. M. B. de (1998). "Disentangling Context Availability and Concreteness in Lexical Decision and Word Translation". In : *The Quarterly Journal of Experimental Psychology, Section A : Human Experimental Psychology* 51.1, p. 41-63. DOI : 10.1080/713755752.

HERBAY, A. C., GONNERMAN, L. M. et BAUM, S. R. (2018). "How Do French–English Bilinguals Pull Verb Particle Constructions Off ? Factors Influencing Second Language Processing of Unfamiliar Structures at the Syntax-Semantics Interface". In : *Frontiers in Psychology* 9, p. 1885. DOI : 10.3389/fpsyg.2018.

01885. URL : https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01885/full.

Hiligsmann, P. et Rasier, L. (2006). *Uitspraakleer Nederlands voor Franstaligen*. Mechelen : Wolters Plantyn.

Honnibal, M. et Montani, I. (2017). "spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In : *To appear*.

Hoover, W. A. et Gough, P. B. (1990). "The simple view of reading". In : *Reading and Writing : An Interdisciplinary Journal* 2, p. 127-160.

Howes, D. H. et Solomon, R. L. (1951). "Visual duration threshold as a function of word-probability". In : *Journal of Experimental Psychology* 41.6, p. 401-410. DOI : 10.1037/h0056020.

Hulstijn, J. H. (2001). "Intentional and incidental second language vocabulary learning : A reappraisal of elaboration, rehearsal and automaticity". In : *Cognition and Second Language Instruction*. Sous la dir. de Robinson, P. Cambridge : Cambridge University Press, p. 258-286.

Hunter, J. D. (2007). "Matplotlib : A 2D graphics environment". In : *Computing in Science & Engineering* 9.3, p. 90-95. DOI : 10.1109/MCSE.2007.55. URL : https://doi.org/10.1109/MCSE.2007.55.

Jordens, P. (2006). "Finiteness in Dutch as a second language". In : *Semantics in acquisition*. Sous la dir. de Geenhoven, V. van. Consulté aux pages 125–130, 135–140, 145–150. Springer, p. 125-150. URL : https://www.lotpublications.nl/Documents/212_fulltext.pdf.

Just, M. A. et Carpenter, P. A. (1980). "A theory of reading : From eye fixations to comprehension". In : *Psychological Review* 87.4, p. 329-354. DOI : 10.1037/0033-295X.87.4.329.

Kemper, S. et al. (1993). "Enhancing older adults' reading comprehension". In : *Discourse Processes* 16.4, p. 405-428. DOI : 10.1080/01638539309544846.

Klare, G. R. (1963). *The Measurement of Readability*. Ames : Iowa State University Press.

Kleijn, S. (2018). *Clozing in on readability : How linguistic features affect and predict text comprehension and on-line processing*. Utrecht : LOT.

Koda, K. (2005). *Insights into Second Language Reading : A Cross-Linguistic Approach*. Cambridge Applied Linguistics. Cambridge University Press.

Koster, J. (1975). "Dutch as an SOV Language". In : *Literatuur en Taal*. Slightly extended version of a paper read at the third annual meeting of the Algemene Vereniging voor Taalwetenschap, Amsterdam, 20 January 1973, p. 111-113. URL : https://www.dbnl.org/tekst/kost004dutc01_01/.

Krashen, S. (1985). *The Input Hypothesis : Issues and Implications*. London : Longman.

Labasse, B. (1999). "La lisibilité rédactionnelle : fondements et perspectives". In : *Communication et langages* 121.1, p. 86-103. DOI : 10.3406/colan.1999.3082.

Laroche, J. (1979). "La lisibilité des textes en français langue étrangère". In : *Études de Linguistique Appliquée* 33, p. 131-143.

Lei, W. et al. (2021). "Have We Solved The Hard Problem ? It's Not Easy ! Contextual Lexical Contrast as a Means to Probe Neural Coherence". In : *Proceedings of*

the AAAI Conference on Artificial Intelligence. T. 35. 15, p. 13208-13216. DOI : 10.1609/aaai.v35i15.17560.

LINGUA.COM (2025). *Learn languages effectively online !* https://lingua.com/. Consulté en juin 2025, disponible sur : https ://lingua.com/.

LIVELY, B. A. et PRESSEY, S. L. (1923). "A method for measuring the "vocabulary burden" of textbooks". In : *Educational Administration and Supervision* 9, p. 389-398. URL : https://www.researchgate.net/publication/224347785_Unlocking_Language_The_Classic_Readability_Studies.

LORGE, I. (1944). "Predicting Readability". In : *The Teachers College Record* 45.6, p. 404-419.

LUPKER, S. J. (2005). "The word frequency effect in visual word recognition : A review". In : *Psychonomic Bulletin & Review* 12.1.

MCCARTHY, P. M. et JARVIS, S. (2010). "MTLD, vocd-D, and HD-D : A validation study of sophisticated approaches to lexical diversity assessment". In : *Behavior Research Methods* 42.2, p. 381-392. DOI : 10.3758/BRM.42.2.381.

MCLAUGHLIN, G. H. (1969). "SMOG grading : A new readability formula". In : *Journal of Reading* 12.8, p. 639-646. URL : https://www.jstor.org/stable/40011226.

MILTON, J. et ALEXIOU, T. (2020). "Vocabulary Size Assessment : Assessing the Vocabulary Needs of Learners in Relation to Their CEFR Goals". In : *Vocabulary Studies in First and Second Language Acquisition*. Chapter 12, DOI : 10.1057/9780230242258$_1$2. Palgrave Macmillan.

NATION, I. S. P. (2006). "How large a vocabulary is needed for reading and listening ?" In : *Canadian Modern Language Review* 63.1, p. 59-82.

NT2 TAALMENU (2025). *NT2 Taalmenu - Oefeningen Nederlands als Tweede Taal (A1–B2)*. Consulté en juin 2025. URL : https://nt2taalmenu.nl/nt2-a1-oefeningen-2/#a1-lezen.

OOSTDIJK, N. et al. (2013). "The construction of a 500-million-word reference corpus of contemporary written Dutch". In : *Essential Speech and Language Technology for Dutch*. Sous la dir. de SPYNS, P. et ODIJK, J. Berlin, Heidelberg : Springer, p. 219-247.

OPENSUBTITLES.ORG CONTRIBUTORS (n.d.). *OpenSubtitles.org : Sous-titres gratuits pour films et séries TV*. https://www.opensubtitles.org/fr. Consulté le 17 juin 2025.

PAIVIO, A. (1971). *Imagery and Verbal Processes*. New York : Holt, Rinehart & Winston.

PAIVIO, A., CLARK, J. et LAMBERT, W. (1988). "Bilingual dual-coding theory and semantic repetition effects on recall". In : *Journal of Experimental Psychology : Learning, Memory, and Cognition* 14.1, p. 163-172. DOI : 10.1037/0278-7393.14.1.163.

PERREZ, J. (2006). "Connectieven, tekstbegrip en vreemdetaalverwerving : Een studie van de impact van causale en contrastieve connectieven op het begrijpen van teksten in het Nederlands als vreemde taal". Thèse de doctorat. Louvain-la-Neuve : Université catholique de Louvain.

RAZON, A. et BARNDEN, J. (2015). "A New Approach to Automated Text Readability Classification based on Concept Indexing with Integrated Part-of-Speech n-gram Features". In : *Proceedings of the International Conference Recent Ad-*

vances in Natural Language Processing. Sous la dir. de MITKOV, R., ANGELOVA, G. et BONTCHEVA, K. Hissar, Bulgaria : INCOMA Ltd. Shoumen, BULGARIA, p. 521-528. URL : https://aclanthology.org/R15-1068/.

RUMELHART, D. E. (1985). "Toward an interactive model of reading". In : *Theoretical Models and Processes of Reading*. Sous la dir. de SINGER, H. et RUDDELL, R. B. 3rd. Newark, DE : International Reading Association, p. 719-750.

SANDERS, T. et SPOOREN, W. (jan. 2012). "Discourse and Text Structure". In : *The Oxford Handbook of Cognitive Linguistics*. DOI : 10.1093/oxfordhb/9780199738632.013.0035.

SCHMITT, N. (2008). "Instructed Second Language Vocabulary Learning". In : *Language Teaching Research* 12.3, p. 329-363.

SMITH, E. (1961). "Devereaux readability index". In : *The Journal of Educational Research* 54.8, p. 289-303. DOI : 10.1080/00220671.1961.10882179.

STANOVICH, K. E. (1980). "Toward an Interactive-Compensatory Model of Individual Differences in the Development of Reading". In : *Reading Research Quarterly* 16, p. 32-71. DOI : 10.2307/747348.

STAPHORSIUS, G. et VERHELST, N. D. (1997). "Indexering van de leestechniek". In : *Pedagogische Studiën* 74.3, p. 154-164.

STAPHORSIUS, G. (1994). *Leesbaarheid en leesvaardigheid : De ontwikkeling van een domeingericht meetinstrument*. Arnhem : Cito.

STEINHILBER, A. (2023). "Modélisation bayésienne de l'apprentissage de la lecture". Consulté aux pages 15–25 et 45–60. Thèse de doctorat. Université Grenoble Alpes. URL : https://theses.hal.science/tel-04198656/file/STEINHILBER_2023_archivage.pdf.

WASKOM, M. et al. (sept. 2017). *mwaskom/seaborn : v0.8.1 (September 2017)*. Version 0.8.1. DOI : 10.5281/zenodo.883859. URL : https://doi.org/10.5281/zenodo.883859.

WETZEL, M., ZUFFEREY, S. et GYGAX, P. (2020). "Second Language Acquisition and the Mastery of Discourse Connectives : Assessing the Factors That Hinder L2-Learners from Mastering French Connectives". In : *Languages* 5.3, p. 35. DOI : 10.3390/languages5030035. URL : https://doi.org/10.3390/languages5030035.

WIKTIONARY CONTRIBUTORS (n.d.). *Wiktionary : Frequency lists/Dutch wordlist*. https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Dutch_wordlist. Consulté le 17 juin 2025.

ZIPF, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Reading, MA : Addison-Wesley.

ZWITSERLOOD, P. (1994). "The role of semantic transparency in the processing and representation of Dutch compounds". In : *Language and Cognitive Processes* 9.3, p. 341-368. DOI : 10.1080/01690969408402123. URL : https://doi.org/10.1080/01690969408402123.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté de philosophie, arts et lettres
Place Cardinal Mercier, 31, 1348 Louvain-la-Neuve, Belgique I **www.uclouvain.be/fial**