Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Det humanistiske fakultet
Institutt for språk og litteratur

Maren Langen Kjellmark

La taille du vocabulaire chez des étudiants de français au niveau universitaire en Norvège

Masteroppgave i lektorutdanning i språkfag
Veileder: Kjersti Faldet Listhaug

Mai 2021

NTNU
Kunnskap for en bedre verden

Maren Langen Kjellmark

La taille du vocabulaire chez des étudiants de français au niveau universitaire en Norvège

Masteroppgave i lektorutdanning i språkfag
Veileder: Kjersti Faldet Listhaug
Mai 2021

Norges teknisk-naturvitenskapelige universitet
Det humanistiske fakultet
Institutt for språk og litteratur

NTNU
Kunnskap for en bedre verden

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué à l'aboutissement et la rédaction de ce mémoire.

Je souhaiterais tout d'abord remercier ma directrice de mémoire, Kjersti Faldet Listhaug, qui m'a encouragée et soutenue pendant mes travaux d'écriture. Elle m'a inspirée et donnée de bons conseils. Je suis très reconnaissante de nos rendez-vous d'orientation éclairant tout au long de cette période de travail.

Mes remerciements s'adressent également à toutes les personnes qui ont généreusement accepté de participer à cette étude.

J'aimerais remercier ma bonne amie Maren Langhaug Gullikstad pour son aide et ses conseils excellents avec les traitements des données sur Microsoft Excel. Je remercie également Dahlia Thomas et mes amis pour m'avoir encouragée et motivée lors de la rédaction de ce mémoire

Finalement, un grand merci à ma famille pour sa patience et son soutien. À la mémoire de ma mère, toujours dans mon cœur.

Résumé

Selon plusieurs études récentes, les connaissances lexicales sont primordiales pour l'acquisition de la langue en général. Dans ce mémoire, nous allons examiner la taille du vocabulaire chez des étudiants du premier semestre d'études universitaires de français en Norvège. 16 étudiants âgés de 18 à 25 ans ont accompli trois versions différentes du test X-Lex (Meara et Milton, 2003). Chaque version du test contient 100 mots français ainsi que 20 mots inventés. Les vrais mots ont été sélectionnés parmi les 5000 mots les plus fréquents de la langue française. La moyenne était ensuite calculée ainsi que la distribution de connaissance selon les cinq bandes de fréquence. Nous avons aussi examiné lesquels parmi les mots inventés les participants ont indiqué qu'ils connaissent. Les mots inventés ressemblant aux vrais mots anglais sont les plus cochés. Les résultats font ressortir que les étudiants norvégiens connaissent en moyenne 2451 mots français. Ceci concorde avec l'estimation de la taille du vocabulaire nécessaire pour atteindre le niveau CECRL de maîtrise de langue A2. Les connaissances préalables recommandées pour étudier le français à l'université en Norvège est aussi le niveau A2. La taille du vocabulaire chez les étudiants en premier semestre d'études universitaires de français en Norvège est similaire à celle des étudiants anglophones dans des études précédentes.

Une étude effectuée en Suède a suggéré que l'étendue du vocabulaire nécessaire pour arriver à un niveau de compréhension acceptable de trois textes littéraires qui sont au programme du premier semestre d'études de français. La connaissance d'entre 7000 et 9000 lemmes donnent une couverture (*coverage)* de 98%. Nos résultats et les similarités entre les programmes universitaires en Norvège et en Suède soulève la question de savoir si les étudiants norvégiens ont un vocabulaire suffisamment large pour lire et comprendre les textes qui figurent au programme.

Table de matières

Chapitre 1. Introduction

La grammaire est souvent considérée comme l'un des éléments les plus importants dans l'apprentissage d'une langue. Cependant, il est impossible de communiquer sans des mots à insérer dans le système syntaxique fourni par la grammaire. De nombreuses recherches menées les dernières décennies ont montré que le développement de la connaissance lexicale est une composante essentielle dans l'acquisition d'une langue étrangère (Milton, 2009; Schmitt, 2000). En outre, contrairement à d'autres aspects de la connaissance du langage, la taille du vocabulaire, elle, peut être quantifiée. Lorsque nous parlons de la taille du vocabulaire, nous faisons référence au nombre de mots connus par un locuteur ou un apprenant d'une langue spécifique. Avoir une estimation de la taille du vocabulaire des apprenants est alors un outil utile qui sert à décider si leur niveau de connaissance lexicale est adéquat pour comprendre des textes, accomplir des exercices et faire d'autres activités d'apprentissage pendant les cours. De plus, les connaissances lexicales se sont avérées être un bon indicateur de maîtrise de langue en général (Laufer et Ravenhorst-Kalovski, 2010; Milton, Wade et Hopkins, 2010; Stæhr, 2008). Plusieurs tests visant à estimer combien de mots connaissent les apprenants d'une langue étrangère ont été développé et ils sont souvent rapides et faciles à faire passer. Les tests de la taille du vocabulaire peuvent alors être un outil pratique et facilement accessible aux enseignants pour donner une indication de la compétence linguistique générale chez les apprenants.

La recherche sur l'acquisition du vocabulaire a connu un essor récemment, mais Milton (2008) constate qu'il y a toujours une lacune d'études examinant systématiquement l'acquisition de vocabulaire dans d'autres langues étrangères que l'anglais. En ce qui concerne la recherche sur la taille du vocabulaire en français en tant que langue étrangère, la plupart des études précédentes ont été conduites sur les apprenants anglophones, voir par exemple Milton (2008), David (2008) et Pignot-Shahov (2014, 2018). Lindqvist (2017) a mené une étude en Suède qui se propose de contribuer à combler cette lacune. L'étude de Lindqvist (2017) a donné un survol quantitatif sur le développement de la taille du vocabulaire chez les élèves suédophones au collège. L'un des buts exprimés dans l'introduction de son étude était d'ouvrir à des futures études sur le même sujet dans les pays scandinaves et ceci a été la principale source d'inspiration pour ce mémoire.

1. 1. Les objectifs de la présente étude

À notre connaissance, il n'y a aucune étude sur la taille du vocabulaire chez les étudiants norvégiens de français. L'objectif principal de la présente étude sera alors d'estimer la taille du vocabulaire en français chez des étudiants en première année des études universitaires en Norvège. Nous allons faire cela à l'aide d'un test bien établi et éprouvé, le test X-Lex développé par Meara et Milton (2003). C'est le même test qui a été utilisé par Milton (2008), David (2008), Pignot-Shahov (2014, 2018) et Lindqvist (2017). Le X-Lex mesure la taille du vocabulaire réceptif et il consiste en 100 mots français ainsi que 20 mots inventés mais qui ressemblent morphologiquement aux vrais mots français. Les participants doivent cocher les mots qu'ils connaissent. Un point qui distingue l'étude de Lindqvist (2017) de la plupart des autres enquêtes précédentes est que pour les élèves suédois, le français est leur deuxième langue étrangère après l'anglais, une L3 pour ainsi dire. Par L3 nous entendons ici la troisième langue chronologiquement acquise par un locuteur. Pour les participants en Norvège le français est leur L3 aussi, après le norvégien et l'anglais. Les élèves suédois et les étudiants norvégiens apprennent donc le français dans un contexte très similaire. Les participants suédois ont obtenu des scores étonnamment élevés par rapport à leurs homologues dans d'autres études. Lindqvist (2020) a mené une étude de suivi pour déterminer si les élèves ont utilisé leurs connaissances en anglais en accomplissant le test de vocabulaire. Ces états de fait nous amènent à poser les deux questions de recherche suivantes :

1. Quelle est la taille du vocabulaire en français chez des étudiants en première semestre à l'université en Norvège ?

2. Quels mots inventés les étudiants norvégiens confondent-ils avec de vrais mots français ?

Nous estimons que la taille du vocabulaire chez les étudiants en Norvège est plus grande que la taille du vocabulaire chez les élèves les plus âgés dans l'étude de Lindqvist (2017) et à peu près à égalité avec le score des groupes correspondants dans les études de Milton (2008), David (2008) et Pignot-Shahov (2014, 2018).

Dans ce mémoire, nous commencerons avec le cadre théorique et nous définirons quelques concepts-clés dans le domaine du vocabulaire. Ensuite, nous allons traiter de ce que dit la recherche sur la taille du vocabulaire et les résultats des études antérieures. Le chapitre quatre sera consacré à la méthodologie de la recherche et nous y présenterons en détail le processus de collecte des données, les participants et le calcul de scores. Au cinquième chapitre nous présenterons les résultats du test. Au sixième chapitre nous discuterons nos trouvailles à la lumière des résultats d'études précédentes et à l'aide de nos questions de recherche. Nous voulons également expliciter les implications pédagogiques de nos découvertes. Finalement, nous exposerons les limites de notre étude et nous offrirons aux chercheurs quelques pistes d'études futures dans le même domaine.

Chapitre 2. Le mot, le lexique mental et le vocabulaire

2. 1. La notion du terme « mot »

Comme nous allons estimer le nombre de mots connus par des étudiants de français, il faut d'abord regarder de plus près ce qui constitue un *mot*. A priori, la notion du terme « mot » semble facile à comprendre, et dans le dictionnaire Le Petit Robert on propose la définition suivante : « Chacun des sons ou groupes de sons (de lettres ou groupes de lettres) correspondant à un sens isolable spontanément, dans le langage ; (par écrit) suite ininterrompue de lettres, entre deux blancs (*Mot*, 2016) ». Autrement dit, la définition « standard » est qu'un mot est une compilation de sons de langage qui a un sens propre et qui peut se tenir indépendamment. Toutefois, il existe en français des termes comme *pomme de terre*. Chaque unité a un sens propre isolable spontanément, et à l'écrit, il y a trois unités séparées par des blancs. Est-ce que cette combinaison des trois unités *pomme, de* et *terre* constitue donc trois mots ? Ou est-ce qu'elles forment un seul mot quand elles se produisent ensemble pour désigner un légume distinct ? Déjà, nous voyons que la définition présentée ci-dessus ne tient pas. Riegel, Pellat et Rioul (2018, p. 887) constatent : « Si le mot est intuitivement identifié comme l'unité de base du système grammatical et dénominatif que forme la langue, son statut n'en reste pas moins problématique. ». La question qui se pose est : comment peut-on estimer la taille du vocabulaire quand la notion du *mot* est si vague ?

Pour explorer le concept du *mot*, nous allons commencer par considérer brièvement ce que compose un mot, notamment les morphèmes. Sa définition varie selon les linguistes mais le morphème est généralement considéré comme le plus petit élément linguistique doté de sens. Le mot *banane* est un morphème puisqu'il a un sens global et il ne peut plus se décomposer en unités signifiantes. En revanche, l'adverbe *injustement* s'analyse en trois morphèmes : *in-* + *juste* + *-ment* (Riegel et al., 2018, p. 890). Il y a deux grandes catégories de morphèmes : les morphèmes lexicaux (noms, adjectifs, verbes et adverbes) et les morphèmes grammaticaux (affixes, déterminants, pronoms, auxiliaires, prépositions et conjonctions). À partir d'un morphème appelé base ou radical, on peut ajouter des autres morphèmes, des affixes. Les affixes flexionnels ne créent pas des mots nouveaux, mais des formes différentes d'un même mot. Les terminaisons d'un verbe conjugué au présent *-e, -es, -ez, -ont, -ent* sont des exemples et la forme conjuguée du verbe s'appelle une inflexion. Les affixes dérivationnels servent à former des mots nouveaux dits dérivations. À partir d'une base par exemple l'adjectif *gentil*, on peut ajouter le suffixe *-ment* pour former l'adverbe *gentiment (Riegel et al., 2018).*

Il faut mentionner que toutes les langues ne se servent pas de mêmes mécanismes pour créer des mots. En finnois, et dans d'autres langues agglutinantes comme le hongrois et le turc, on peut former des mots complexes à partir d'un mot-racine et des affixes. Ces affixes sont des morphèmes ayant des fonctions grammaticales et ils peuvent exprimer le nombre, le lieu, la personne et plein d'autres nuances. Le mot finnois *vaimollenikin* est constitué d'unités *vaimo-lle-ni-kin*. Ceci correspond en français à *femme-à-ma-aussi*, ou plutôt à la phrase *à ma femme aussi*. Un seul « mot » peut alors contenir la même information qu'une phrase entière en français. Cela veut dire que quelques prépositions ne sont pas des « mots » indépendants en finnois, mais des affixes ajoutés aux mots. En outre, le chinois est une langue qui combine des idéogrammes pour créer des mots. Milton (2010) remarque que dans cette langue les limites des mots à l'écrit ne sont pas marquées, de sorte qu'il n'est pas toujours clair de définir où un mot se termine et où commence l'autre. La façon dont on définit ce qu'est un mot peut donc dépendre de la langue que l'on examine.

Revenons à la langue française, pour la plupart des verbes, les différentes inflexions ne changent pas la base du mot, mais pour certains le radical change considérablement. Le verbe *être* a par exemple des inflexions possibles comme *suis*, *fut* et *seront*. Ces formes se démarquent considérablement de la forme infinitive. Est-ce qu'il faut compter chaque radical différent des verbes comme un mot à part ? Un problème semblable se pose avec les adjectifs. Un exemple est l'adjectif *vieux* où il n'est pas forcément évident au premier regard qu'il s'agit du même mot quand on voit sa forme féminine *vieille*. Dans la prochaine section, nous allons voir comment ces questions sont traitées dans le domaine de la recherche de vocabulaire.

2. 2. Regroupement des mots ; les familles de mots et les lemmes

Mesurer la taille du vocabulaire dans une langue étrangère chez un apprenant oblige à prendre des décisions sur ce qui doit être compté comme un mot. Il y a maintenant des conventions à ce sujet mais pas de « règles » strictes. Cela est dû en grande partie au fait qu'il y a des différences en structure entre les langues (voir 2. 1.) et de différentes conventions selon le champ de recherche. Généralement, les mots composés tels que *pomme de terre* mentionné plus haut sont considérés comme un seul mot dans le contexte de dénombrement des mots. De même, les mots ayant la même forme mais qui ont plusieurs significations différentes comme *avocat*, désignant soit un auxiliaire de justice, soit un fruit, sont comptés comme deux mots

séparés. En ce qui concerne les dénombrements des mots en général, deux grandes conventions de classement ont été développées : la famille des mots et la lemmatisation.

La lemmatisation est une méthode où l'on regroupe les mots en *lemmes*. Un lemme consiste en un mot principal et ses inflexions les plus fréquentes. En français, le lemme de l'adjectif *grand* renvoie à quatre formes : *grand, grands, grande, grandes.* L'adverbe *grandement* et le verbe *grandir* n'appartiennent pas à ce lemme parce qu'ils appartiennent à d'autres classes du mot. Ces dérivations (et leurs inflexions respectives) sont comptées comme deux lemmes à part. Pour les verbes, le lemme sera représenté par l'infinitif. La convention de lemmatisation est particulièrement utile pour estimer la taille ou la connaissance du vocabulaire chez des apprenants en niveau élémentaire et intermédiaire car on présuppose que les apprenants à ces niveaux ne maîtrisent que les inflexions les plus fréquentes (Milton, 2009, p. 11). Ces inflexions sont souvent acquises tôt dans le processus d'apprentissage et une fois qu'un apprenant connaît une règle de conjugaison, elle peut être appliquée à un grand nombre d'autres mots sans que chaque nouvelle forme doive être apprise séparément. Les verbes réguliers en français qui se terminent par *-er* sont un bon exemple. Ceci est l'une des principales raisons pour lesquelles les tests de taille du vocabulaire s'appuient souvent sur des études de fréquence de mots où les mots comptés sont lemmatisés (Milton, 2010).

L'autre convention utilise une unité de mesure moins restrictive. Les chercheurs anglophones regroupent souvent la forme de base d'un mot (le radical ou la racine) et ses formes associées dans une famille de mots (*word family*). Les inflexions et les dérivations des mots, qui partagent le même radical, sont regroupées dans la même famille de mots. Le raisonnement qui sous-tend ce classement est qu'avec des connaissances morphologiques élémentaires, on peut discerner les sens des autres membres de la famille de mots, étant donné qu'on connait déjà la signification de l'un des membres (Ramnäs, 2019). Par conséquent, les mots *grandir, grand, grandement* et *grandeur* font partie de la même famille de mots. Le cadre de la famille de mots est plutôt grand, et il comprend aussi des dérivations peu fréquentes que les locuteurs non natifs connaissent rarement. Les comptages s'appuyant sur la famille de mots comme unité de mesure produiront alors des chiffres plus petits dans une mesure de la taille du vocabulaire que les calculs effectués à l'aide d'un comptage lemmatisé. Afin de comparer une mesure de taille du vocabulaire utilisant des familles de mots avec un autre utilisant des lemmes, Milton (2009) propose une règle approximative : multipliez le score dans les familles de mots par 1,6 pour obtenir un score équivalent très approximatif en lemmes.

Quand on compare des études sur la taille du vocabulaire, il faut tenir en compte les diversités en la construction des mot dans des langues différentes et les différentes manières de regrouper des mots. Milton (2009) affirme que nous avons tendance à utiliser le mot «mot», probablement par souci de facilité et de commodité, alors que nous référons en réalité à des définitions très spécialisées du terme. À cause de toutes ces ambiguïtés, les linguistes préfèrent souvent aujourd'hui utiliser des termes scientifiques plus précis. Par souci de simplicité, nous utilisons dans la suite le terme *mot* pour signifier les lemmes à moins qu'une terminologie plus précise ne soit requise.

2. 3. Qu'est-ce que veut dire de connaitre un mot ?

Dans la section précédente, nous avons examiné les différentes définitions de ce qu' un mot. Nous allons continuer avec l'une des autres difficultés principales concernant la mesure de la taille du vocabulaire, notamment explorer ce que *connaître* un mot signifie, et ceci plus précisément dans une langue étrangère. Essayer de définir ce qu'implique la connaissance d'un mot est compliqué, car ceci est un domaine très complexe et les études menées sur ce thème impliquent de nombreux domaines scientifiques tels que la psycholinguistique et la neurolinguistique. Ce domaine de recherche n'en devient pas moins complexe lorsqu'il s'agit de la connaissance de mots dans une deuxième ou troisième langue. Dans la section suivante, nous présenterons brièvement quelques concepts-clés et trouvailles de la recherche abordant la notion de ce que signifie connaître un mot dans une langue étrangère.

2.3.1. Le lexique mental et les recherches empiriques sur l'acquisition du vocabulaire

La plupart des théories sur la taille du vocabulaire reposent sur l'idée que nous posséderons tous un « lexique mental » où les mots que nous connaissons sont stockés et organisés de manière à faciliter leur récupération rapide afin que nous puissions communiquer efficacement. Les études psycholinguistiques sur le lexique mental ont commencé à mettre en lumière une partie de son organisation, mais la façon dont le lexique mental est organisé dans son ensemble n'est toujours pas claire. Pignot-Shahov (2018) constate néanmoins qu'il doit être organisé d'une manière ou d'une autre, sinon il serait très difficile pour les locuteurs de garder en mémoire des mots organisés au hasard. Nous n'entrons pas dans les détails dans ce mémoire, mais de manière générale, il est supposé que, pour interpréter le monde on utilise des «concepts» liés à des mots tels que *chaussure* et *maison*. Ces concepts sont des amalgames abstraits des idées et des expériences de ce que constitue par exemple une

chaussure ou une maison (Aitchison, 2012; Pignot-Shahov, 2018). La manière dont ces concepts et ces mots interagissent n'est pas claire et plusieurs modèles du processus ont été suggérés. Pour plus de détails et des résumés sur ce domaine de recherche et les modèles du lexique bilingue, voir par exemple Listhaug (2015) et Pignot-Shahov (2018).

On ne se sait pas vraiment si les morphèmes sont stockés séparément et puis assemblés pour former des mots complets ou complexes au moment de l'emploi, ou si chaque dérivation ou inflexion d'un mot est stocké une par une comme des unités distincts. Les recherches psycholinguistiques suggèrent que le cerveau regroupe ensemble au moins quelques formes différentes d'un mot (Clahsen, Eisenbeiss, Hadler et Sonnenstuhl, 2001; Thornbury, 2002). Le concept de lemme se base sur ces découvertes, que la forme de base d'un mot est stockée avec son rôle syntaxique et ses inflexions (Pignot-Shahov, 2018). Une difficulté liée au lemme est qu'il existe des formes irrégulières où les racines des mots se modifient selon la conjugaison. On débat toujours pour savoir si ces formes doivent être comptées comme des lemmes séparés ou non, et on ne sait pas exactement comment ces formes sont stockées dans le cerveau.

Quant au lexique bilingue, l'une des questions principales dans la recherche sur le sujet est de savoir si les lexiques en L1 et L2 sont organisés séparément, ou si tous les mots connus par un individu appartiennent à un grand lexique combiné. Une autre question importante est de savoir si les apprenants peuvent accéder aux concepts directement à partir de leur lexique en L2 (ou en L3), ou s'ils y ont accès seulement via leur lexique en L1. Plusieurs modèles visant à expliquer la cadre du lexique plurilingue ont été proposés, voir par exemple de Bot (2004), Ameel, Storms, Malt et Sloman (2005) et Kroll, Gullifer et Rossi (2013). Actuellement la recherche au sujet du lexique mental chez des locuteurs connaissant plus de deux langues est limitée, mais il y a un intérêt croissant pour le phénomène du plurilinguisme. Aujourd'hui, l'hypothèse dominante dans ce domaine affirme que toutes les langues d'un plurilingue sont actives lors du traitement langagier (Kroll et al., 2013). En outre, il semble que les lexiques des différentes langues sont connectés, assez étroitement, les uns aux autres (de Bot, 2004).

Sur la base de recherches antérieures, il a été constaté que la L1 et la L2 influencent la L3 au niveau lexical. Lors d'activités dans la troisième langue, les apprenants se servent souvent d'une autre langue étrangère plutôt que de leur langue maternelle, mais cette influence semble diminuer à mesure que l'apprenant devient plus compétent en L3 (Lindqvist, 2016). Plusieurs études ont également montré que les similarités entre la langue maternelle (et éventuellement les autres langues connues) et la langue cible facilitent la compréhension et l'acquisition de nouveaux éléments lexicaux (Szabo, 2020). Or, de nombreuses variables linguistiques, telles

que le niveau de maîtrise de langue et la fréquence d'occurrence des mots en question, peuvent avoir une incidence sur la nature du cadre multilingue et l'effet des influences inter-linguistiques (Krautz, 2020; Wlosowicz, 2010). Bardel, Gudmundson et Lindqvist (2012) ont montré par exemple que les apprenants suédophones peuvent reconnaitre quelques mots français moins fréquents grâce à leurs similitudes avec des mots suédois empruntés au français. Nous reviendrons plus en détails sur l'un de ces aspects, notamment la notion de mots cognats, dans la section 3.3.2.

2. 4. Propriétés du vocabulaire

L'acquisition d'un mot est un processus complexe. Nous avons vu que la recherche empirique a montré que l'esprit utilise des concepts pour comprendre le monde et que ces concepts sont à leur tour liés à des mots. Ces mots sont en quelque sorte stockés dans le cerveau. Qu'est-ce que cela sous-entend précisément ? Il est possible de pouvoir reconnaître la forme phonologique ou sonore d'un mot, sans pouvoir le comprendre ou l'épeler à l'écrit. De plus, il est possible de saisir le sens d'un mot lorsqu'on le rencontre dans un contexte sans pouvoir s'en servir soi-même. Afin de pouvoir utiliser un mot correctement dans une langue étrangère, de nombreux types de connaissances sont nécessaires.

Une convention courante distingue entre les connaissances *actives* ou *productives* et *passives* ou *réceptives* des mots. La connaissance réceptive fait référence au moment où un apprenant est capable de reconnaître un mot lorsqu'il est entendu ou lu. Être capable d'utiliser un mot dans la parole ou à l'écrit, relève de la connaissance productive. On estime généralement que le vocabulaire réceptif est plus large que le vocabulaire productif (Milton, 2009). Une autre convention courante, d'abord proposée par Anderson et Freebody (1981), est la distinction entre *l'étendue du vocabulaire* et *la profondeur du vocabulaire.* L'étendue se réfère au nombre de mots connus par l'apprenant et la profondeur renvoie à la connaissance que l'apprenant a de chaque mot. Ces deux termes semblent intuitivement faciles à comprendre, mais il n'est pas facile par exemple de définir quel type de connaissance est nécessaire chez l'apprenant pour qu'un mot soit considéré comme «connu» quand on parle de l'étendue du vocabulaire. Il y a une grande différence entre être capable de donner une définition d'un mot, et la capacité à reconnaître qu'un mot est un mot dans une langue étrangère, même s'il ne sait pas ce que signifie ce mot, ou s'il ne peut pas en fournir une traduction. Quant à la profondeur du vocabulaire, dans un sens, le terme pourrait faire référence à la connaissance des

caractéristiques du mot, comme les caractéristiques orthographiques et morphologiques. D'autre part, il peut également être interprété comme la connaissance spécifique des différentes significations d'un mot, les collocations et des contextes dans lesquels le mot est utilisé (Read, 2004).

Ce qui est impliqué dans la connaissance d'un mot			
R = connaissance réceptive ; P = connaissance productive			
Forme	Orale	R	Quels sont les sons du mot ?
		P	Comment le mot est-il prononcé ?
	Écrite	R	À quoi le mot ressemble-t-il ?
		P	Comment le mot est-il écrit et orthographié ?
	Parties des mots	R	Quelles sont les parties reconnaissables dans ce mot ?
		P	Quelles sont les parties de mots nécessaires pour exprimer le sens ?
Sens	Forme et sens	R	Quel sens la forme de ce mot signale-t-elle ?
		P	Quelle forme de mot peut être utilisée pour exprimer ce sens ?
	Concept et référents	R	Qu'est-il inclus dans ce concept ?
		P	À quels items ce concept peut-il se référer ?
	Associations	R	À quels autres mots cela nous fait-il penser ?
		P	Quels autres mots pourrions-nous utiliser à la place de celui-ci ?
Utilisation	Fonctions grammaticales	R	Dans quelles structures le mot apparaît-il ?
		P	Dans quelles structures devons-nous utiliser ce mot ?
	Collocations	R	Quels mots ou types de mots apparaissent avec celui-ci ?
		P	Quels mots ou types de mots devons-nous utiliser avec celui-ci ?
	Contraintes d'utilisation	R	Où, quand et à quelle fréquence pourrions-nous nous attendre à rencontrer ce mot ?
		P	Où, quand et à quelle fréquence pouvons-nous utiliser ce mot ?

Tableau 2-1 Ce qui est impliqué dans la connaissance d'un mot (Nation, 2001) (traduction par Vinet (2011)

Nation (2013) a proposé un modèle visant à résumer les complexités et les différents aspects qui sont impliqués dans la connaissance d'un mot. Le tableau 2-1 ci-dessous montre les trois aspects principaux de la connaissance d'un mot : sa forme, son sens et son emploi. Ces trois aspects sont ensuite divisés en neuf sous-catégories qui, à leur tour ont deux faces, une dimension réceptive et une dimension productive. Daller, Milton et Treffers-Daller (2007) ajoutent également une troisième dimension, à savoir *la fluidité.* Cet aspect est défini comme la capacité d'un locuteur à reconnaître et à utiliser des mots avec rapidité et facilité. Comme nous pouvons le voir, la notion de connaissance d'un mot est complexe, et Lindqvist et Ramnäs (2016, p. 57) le résume avec précision : « Le fait de connaître pleinement un mot fait donc intervenir une multitude de connaissances qui sont parfois à la limite entre le lexique et la grammaire. ». Il est évident qu'un seul test ne pourrait jamais mesurer tous les aspects de la connaissance lexicale. Nous avons cependant un petit nombre de tests conçus pour mesurer

des domaines spécifiques de la connaissance des mots. L'étendue du vocabulaire réceptif est l’aspect qui est probablement le plus facile à tester car c’est une qualité qui au moins est dénombrable ou mesurable dans un sens significatif (Milton, 2009). Dans le chapitre suivant, nous examinerons combien de mots il faut connaitre dans une langue étrangère, comment on peut mesurer la taille du vocabulaire réceptif et les résultats des études précédentes dans ce domaine de recherche.

Chapitre 3. La taille du vocabulaire réceptif

Dans le chapitre précèdent, nous avons traité la notion du terme mot et les propriétés du vocabulaire. Dans la suite, nous allons présenter ce que dit la recherche sur la taille du vocabulaire nécessaire pour bien maîtriser une langue. Nous allons également présenter quelques études antérieures, menées à travers le test X-Lex, sur la taille du vocabulaire chez des apprenants de français à des stades d'apprentissage différents.

3. 1. Combien de mots faut-il connaître ?

Même un locuteur natif ne connaît pas tous les mots de sa langue et il serait irréaliste de s'attendre à ce qu'il connaisse tout le vocabulaire spécialisé, tous les archaïsmes, tous les noms propres, ou tout autre vocabulaire très peu utilisé. Ce serait donc un objectif encore plus irréaliste pour un apprenant d'une langue L2 ou L3 d'apprendre tous les mots de sa langue cible. Pourtant, avoir une estimation du nombre de mots qu'il faut connaître pour communiquer et lire avec fluidité dans une langue quelconque est très utile dans le contexte d'enseignement et d'apprentissage de langues. La plupart des études traitant de ce domaine ont été effectuées soit sur les locuteurs natifs d'anglais, soit sur l'anglais en tant que langue étrangère.

3.1.1. Couverture

En anglais, le pourcentage des mots connus d'un texte donné est appelé *coverage*. Selon notre connaissance, il n'existe pas un terme français équivalent mais dans la suite nous utilisons la traduction française *couverture*. Dans une étude effectué par Nation (2006), on a trouvé que si un apprenant maîtrise les mille familles de mots (voir 2. 2.) les plus fréquentes en anglais, il peut comprendre environ 80% des mots d'un texte anglais ordinaire. Lindqvist et Ramnäs (2016) remarquent que ce nombre de familles de mots assez restreint donne une couverture qui peut sembler étonnamment élevée. Toutefois, Ramnäs (2019) a analysé le vocabulaire de trois textes littéraires qui sont étudiés par l'ensemble des étudiants pendant le premier semestre d'études de français à l'université de Göteborg : *La Petite Bijou* (Modiano, 2001), *L'étranger* (Camus, 1995 [1942]) et *No et moi* (de Vigan, 2007). Le but de son travail était de déterminer l'étendue du vocabulaire nécessaire pour lire ces textes avec une fluidité acceptable et sans recours au dictionnaire. Une couverture de 80% correspond à environ un mot inconnu sur cinq, et les exemples ci-dessus en figure 3-1[1] , montrent qu'il est très difficile

[1] Le texte est un extrait tiré du roman *La Petite Bijou* (Modiano, 2001, p. 9-10) et les travaux de Lonsdale et Le Bras (2009) ont servi de référence concernant les informations sur la fréquence des mots.

de se faire une idée du contenu d'un texte quand un mot sur cinq est inconnu. Une couverture de 80% n'est par conséquent pas suffisante.

Laufer et Ravenhorst-Kalovski (2010) et Nation (2013) estiment qu'une couverture de 95% (voir l'exemple avec quatre mots inconnus en figure 3-1), est nécessaire pour arriver à la compréhension minimale acceptable. Ce qui constitue un niveau de compréhension acceptable a été établi sur la base des scores obtenus à un test de compréhension de lecture à choix multiples et par un rappel écrit d'un texte. Laufer et Ravenhorst-Kalovski (2010) maintiennent en outre que dans de nombreuses circonstances même un tel pourcentage de couverture ne suffit pas à une véritable maîtrise d'un texte, y compris dans des études universitaires. Pour atteindre un niveau acceptable de compréhension, ils maintiennent que 98% de couverture est essentiel. Cela correspond à deux mots inconnus sur 100, autrement dit environ un mot inconnu sur cinq lignes de texte (supposant qu'il y a environ10 mots par ligne). Le tableau 3-1 montre la relation entre le pourcentage de couverture du texte, le nombre de mots inconnus sur 100 et le nombre de lignes de texte pour un mot inconnu.

En fait, le contexte autour d'un mot inconnu soutient la compréhension global d'un texte. La probabilité que le lecteur se débrouille face à un mot inconnu est d'autant plus grande quand la densité de mots inconnus dans le texte est faible (Nation, 2013). Les apprenants d'une langue étrangère sont souvent encouragés à lire beaucoup dans leur langue cible. La lecture de textes littéraires présente de nombreux avantages. Elle permet de découvrir des mots en contexte et s'y exposer, mais la lecture ne suffit pas pour l'apprentissage implicite de nouveau mots. Dans un contexte d'apprentissage, il est conseillé de choisir des textes ayant un niveau qui correspond au 98% de couverture chez le lecteur (Nation, 2013, p. 207). Nation (2013, p. 207-208) estime qu'en anglais, il faut connaître entre 8000 et 10 000 familles de mots pour obtenir une couverture de 98% pour des textes tels que les roman écrits tandis que la compréhension de 6000 à 7000 familles de mots est nécessaire pour des textes oraux. Il est important de faire remarquer que la recherche dans le tableau citée ci-dessus a été faite sur l'anglais.

% Couverture de texte	Le nombre de mots inconnus sur 100	Le nombre de lignes de texte pour 1 mot inconnu
99	1	10
98	2	5
95	5	2
90	10	1

Tableau 3-1 Aperçu sur la relation entre la couverture et le nombre de mots inconnus sur 100 (Nation, 2013, p. 206)

1 000 lemmes (18 mots inconnus)
Une ..... m'était revenue en ....., l'une des quelques ..... que j'ai gardées de ma mère. Son visage est ..... comme si un ..... l'avait fait ..... de la nuit. J'ai toujours ..... une ..... devant cette ..... Dans mes ....., chaque fois, c'était une .... que quelqu'un me ..... — un ..... de police, un ..... de la ..... — pour que je puisse ..... cette personne. Mais je restais ..... Je ne savais rien d'elle.

3 000 lemmes (4 mots inconnus)
Une *photo* m'était revenue en *mémoire*, l'une des quelques *photos* que j'ai gardées de ma mère. Son visage est *éclairé* comme si un ..... l'avait fait *surgir* de la nuit. J'ai toujours *éprouvé* une *gêne* devant cette *photo*. Dans mes rêves, chaque fois, c'était une *photo* ..... que quelqu'un me tendait — un *commissaire* de police, un *employé* de la ..... — pour que je puisse *identifier* cette personne. Mais je restais ..... Je ne savais rien d'elle.

L'extrait sans trous
Une *photo* m'était revenue en *mémoire*, l'une des quelques *photos* que j'ai gardées de ma mère. Son visage est *éclairé* comme si un *projecteur* l'avait fait *surgir* de la nuit. J'ai toujours *éprouvé* une *gêne* devant cette *photo*. Dans mes rêves, chaque fois, c'était une *photo anthropométrique* que quelqu'un me tendait — un *commissaire* de police, un *employé* de la *morgue* — pour que je puisse *identifier* cette personne. Mais je restais *muette*. Je ne savais rien d'elle. (Modiano 2001, 9–10)

Figure 3-1 Exemple différents niveaux de couverture, pris de Ramnäs (2019)

Revenons à l'analyse de Ramnäs (2019) qui montre qu'une couverture d'environ 95% nécessite la connaissance de 3000 à 4000 lemmes. Pour atteindre une couverture de 98% de ces romans, le lecteur doit connaitre entre 7000 et 9000 lemmes. Il faut alors apprendre beaucoup plus de mots pour passer d'une couverture de 95% à une couverture de 98%. Comme nous l'avons vu ci-dessus, Nation estime que la connaissance de 8000 à 9000 familles de mots est nécessaire pour lire des romans. Cependant, Ramnäs (2019) souligne qu'il s'agit là d'une autre langue (l'anglais) et d'une autre unité de mesure (familles de mots et non pas les lemmes). En plus, l'étude de Nation s'intéresse aux romans en général alors que la sienne ne porte que sur trois romans spécialement choisis pour leur relative simplicité. Même si l'étude de Ramnäs (2019) avait peu d'envergure, les résultats indiquent que les étudiants en Suède (et potentiellement en Norvège ?) ont besoin d'un vocabulaire environ 7000 à 9000 lemmes pour lire les textes littéraires au programme avec une certaine fluidité, ou sans trop de difficulté. Il serait intéressant de faire une étude semblable pour les textes académiques authentiques (et non pas littéraires) au programme aussi.

3.1.2. Les niveaux de maîtrise de langue défini par le CECRL

Le Cadre européen commun de référence pour les langues (CECRL) est un document publié par le Conseil de l'Europe en 2001. Ce document définit les niveaux de maîtrise d'une langue étrangère en fonction de savoir-faire dans différents domaines de compétence. Il fournit donc une base commune pour la conception de programmes, de diplômes et de certificats. En 2018, un volume complémentaire avec de nouveaux descriptifs a été publié. Le volume complémentaire introduit de nouvelles échelles, concernant entre autres un enrichissement de la description des niveaux A1 (notamment le niveau pré-A1) et des niveaux C, particulièrement C2 et des descriptions plus complètes des échelles pour la compréhension orale et écrite (Conseil de l'Europe, 2018. Dans le volume complémentaire, des notions clés concernant l'étendue du vocabulaire ont été concrétisées pour chaque niveau de l'échelle (Conseil de l'Europe, 2018). Aucun nombre concret n'est indiqué, mais comme on peut le voir dans la tableau 3-2 ci-dessous, la capacité à périphraser et à varier sa formulation sont des points importants à partir du niveau B1. Au niveau B2, un locuteur est censé être capable de varier sa formulation pour éviter des répétitions fréquentes.

	ÉTENDUE DU VOCABULAIRE
C2	Possède une bonne maîtrise d'un vaste répertoire lexical incluant des expressions idiomatiques et des termes familiers ; est conscient des niveaux de connotation sémantique.
C1	Possède une bonne maîtrise d'un vaste répertoire lexical lui permettant de surmonter facilement les lacunes par des périphrases avec une recherche peu apparente d'expressions et de stratégies d'évitement. Peut choisir entre plusieurs possibilités lexicales dans pratiquement toutes les situations en utilisant des synonymes même pour des mots non familiers. Maîtrise bien les expressions idiomatiques familières et fait des jeux de mots avec facilité. Peut comprendre et utiliser de façon appropriée la gamme de vocabulaire technique et d'expressions idiomatiques propres à son domaine de spécialité
B2	Peut comprendre et utiliser les termes techniques généraux de son domaine, quand il/elle en discute avec d'autres spécialistes.
	Possède une bonne gamme de vocabulaire pour les sujets relatifs à son domaine et les sujets plus généraux. Peut varier sa formulation pour éviter des répétitions fréquentes, mais des lacunes lexicales peuvent encore provoquer des hésitations et l'usage de périphrases. Peut produire assez systématiquement de nombreux mots adéquats dans la plupart des contextes. Peut comprendre et utiliser une grande partie du vocabulaire spécialisé de son domaine mais a des difficultés avec la terminologie d'une spécialité différente de la sienne.
B1	A une bonne gamme de vocabulaire en rapport avec des sujets familiers et des situations quotidiennes. Possède un vocabulaire suffisant pour s'exprimer à l'aide de périphrases sur la plupart des sujets relatifs à sa vie quotidienne tels que la famille, les loisirs et les centres d'intérêt, le travail, les voyages et l'actualité.
A2	Possède un vocabulaire suffisant pour mener des transactions quotidiennes courantes dans des situations et sur des sujets familiers.
	Possède un vocabulaire suffisant pour satisfaire les besoins communicatifs élémentaires. Possède un vocabulaire suffisant pour satisfaire les besoins primordiaux.
A1	Possède un répertoire élémentaire de mots isolés et d'expressions relatifs à des situations concrètes précises

Tableau 3-2 Les notions clés concernant l'étendue du vocabulaire dans le CECRL Conseil de l'Europe (2018)

Milton (2010) a étudié la taille du vocabulaire chez des apprenants d'anglais et français langue étrangère (FLE) afin d'estimer quelle taille du vocabulaire est nécessaire pour chaque niveau de compétence linguistique défini dans le CECRL. Dans cette étude, les niveaux de langue ont été déterminés par des enseignants qui ont placé les apprenants testés dans des

filières d'étude à chacun des niveaux du CECRL. Après, les apprenants ont effectué la version française du test X-Lex afin d'estimer leur taille du vocabulaire dans la langue cible. Le tableau 3-3 montre la relation entre le niveau CECRL des apprenants de français L2 en Grèce et en Espagne et leurs scores sur le test X-Lex, ainsi qu'une estimation de la taille du vocabulaire nécessaire pour atteindre les mêmes niveaux CECRL en anglais.

Niveau CECRL	Estimation de taille de vocabulaire en anglais	Score moyen X-Lex chez des apprenants de FLE en Grèce	Score moyen X-Lex chez des apprenants de FLE en Espagne
A1	<1500	1125	894
A2	1500-2500	1756	1700
B1	2500-3250	2422	2194
B2	3250-3750	2630	2450
C1	3750-4500	3212	2675
C2	4500-5000	3525	3721

Tableau 3-3 Aperçu de la relation entre le niveau CECRL et la taille du vocabulaire chez des apprenant de FLE en Grèce et en Espagne (Milton, 2010)

Cela soulève la question de savoir comment les tailles du vocabulaire peuvent être comparées de manière significative à travers des langues. Une étude sur des corpus français et anglais par Cobb et Horst (2004) suggère que la connaissance des 2000 mots français les plus fréquents dans leur corpus offre un niveau de couverture plus élevé que les 2000 mots anglais correspondants. Cela implique que les résultats des recherches effectuées sur une langue ne seront pas nécessairement les mêmes pour des autres langues. En anglais, les pronoms et les prépositions sont parmi les mots les plus fréquents. D'autre part, en finnois les fonctions remplies par les prépositions sont réalisées par l'addition des suffixes à la forme racine d'un nom ou d'un verbe. Cela conduit au fait qu'en finnois une seule famille de mots peut inclure beaucoup plus de formes de mots que ce ne serait le cas en anglais (Milton, 2010). Ces différences entre les langues produisent alors des nombres de mots différents lorsqu'elles seraient systématisées sur tout un corpus. Il serait donc difficile de tirer la conclusion que la connaissance du même nombre de mots dans des langues différentes signifie que l'on possède forcément la même compétence dans chacune des langues en question (Milton, 2009).

3. 2. Mesurer la taille du vocabulaire réceptif

Nous avons vu au plusieurs reprises que la notion de la taille du vocabulaire est un concept très complexe. Néanmoins, Milton (2009) constate que l'acquisition du vocabulaire dans une langue étrangère est, au moins superficiellement, une qualité qui semble être mesurable ou comptable. De plus, il a été démontré que les estimations de l'étendue du vocabulaire sont de bons indicateurs des compétences linguistiques générales (Batista et Horst, 2016; Milton, 2009; Stæhr, 2008). Par conséquent, plusieurs tests visant à mesurer la taille du vocabulaire ont été développés afin d'estimer le niveau de langue d'un locuteur plus facilement qu'avec des tests de compétence plus extensifs. Les tests mesurant la taille du vocabulaire servent aussi à indiquer où les apprenants ont des lacunes (tests diagnostiques), vérifier si les étudiants ont appris le lexique étudié auparavant (à court terme), ou pour vérifier si un cours a eu du succès en tant qu'enseignement des mots et augmentation du vocabulaire des apprenants (à long terme) (Nation, 2013, p. 515). Comme l'anglais est maintenant l'une des langues les plus utilisées dans le monde, plusieurs tests ont été développés pour estimer la taille du vocabulaire réceptif chez les apprenants d'anglais. Pour les apprenants d'autres langues, les options ne sont pas aussi nombreuses et la plupart des tests visant à mesurer la taille du vocabulaire d'autres langues étaient initialement développés pour l'anglais. Au début de ce projet, nous avons examiné deux options principales pour tester la taille du vocabulaire en français : le test X-Lex et le TTV. Dans la suite, nous allons voir d'abord comment ces tests ont été développés. Ensuite, nous traiterons plus en détail les avantages et les inconvénients du TTV et du X-Lex.

3.2.1. Comment mesurer la taille du vocabulaire réceptif ?

Aujourd'hui, le standard est que les tests visant à mesurer la taille du vocabulaire se basent sur des listes de fréquence d'occurrence des mots dans un corpus. Ceci est dû surtout au fait que les mots les plus fréquents ont tendance à être appris plus tôt dans le processus d'apprentissage. Milton (2009) fait valoir que ceci n'est pas une règle absolue, car les manuels d'apprentissage sont généralement organisés par thème. Par conséquent, les apprenants d'une L2 rencontrent souvent un lexique qui traite les noms des animaux ou les vêtements assez tôt, même si ces mots sont peu fréquents dans le vocabulaire quotidien d'un adulte. Cela vaut particulièrement pour ceux qui apprennent dans un cadre non-naturaliste, autrement dit dans un contexte d'apprentissage formel. Ces apprenants peuvent avoir un profil du vocabulaire déplacé par rapport aux locuteurs natifs. Ils peuvent également avoir des lacunes parmi les

mots les très fréquents, tandis qu'ils connaissent plusieurs mots beaucoup moins fréquents (Milton, 2009; Petitpas, 2010).

Milton (2009) indique que les mots les plus fréquents dans une langue sont presque toujours des mots de structure ou de fonction. Dans le corpus de Baudot (1992) (voir la section 3.2.4), consistant en environ 1,1 million de mots français, les deux mots les plus fréquents représentent 25% du corpus (Milton, 2009, p. 8). Ces mots sont très importants pour former des phrases qui sont grammaticalement correctes et qui portent du sens, mais ils ne portent pas beaucoup de sens en eux-mêmes. Des mots portant plus de sens, comme des noms, des verbes principaux (contrairement aux verbes auxiliaires), des adjectifs et des adverbes sont moins fréquents.

Ce que la fréquence d'un mot nous dit, c'est plutôt la probabilité qu'un apprenant rencontre un mot, et que ce mot est répété si souvent qu'il est appris. Pour identifier la fréquence des mots, ceux-ci sont généralement organisés dans des groupes de fréquence, souvent en groupes de 1000 mots dans chaque niveau. Les 1000 mots les plus fréquents, c'est-à-dire du mot le plus fréquent jusqu'au 999ème mot le plus fréquent, sont regroupés dans ce que nous appellerons désormais la bande de fréquence 1K. Les 1000 mots les plus fréquents suivants (c'est-à-dire le 1000ème mot le plus fréquent jusqu'au 1999ème mot le plus fréquent) sont dans la bande de fréquence 2K et ainsi de suite. Des mots échantillonnés de chaque bande de fréquences sont ensuite utilisés pour constituer les tests de taille du vocabulaire.

3.2.2. Vocabulary Levels Test et Teste de la Taille du Vocabulaire

Le *Vocabulary Levels Test* (VLT) (Nation, 1983; Schmitt, Schmitt et Clapham, 2016) a été conçu pour donner une estimation de la taille du vocabulaire chez les apprenants de l'anglais langue seconde (L2). Le VLT mesure la connaissance des mots appartenant aux bandes de fréquence 2K, 3K, 5K et 10K ainsi qu'un niveau spécial pour les mots anglais académiques. Dans ce test, les participants doivent identifier la définition correcte de 150 mots anglais. Les mots et les définitions sont présentés en grappes. Chaque grappe comprend trois définitions simples et six mots, voir figure 3-2 [2] pour un exemple du format.

À partir de ce format de grappes, Batista et Horst (2016) ont développé un test de vocabulaire français : le *Test de la taille du vocabulaire* (TTV). Ce teste repose sur le même modèle que VLT et les mots dans le TTV viennent des bandes de fréquence 2K, 3K, 5K et 10K. Les mots

[2] La **Feil! Fant ikke referansekilden.** est un exemple tiré du test TTV, mais le VLT a le même format.

dans le TTV ont été sélectionnés à partir de la liste fréquence de Lonsdale et Le Bras (2009) pour les niveaux 2K, 3K et 5K. Cette liste, contenant les 5000 lemmes français les plus fréquents, se base sur un corpus de 23 millions de mots. Les mots dans le corpus sont tirés des textes contemporains écrits et oraux du français international. Contrairement aux autres corpus semblables, cette collection se base sur 50% de sources orales. Afin de faciliter la comparaison entre des études sur l'acquisition du vocabulaire L2 en anglais et en français, Batista et Horst (2016) ont utilisé les travaux de Baudot (1992) pour inclure un niveau des mots dans la bande de fréquence 10K (voir section X pour une description complétive de ce travail). Comme il n'y a pas de liste répertoriant des mots français académiques, le TTV n'inclut pas ce niveau (voir Cobb et Horst (2004) pour une discussion pour savoir si une telle liste est nécessaire en français).

Un principe important pour le format du TTV est que les définitions ont été composées uniquement par des mots appartenant aux niveaux plus fréquents que les mots testés. Autrement dit, la définition pour un terme appartenant à la bande de fréquence 2K se compose uniquement par des mots venant de la bande de fréquence 1K. Les mots testés des autres sections (3K, 5K et 10K) sont définis à l'aide de mots tirés des listes 1K et 2K. Cela permet de s'assurer que les participants peuvent comprendre les définitions fournies. Cependant, le fait qu'un mot est plus fréquent d'un autre ne garantisse pas qu'un locuteur comprends le mot le plus fréquent s'il connaît déjà un mot moins fréquent.

Box 4. A noun cluster from 5K frequency section of the TTV

1. brouillard	
2. coïncidence	
3. farce	_____ une histoire qui fait rire
4. instituteur	_____ ce qui empêche de voir loin
5. pneu	_____ un professionnel de l'éducation
6. soumission	

Figure 3-2 Exemple de grappe de mots dans le test TTV (Batista et Horst, 2016)

3.2.3. Eurocentres Vocabulary Size Test

Le test *Eurocentres Vocabulary Size Test* (EVST), créé par Meara et Jones (1990) a été à l'origine développé pour estimer la proportion des 10 000 mots anglais les plus fréquents connus par un apprenant. Le format du test est une liste de contrôle « oui/non » de 150 mots. Autrement dit, les participants de l'expérience doivent simplement cocher la case à côté d'un mot s'ils connaissent sa signification. Nation (2013) note qu'un tel format de test est facile à gérer et à informatiser pour un marquage rapide. En revanche, il n'est pas possible de vérifier si les participants connaissent vraiment les mots qu'ils ont indiqué connaître. Une caractéristique notable de ce test est l'intégration des mots faux mais plausibles, c'est-à-dire des mots qui ressemblent à de vrais mots dans la langue cible. Ces faux mots sont dispersés parmi les vrais mots. Ils permettent d'estimer le degré de surestimation que fait un apprenant, et sur cette base on peut ajuster les scores. Ils fonctionnent comme une sorte de contrôle des surestimations chez les participants.

1	galpin	[ ]	2	impulse	[ ]	3	suggest	[ ]
4	advance	[ ]	5	peculiar	[ ]	6	benevolate	[ ]
7	indicate	[ ]	8	needle	[ ]	9	destruction	[ ]

Tableau 3-4 Exemple de format du test de la taille du vocabulaire EVST (Batista et Horst, 2016)

3.2.4. Le test X-Lex

En 2003, Meara et Milton (2003) a développé le test X-Lex qui est numérique et très proche au ESVT. Comme le test *Eurocentres Vocabulary Size Test*, le test X-Lex a originellement été développé pour l'anglais. Le test X-Lex estime la connaissance des 5000 mots les plus fréquents dans la langue cible. Le test contient 120 mots venant de six catégories différentes. Les cinq premiers catégories consiste en 20 mots chacun échantillonnés des bandes de fréquence 1K, 2K, 3K, 4K et 5K. Les mots dans le premier groupe ont été choisis parmi les mots dans la bande de fréquence 1K, dans le deuxième groupe ils appartiennent à la bande de fréquence 2K et ainsi de suite. Le dernier groupe consiste en 20 mots inventés, ressemblant à de vrais mots dans la langue cible.

Milton (2009, p. 257-259) a élaboré trois versions françaises du test X-Lex ayant le même format que la version anglaise originale, mais les mots sont tirés d'un ouvrage de Baudot (1992) répertoriant les fréquences d'usage des mots en français écrit. Les mots faux dans les trois versions françaises du test X-Lex ont été créés pour ressembler à de vrais mots français, soit par la composition de syllabes et affixes qui existent déjà dans la langue française

(**crétale, *formirique*), soit par la dérivation de vrais mots français (**provocatif* vs. *provocant/provocateur)*. Certains de ces mots inventés ressemblent à des mots anglais ayant été « francophonisés », par exemple **slendre* (anglais : slender), **vicinité* (vicinity), et **rescuer* (to rescue). Le corpus et les listes de fréquence rédigés par Baudot (1992) se basent sur la lemmatisation et ont été produits à partir d'un corpus de 803 échantillons de textes. Chacun des extraits comptait entre 1000 et 1500 mots, et la plupart des textes dont les extraits sont tirés ont été rédigés entre 1960 et 1967. Les textes se répartissent en 15 genres discursifs et leurs origines se distribuent entre la France (62%), le Canada (37%) et d'autres pays francophones (1%). Quant aux types de publication, 42% des textes viennent de revues et de magazines, 25% de livres et de manuels, 24% de journaux, 7% de bulletins et de rapports et 2 % de brochures et de circulaires.

3.2.5. La qualité des tests TTV et X-Lex

Dans la section 2. 4. nous avons vu que le concept du vocabulaire réceptif n'est pas sans ambiguïté. Tous les tests de vocabulaire réceptifs ont par conséquent ces défauts. Cependant, on considère que les résultats de ce type de test sont valides pour les individus, mais surtout pour les groupes d'apprenants (David, 2008; Milton et Alexiou, 2020). Ici nous n'entrerons pas dans les détails, mais nous mettrons en évidence quand même certaines des faiblesses est des différences entre le test X-Lex et le TTV. Les défauts de ces test ont été analysés plus en profondeur dans des autres études, voir par exemple Beeckmans, Eyckmans, Janssens, Dufranne et Velde (2001), Batista et Horst (2016) et Harsch et Hartig (2016).

L'un des défauts potentiels est les listes de fréquences sur lesquelles s'appuient les tests de vocabulaire. L'un des points forces du test TTV est le fait qu'il se base (sauf le niveau 10K) sur un corpus plus extensif, représentatif et moderne que le X-Lex. Les sections 2K, 3K et 5K du test TTV se base sur les listes de Lonsdale et Le Bras (2009) (voir 3.2.2 pour une description plus détaillé). La section 10K se penche sur le même travaux de Baudot (1992) que le test X-Lex Une caractéristique notable dans le travail de Baudot (1992) est que son corpus se base uniquement sur des sources écrites. Pourtant, le lexique oral est un aspect non négligeable de toute langue. Dans le langage oral, le lexique employé est souvent moins formel. Un exemple est le mot *truc* qui est très fréquent à l'oral mais dans le volume de Baudot, le mot est répertorié comme le 7232ème mot le plus fréquent. Milton (2009) souligne que les informations de fréquence d'un corpus basés sur des sources orales peuvent différer de celles obtenues à partir d'un corpus écrit. C'est pourquoi la plupart des grands corpus

modernes contiennent généralement d'importantes portions de textes transcrits à partir de sources orales variées.

Dans un compte rendu du travail de Baudot, paru la même année que la publication de celui-ci, Lenoble (1992, p. 323) remarque que : « Le répertoire de Baudot se veut représentatif du français écrit contemporain, ce qui peut paraître bizarre puisque la majorité des textes ont une date de publication qui remonte au moins à un quart de siècle. ». Si le lexique dans le corpus paraissait un peu démodé en 1992, c'est bien le cas aujourd'hui trente ans plus tard. Certains mots sont certainement devenus dépassés et des nouveaux mots comme *Internet*, *téléphone portable*, *cotravail* et *écotoxique* ont été introduits dans le langage du quotidien. Pour d'autres mots, comme *application* (ou *appli*), ils ont pris un nouveau sens. Cela peut mener à des changements dans la fréquence des mots. Cependant, les mots les plus fréquents sont pour la plupart des mots de structure et de fonction (Milton, 2009). Leurs profils d'occurrence sont alors peu susceptibles de changer considérablement en 60 ans. Néanmoins, cela renforcerait probablement le test X-Lex-test et la section 10K du test TTV s'ils se basaient sur un corpus plus moderne comprenant des sources orales aussi.

Même si le TTV est basé sur un corpus plus moderne et présente l'avantage supplémentaire de pouvoir contrôler si les participants connaissent vraiment les mots dans le test, nous avons fini par choisir le test X-Lex. Notre choix se justifie pour deux raisons, la première étant qu'il ne suffit pas que les participants accomplissant le test TTV connaissent les mots ciblés, ils doivent aussi comprendre les mots utilisés dans les définitions fournis afin de répondre correctement aux questions. Toutefois, la raison principale pour laquelle nous avons fini par choisir le test X-Lex est qu'il est plus largement utilisé et que la base de comparaison avec les études précédentes est donc plus grande. Autres points à considérer concernant ces tests de vocabulaire disponibles pour le français sont qu'ils nécessitent que les participants aient les compétences de lecture suffisantes et les tests X-Lex et TTV ne conviennent pas à de très jeunes apprenants. Il faut aussi considérer que ces tests ne conviendraient peut-être pas aux participants ayant certains handicaps comme une déficience visuelle ou des troubles spécifiques de l'apprentissage telles que la dyslexie.

3. 3. Etudes précédentes de la taille du vocabulaire chez des apprenants de français

Dans la suite, nous allons d'abord présenter quelques études précédentes visant à estimer la taille du vocabulaire réceptif chez des apprenants de français à l'aide du test X-Lex. La plupart de ces études ont été faites avec des participants anglais, ayant le français comme L2, sauf d'une étude menée en Suède. Ensuite, nous allons examiner comment l'effet de mots cognats peut influencer les résultats des tests de vocabulaire, surtout chez des apprenants d'une L3. Finalement, nous allons considérer quelques études concernant la taille du vocabulaire nécessaire pour les études universitaires de français en Suède.

3.3.1. Études utilisant le test X-Lex

Milton (2008) a utilisé le test X-Lex pour tester la taille du vocabulaire en français chez des élèves et des étudiants au Royaume-Uni. 449 apprenants de tous niveaux dans une école britannique ont passé ce test, tous à la fin de l'année scolaire. Les élèves dans l'étude de Milton (2008) connaissent en moyenne 592 mots après quatre ans d'études (year 10, âgés de 14-15 ans). À la fin de la dernière année de lycée (year 13, âgés de 17 à 18 ans), leur niveau a été estimé à 1930 mots. En utilisant la même méthodologie, Milton a étendu son étude par 29 étudiants dans leur première année d'études de français à l'université. Les étudiants ont accompli le test une fois au début de l'année, en octobre, et de nouveau à la fin de l'année universitaire, en juin. En octobre, leur connaissance a été estimée à 1950 mots. Après leur première année à l'université, leur connaissance moyenne du vocabulaire a augmenté jusqu'à 2555 mots. Les résultats de Milton (2008) montre que les étudiants à l'université continuent à apprendre des mots au même rythme que pendant leurs deux dernières années de lycée (year 12 et 13, A-levels), environ 500 mots par an.

David (2008) a effectué une étude semblable où 483 élèves et étudiants âgés de 12 à 23 ans en Royaume-Uni ont complété le test X-Lex. Dans l'analyse des résultats, les participants qui ont coché plus de cinq mots inventés ont été supprimés des données. David (2008) propose qu'un tel nombre de réponses incorrectes puisse indiquer des niveaux élevés de surestimation et que le participant devine. 66 copies ont par conséquent été exclues, et les analyses se font sur les données de 417 participants. Les données de l'étude de David (2008) ont été recueillies au milieu de l'année scolaire. En moyenne, les élèves en dernière année de lycée (year 13,) connaissaient 2108 mots. Les étudiants en première année à l'université ont obtenu un résultat moyen de 2524 mots. Les résultats de l'étude de David (2008) correspondent bien avec ceux de Milton (2008).

Pignot-Shahov (2014, 2018) a fait deux études plus récentes, également réalisées au Royaume-Uni, sur le développement lexical chez des étudiants de français à l'université. Le projet de Pignot-Shahov consiste en une étude pilote (2014) et une étude principale (2018). Le groupe qui nous intéresse dans ses études est celui des étudiants en première année (first year undergraduate). Comme dans le travail de David (2008), les participants ayant indiqué connaître plus de cinq mots inventés ont été exclus dans ces deux études. Dans l'étude pilote, après l'exclusion, six participants en première année ont accompli la version informatisée du test X-Lex. Les participants ont obtenu un résultat moyen de 3233 mots, le score minimum était 2400 mots et le maximum était 4750 mots. Dans l'étude principale, les participants ont fait ce test du vocabulaire une fois en novembre et encore une fois à la fin du semestre en mai. 12 étudiants en première année ont participé à l'étude principale en novembre et neuf d'entre eux ont accompli le test en mai aussi. Le score moyen en novembre était 2570,83 mots. La version utilisée à la fin du semestre contenait également 20 mots de chacune des bandes de fréquences 6K et 7K ainsi que huit mots inventés supplémentaires. C'est-à-dire que cette version contenait 168 mots par rapport aux 120 mots dans la version standard du test X-Lex. Le score moyen en mai pour les bandes de fréquence 1K à 5K était de 3044,44 mots. Les scores minimum et maximum étaient de 1900 à 4250 mots. Le score moyen, sur 1000 points possibles, dans les bandes de fréquence 6K et 7K était de 327,77 points et 438,88 points, respectivement.

Les études mentionnées ci-dessus ont estimé la taille du vocabulaire chez des apprenants de français en tant que L2. Il existe d'après nos connaissances peu de recherches sur la taille du vocabulaire en français en tant que L3, exception faite de l'étude de Lindqvist (2017). Afin de permettre une comparaison des résultats, Lindqvist (2017) a suivi le modèle proposé par David (2008) de ne pas prendre en compte les tests où plus de cinq mots inventés ont été cochés par le participant. Par conséquent, quatre tests ont été omis et 152 tests sont finalement inclus dans l'analyse. Cette étude a indiqué une taille du vocabulaire moyenne de 1150 mots (n= 34) chez les élèves les plus âgés (en 9ème classe) dans l'étude, c'est-à-dire après quatre ans d'études. Pour comparaison, après quatre ans d'études, les élèves dans l'étude de Milton (2008) connaissaient en moyenne 592 mots. Le chiffre correspondant dans David (2008) est 564 mots (après cinq ans d'études). La différente entre les résultats de Milton (2008) et David (2008) et de Lindqvist (2017) indique que les élèves en Suède ont un vocabulaire considérablement plus large (1150 mots vs. 592 mots) que les élèves au Royaume-Uni, et que leur taille du vocabulaire augmente à un rythme plus rapide aussi.

Il est difficile de trouver des explications pour l'écart entre les résultats dans les deux pays. En Royaume-Uni et en Suède, les élèves apprennent tous le français en tant que langue étrangère dans un contexte formel. Dans la vie quotidienne, ils reçoivent probablement peu d'input français en dehors des cours. Lindqvist (2017) mentionne que les différences pourraient s'expliquer, au moins partiellement, par des méthodes d'enseignement différentes (voir Lindqvist (2017) et Lindqvist et Ramnäs (2016)). Une autre distinction entre les deux pays est que les élèves suédois ont déjà appris une langue étrangère, l'anglais, et Lindqvist (2017) propose la possibilité que les participants suédois perçoivent les similarités entre l'anglais et le français et qu'ils en bénéficient dans cette tâche particulière. Nous reviendrons plus en détails sur cette question dans la section suivante.

Milton (2008), David (2008) et Pignot-Shahov (2018) soulignent tous qu'il y a beaucoup de variations individuelles au sein des groupes de participants. Les étudiants participant aux études au Royaume-Uni ont une connaissance moyenne variant de 1950 mots (Milton, 2008) à 3233 mots (Pignot-Shahov, 2014). Il en va de même pour l'étude suédoise où les scores du groupe de participants le plus âgé (en 9ème classe) vont d'une connaissance de 400 mots à 2500 mots. Dans toutes les études, les résultats montrent que la taille du vocabulaire augmente au fur et à mesure des études du français.

3.3.2. L'effet des mots cognats

Comme nous l'avons vu dans la section 2.3.1, plusieurs recherches ont montré que les similarités entre les langues déjà connues par un apprenant et la langue cible peuvent faciliter la compréhension et l'acquisition de nouveaux mots. En linguistique, le terme *mots cognats* fait référence à des mots qui sont orthographiquement et phonétiquement identiques ou similaires qui se chevauchent sémantiquement, y compris les emprunts et les internationalismes, dans deux ou plusieurs langues (Szabo, 2020). Il existe un bon nombre de mots cognats en anglais et français et il a été estimé qu'il y a environ 6500 mots cognats identiques anglais-français, et environ 17 000 cognats partiels, c'est-à-dire qu'il y a un certain chevauchement en forme ou en sens (Meara, 1993). Au vu de cela, la question se pose de savoir dans quelle mesure ce fait peut influencer les résultats d'un test de la taille du vocabulaire réceptif dans une langue étrangère.

L'effet des mots cognats entre les langues a fait l'objet de plusieurs études actuelles, y compris l'impact que ces mots pourraient avoir sur les tests de connaissance du vocabulaire. Dans un travail récent, Lindqvist (2020) a examiné de plus près l'effet de mots cognats dans son étude

de 2017. Son étude a montré que la L1 et la L2 semblent exercer une influence sur la connaissance des mots en L3. Il est donc possible qu'un participant ait indiqué qu'il connaissait un mot parce qu'il l'a reconnu grâce à sa compétence en anglais L2 au lieu de sa compréhension de français. Par conséquent, les scores des participants suédois sont potentiellement trop gonflés par rapport à leurs niveaux de maîtrise de langue en général. Cela dit, on peut se demander si le nombre élevé de mots cognats en anglais et en français constitue un avantage égal pour les participants britanniques. Allen (2019) et Szabo (2020) proposent que les mots composant des tests de la taille du vocabulaire soient sélectionnés de telle sorte que la proportion de mots cognats dans la langue maternelle et la langue cible soit représentée proportionnellement et répartie de manière appropriée. Ces mesures pourraient améliorer la précision de ces tests et éviter une surestimation de la connaissance du vocabulaire.

3.3.3. Taille du vocabulaire nécessaire aux études universitaires

En Suède, il existe depuis une dizaine d'années une liste de vocabulaire commune à toutes les universités (Lindqvist et Ramnäs, 2016). Cette liste contient des lemmes que les étudiants sont censés apprendre au cours de leur premier semestre à l'université. La liste originale, contenant environ 2700 lemmes, a été élaborée en 2008 par (Lindqvist, Gudmundson et Bardel, 2013) et elle se basait sur le corpus oral CorpAix. Quelques années plus tard, Per Förnegård de l'université de Stockholm effectuait une révision à des buts d'enseignement (Ramnäs, 2019). Il a fait quelques exclusions, mais environ 1300 mots perçus comme utiles pour les étudiants ont été rajoutés. La sélection de ces 1300 lemmes n'a pas été faite à partir de critères scientifiques. La liste compte aujourd'hui un peu moins de 4000 lemmes (Lindqvist et Ramnäs, 2016). Une nouvelle liste pour les étudiants au deuxième semestre a aussi été élaborée en se basant sur *A Frequency Dictionary of French* par Lonsdale et Le Bras (2009) (cf. section 3.2.2). Cette liste consiste en environ 2000 nouveaux lemmes qui ne sont pas répertoriés sur la liste du premier semestre. Après une année d'études à l'université, les étudiants suédois doivent avoir une connaissance d'environ 6000 lemmes (Lindqvist et Ramnäs, 2016).

Rappelons que dans la section 3.1.1 nous avons vu une analyse du vocabulaire de trois textes littéraires au programme du premier semestre d'études de français à l'université de Göteborg. Dans son étude, Ramnäs (2019) souligne que les 4000 lemmes de la liste commune ne donnent pas une couverture aussi élevée que les 4000 lemmes les plus fréquents du dictionnaire de Lonsdale et Le Bras (2009), un fait qui s'explique probablement par le mode de sélection des mots. De plus, la liste de vocabulaire commune des universités suédoises de

4000 lemmes est loin de fournir une lecture fluide bien que les romans aient été sélectionnés pour leur relative simplicité. Elle équipe les étudiants avec une couverture de presque 95%, ce qui est quand même un bon fondement, mais Ramnäs (2019) conclut que ces textes sont trop difficiles pour permettre aux étudiants de ce niveau un apprentissage du vocabulaire implicite.

En ce qui concerne la taille du vocabulaire nécessaire pour les études universitaires de français en Norvège, nous ne connaissons aucune recherche antérieure sur le sujet. En outre, il n'existe pas de telle liste de vocabulaire commune pour les étudiants norvégiens. Toutefois, nous savons que les élèves sont censés avoir atteint le niveau A1 après 10ème, c'ést-à-dire après collège (Utdanningdirektoratet, 2020b). Après lycée, après cinq ans d'étude au total, les élèves qui ont choisi un deuxième langue étrangère sont censés avoir atteint le niveau A2, (Utdanningdirektoratet, 2020a). De plus, les connaissances préalables recommandées pour étudier le français à l'université en Norvège sont le niveau CECRL A2. Nous pouvons supposer que plupart des étudiants commencent leurs études de français avec un niveau de langue A2, surtout parce qu'ils ont choisi d'étudier le français et sont alors probablement motivés et intéressés à apprendre le français.

Chapitre 4. Méthodologie

Dans ce chapitre, nous décrirons les participants de notre enquête et les instruments de mesure choisis, notamment le test de vocabulaire et le questionnaire biographique. De plus, nous rendrons compte de la collecte de données et de la méthode d'analyse.

4. 1. Participants

Tout d'abord, l'étude a été annoncée auprès du médiateur pour la recherche en charge de la confidentialité, Norsk senter for forskningdata (NSD). Tous les étudiants de la première année d'études de français à une université en Norvège ont été invités à participer et 24 étudiants âgés de 18-25 ans ont accompli le test de vocabulaire et le formulaire biographique. Ils ont utilisé en moyenne environ 10 minutes. Après l'exclusion des participants ayant indiqué plus de cinq mots inventés, notre échantillon d'analyse compte 16 étudiants, âgés de 18 à 25 ans. La plupart d'entre eux, dix étudiants, ont commencé à étudier le français à l'école en 8ème (âgés de 12 à13 ans) et quatre étudiants commençaient au lycée (VGS, âgés de 16 à 19 ans). Deux étudiants n'avaient pas eu du tout de cours de français à l'école. 3 participants ont passé plus de 3 mois dans un pays ou une région francophone : la France, le Québec au Canada et la Wallonie en Belgique. Il est difficile d'estimer combien d'heures d'étude du français chaque étudiant a à son actif, mais ceux qui ont suivi le parcours « standard » de cinq ans de français à l'école ont bénéficié de 447 heures, 222 heures au collège (ungdomsskolen, 8.-10.) et 225 heures au lycée (VG1 + VG2).

Nous avons demandé aux participants d'estimer combien d'heures par semaine ils entendent le français en dehors de l'université. La plupart, sept étudiants, affirment qu'ils entendent parfois le français hors de l'université, mais moins d'une heure par semaine en moyenne, six étudiants ont indiqué entre une et trois heures, deux entendent plus de 3 heures par semaine en moyenne et un participant avait coché qu'il n'entendait jamais le français à l'extérieur de l'université. Ceci démontre une certaine hétérogénéité parmi les participants de sorte qu'il sera difficile d'établir des liens entre ces données de base et les résultats de l'étude.

Nous leur avons demandé d'évaluer leur propre niveau de compétence en anglais et en français sur une échelle de cinq niveaux (voir la section 5. 3. et l'annexe II) ; basique, intermédiaire, indépendant, compétent et courant (*grunnleggende, middels, selvstendig, kompetent, flytende*). Tous les participants dans notre analyse ont le norvégien pour seule langue maternelle. Les Norvégiens ont généralement une compétence relativement élevée en anglais grâce à l'input pratiquement quotidien provenant de différents médias et réseaux

sociaux. Les réponses au formulaire biographique confirment que nos participants estiment bonne leur propre connaissance en anglais.

4. 2. Procédure

Les participants dans notre étude sont tous des étudiants dans leur premier semestre d'études françaises à une université en Norvège. Tout d'abord, nous avons contacté les professeurs de différents cours de première année pour leur demander la permission de présenter l'étude aux étudiants pendant la pause d'un cours magistral. Les étudiants ont été informés des conditions de participation, des objectifs de l'étude et du temps estimé nécessaire pour remplir l'enquête. Ils ont été assurés que les résultats du test n'affecteraient pas leur notation dans les cours et que seul le chercheur et la responsable de l'étude allaient avoir accès aux résultats des tests. Après la présentation du projet, une feuille a été distribuée aux étudiants avec un résumé des détails sur l'étude et un code QR. Scanner le code avec un téléphone portable menait au lien pour participer à l'étude. Les professeurs ont également publié le lien vers l'étude à la plateforme d'apprentissage en ligne.

À cause du risque concernant le COVID-19, nous avons effectué l'étude à l'aide d'un questionnaire digital, *Nettskjema*[3]. Le questionnaire digital comportait trois parties ; a) des informations sur l'étude et le formulaire de consentement (voir annexe I) , b) le test de vocabulaire (cf. la figure 4-1et la figure 4-2 et c) un questionnaire biographique (voir annexe II). Après avoir signé le formulaire de consentement, les participants ont été réexpédiés de façon aléatoire vers l'une des trois versions du test de vocabulaire par le site Nettskjema. Finalement, les participants ont rempli le questionnaire visant à recueillir des informations supplémentaires concernant : leur âge, leur langue maternelle, l'âge auquel ils ont commencé à apprendre le français, une auto-évaluation de leur niveau de français etc. La collecte de données a eu lieu entre mi- et fin octobre, c'est-à-dire environ deux mois après le début du semestre.

4.2.1. Test pour mesurer la taille du vocabulaire

Pour mesurer la taille du vocabulaire de nos participants, nous avons choisi la version française du test X-Lex, élaboré par Meara et Milton (2003). Ce test mesure le vocabulaire réceptif, c'est-à-dire qu'il cible le nombre de mots compris par les participants. Même si ce test a ses inconvénients, comme nous venons de le voir dans la section 3.2.5, le X-Lex est

[3] https://www.uio.no/tjenester/it/adm-app/nettskjema/

largement utilisé, ce qui nous permet de comparer nos résultats avec bon nombre d'études antérieures.

Nous avons suivi la méthodologie de Lindqvist (2017) au plus près afin de former la meilleure base de comparaison de résultats. Nous avons par conséquent adopté le format où les participants cochent soit *oui, je peux reconnaître, comprendre ou utiliser ce mot en français* soit *non, je ne reconnais pas, ne comprends pas ou ne peux pas utiliser pas ce mot en français* pour chaque mot dans le test. Comme David (2008) et Lindqvist (2017), nous avons distribué trois versions du test X-Lex, qui alors constitue trois versions différentes du même test. Utilisant trois versions différentes nous permet de tester plus de mots et cela nous a donné des options en cas de problème avec l'une des versions. David (2008) a conclu qu'il n'y avait pas de différence statistiquement significative dans les résultats obtenus à partir des trois versions. Nous avons utilisé les mêmes trois versions du test français fourni par Milton (2009, p. 257-259), avec quelques modifications.

Rappelons que le test X-Lex consiste en une liste de mots. Chaque version comprend 100 mots français réels et 20 mots inventés qui ressemblent à de vrais mots français du point de vue de la morphologie et de l'orthographie. Ces mots faux sont une manière de remédier à la surestimation parce que des points sont déduits du score total pour chaque faux mot coché. Nous avons voulu garder le contenu du test dans sa version originale. Cependant, étant donné que certains mots inventés sont maintenant répertoriés dans le dictionnaire, nous avons décidé de les remplacer. Il s'agit des mots *pédiment* et *satisfactoire* de la version 2 et 3 du test. Ces mots ont été remplacés par des mots inventés de la version 1 du test, en l'occurrence **manchir* et **diroir,* respectivement. Nous avons aussi remplacé **spirité* dans la version 3 par **jerette* (mot inventé en version 1), car le mot *spirite* est un vrai mot français ; c'est donc seulement l'accent aigu qui rend le mot faux.

Dans notre étude, pour chaque version du test, l'ordre interne de tous les 120 mots a été quasi-randomisé. D'abord, tous les mots d'une version du test ont été mis dans un ordre complètement aléatoire à l'aide d'un site internet, ListRandomizer[4]. Puis, nous avons fait certains changements pour placer quelques mots simples, que les étudiants connaissent probablement, tels que *tante, que, métro* et *nuit,* au début et à la fin du chaque tableau de mots. Ceci était une mesure pédagogique pour ne pas décourager les participants, étant donné

[4] https://www.random.org/lists/

que le test X-Lex contient un bon nombre des mots difficiles que les participant ne connaissent probablement pas.

Les instructions aux participants expliquaient qu'il y avait des mots faux parmi les mots vrais. Lors de l'enquête, les participants devaient cocher les mots qu'ils reconnaissaient ; voir des exemples ci-dessous en figure 4-1 et figure 4-2. Les instructions et les options de réponse étaient fournies en norvégien pour assurer que les participants comprennent correctement la tâche.

Il était possible d'accomplir le test sur un ordinateur ou sur un smartphone. La mise en page du test est apparue assez différemment sur un ordinateur (voir figure 4-2) par rapport à un téléphone portable ou une tablette (figure 4-1). Nous ne nous sommes rendu compte de cette différence qu'après la réalisation du test. Aucune option n'était disponible pour modifier la mise en page afin de les rendre identiques, ni pour voir quel type d'appareil les étudiants utilisaient pour répondre à l'enquête. Cependant, il est peu probable que cette différence de mise en page ait eu de l'effet sur le résultat. Dans la version d'ordinateur, il y avait 2 listes par page contenant 15 mots chacune (au total 30 mots par page), ce qui donne un total de quatre pages de test de vocabulaire. Nous avons choisi ce format afin de briser la monotonie de cliquer un mot après l'autre 120 fois sur une longue liste. Dans la mise en page sur un smartphone, la partition en deux listes n'était pas très évidente. Les mots se suivaient sur une liste de 30 mots par page, et donc sur quatre pages au total. Les participants devaient cocher « oui » ou « non » pour tous les mots d'une page afin de continuer à la page suivante.

Figure 4-2 La mise en page à un téléphone portable ou une tablette

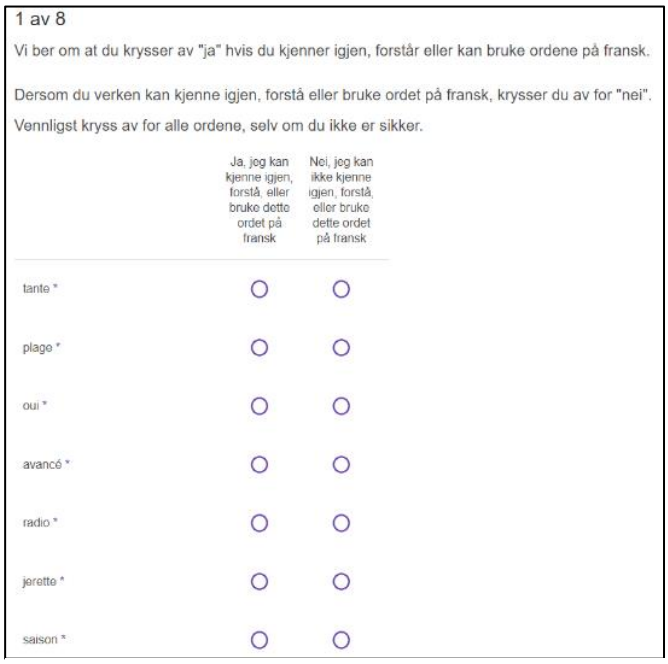


Figure 4-1 La mise en page à un ordinateur

Après la collecte de données, nous avons malheureusement trouvé quelques fautes de frappe dans les tests de vocabulaire. Dans la version 2, le vrai mot *extrêmement* est devenu le faux mot **extrêmenet*. Dans la version 3 le mot qui aurait dû être *révéler* est devenu **réléver*, *participer* est devenu **perticiper, salarié* est devenu **salairé* et *débrouiller* est devenu **débroiller*. La section 4.2.2 traite en détail de l'impact que ces défauts ont pu avoir.

4.2.2. Calcul de scores

Le test X-Lex donne un score à chaque participant. Une réponse correcte vaut 50 points. Le score maximal du test est donc 5000 (50 points × 100 mots vrais). Ceci correspond avec le fait que le test vise les 5000 mots les plus fréquents de la langue française. Une réponse incorrecte, c'est-à-dire cocher la réponse « oui » pour un mot inventé, implique une déduction de 250 points. Les réponses « non » pour un vrai mot ne donnent pas de points mais aucune réduction de points non plus, c'est-à dire 0 points.

Dans son étude en Suède, Lindqvist (2017), a choisi de suivre le modèle de David (2008). Ce modèle propose de ne pas prendre en compte les réponses où un participant a indiqué connaître plus de cinq mots inventés. Une telle fraction de réponses incorrectes pourrait indiquer que le participant en question devine. Nous avons décidé de suivre ce modèle, ce qui nous a mené à omettre huit réponses: deux réponses de la version 1, quatre de la version 2 et deux de la version 3. Nous avons par conséquent analysé les résultats des seize participants, quatre en version 1, et six chacun des versions 2 et 3.

4.2.3. Calcul de points par suite des fautes de frappe

En ce qui concerne les fautes de frappe mentionnées en 4.2.1, elles ont des implications pour le calcul de scores des participants ayant répondu aux versions 2 et 3 du test X-Lex. Comme toutes les fautes de frappe sauf une se trouvent dans la version 3, nous avons considéré la possibilité que les résultats issus de cette version du test ne soient pas valides, et par conséquent qu'ils ne puissent pas être comparés avec les résultats issus des deux autres versions. Nous avons considéré supprimer les résultats de la version 3 et de ne conserver que les résultats des versions 1 et 2. Cela ne nous laissait que 10 réponses à analyser, mais nous pouvions peut-être encore en extrapoler des résultats intéressants. Cependant, nous préférions conserver les résultats de la version 3 afin d'inclure autant de participants que possible dans

notre étude. Nous avons donc examiné plusieurs alternatives afin de pouvoir ajuster les scores de manière à assurer la validité des résultats de la meilleure façon possible.

Version	Mot correcte	Faute de frappe	Nombre de participants ayant coché « oui »	Bande de fréquence
2	extrêmement	**extrêmenet*	0	2K
3	révéler	**réléver*	2	1K
3	participer	**perticiper*	1	1K
3	salarié	**salairé*	4	4K
3	débrouiller	**débroiller*	0	5K

Tableau 4-1 Aperçu des fautes de frappe

Le tableau 4-1 donne un aperçu des fautes de frappe. La plupart de fautes sont très proches des vrais mots français. Les participants savent qu'il y a des mots inventés ressemblant des vrais mots français intégrés dans le test, mais pas combien et à quel point ces mots inventés sont proches des vrais mots. La faute **réléver* par exemple, est très proche du verbe *relever*. Il y a seulement les accents aigus qui rendent cette faute de frappe fausse. Le mot *relever* est aussi dans la bande de fréquence 1K selon les travaux de Baudot (1992). La faute de frappe indiquée comme connue par le plus grand nombre de participants est **salairé* ; 4 sur les 6 participants assignés à la version 3 du test ont répondu qu'ils connaissaient ce mot. Ce mot mal orthographié est très proche des vrais mots *salaire* et *salarié.*

Un autre facteur que nous avons pris en compte est le fait que le mot *débrouiller* se trouve dans la bande de fréquence 5K, on peut ainsi supposer que les participants ne connaissent probablement pas ce mot quoi qu'il en soit, et que leur score sur ce mot serait resté le même indépendamment de la faute de frappe. Nous avons alors considéré la possibilité que les participants n'avaient pas remarqué certaines fautes de frappe parce qu'ils sont si proches de l'orthographe correcte. Une autre éventualité est que les participants se sont rendu compte qu'il s'agit d'une faute de frappe pour certains des mots mal orthographiés. Toutefois, examiner ces mots mal orthographiés individuellement et essayer de déterminer si les participants se sont rendu comptes de fautes ou pas, et s'il est probable qu'ils connaissent le mot en premier lieu, ce ne sera que supputations et conjectures.

Nous avons par conséquent examiné quelques stratégies possibles pour calculer les points. Un principe central pour toutes ces stratégies est que le score maximum reste 5000 points pour chacune des trois versions du test. Le tableau 4-2 ci-dessous présente les scores moyens de chaque version du test et la figure 4-3 présent les scores de chaque participant en version 3

obtenus en suivant les différentes stratégies du calcul de points proposées dans la liste suivante :

1. Considérer les mots contenant des fautes de frappe comme des mots inventés. Cela se traduira en une déduction de 250 points si le participant a coché « oui » pour les mots mal orthographiés et une récompense de 50 points si le participant a coché « non » pour ces mots.
2. Seulement accorder 50 points si un étudiant a coché « non » pour les mots mal orthographiés, mais aucune déduction pour avoir coché « oui ».
3. Supprimer les mots contenant des fautes de frappe, c'est-à-dire ne pas donner de points pour ces mots, mais aucune déduction de points non plus. Dans ce cas, il faut ajuster le score pour chaque mot correct afin que le score maximum soit toujours 5000 points.
4. Suivre la même stratégie que pour l'option 3, mais en outre retirer 250 points pour avoir coché un mot contenant une faute de frappe.
5. Exclure cinq mots des bandes de fréquences correspondant aux mots mal orthographiés dans chaque version du test. De cette manière, il nous restera 95 mots corrects dans chaque version du test. Les points pour ces mots doivent être ajustés afin que le score maximum reste 5000 points.

Si l'on suit la première option, le score maximum restera 5000 points parce que les participants seront récompensés de 50 points en cochant « non » pour les mots contenant des fautes de frappe. Cette option n'est pas conforme au principe du design du test car elle attribue des points aux participants indiquant qu'ils *ne connaissent pas* certains mots. En outre, les participants assignés à la version 3 risquent une déduction supplémentaire de 1000 points parce qu'il y a plus de non-mots que dans les autres variantes du test. Cette méthode de calcul donnera aux participants un score artificiellement faible. Ce risque d'une trop grande déduction de points est éliminé si nous suivons la deuxième stratégie. Pourtant, cette option rompt également avec le principe du test.

En suivant la troisième stratégie proposée, il faut ajuster les points attribués aux mots corrects restants afin de garder un score maximum de 5000 points. Cela veut dire que les mots corrects en version 3 donnent 52,08 points ($\frac{5000\ points}{96\ mots\ corrects} \approx 52{,}08$) et les mots corrects en version 2 donnent 50,51 points chacun au lieu de 50 points comme dans la version 1. Cette stratégie implique que la répartition entre mots corrects et mots faux est décalée. Par conséquent, les participants de la version 3 peuvent obtenir un score légèrement élevé comparé aux participants de la version 1 et 2, étant donné qu'il y a moins de mots corrects qui doivent être connus pour obtenir le score maximum. Nous avons alors considéré la quatrième option, qui en plus implique une déduction de 250 points pour chaque mot mal orthographié indiqué comme connu, mais comme nous l'avons vu pour l'option 1, cette stratégie présente le risque d'une très grande déduction de points pour les participants assignés à la version 3.

La dernière alternative que nous avons considérée est d'éliminer les mots mal orthographiés en excluant cinq mots de chacune des versions du test. Deux mots de la bande de fréquence 1K, et un mot chacun des bandes de fréquence 2K, 4K et 5K seraient, dans ce cas, exclus de la version 1. Le mot **extêmenet* (1K), deux mots de la bande de fréquence 2K et un mot chacun de bandes de fréquence 4K et 5K seraient exclus de la version 2. Finalement, dans la version 3, les quatre mots contenant des fautes de frappe ainsi qu'un mot de la bande de fréquence 2K seraient exclus. Pour garder le score maximal possible à 5000 points, le score pour les 95 mots qui restent doit être ajusté de cette manière : $\frac{5000\ points}{95\ mots} \approx 53{,}63$ points. L'avantage de cette stratégie est qu'il y aura le même nombre de mots corrects dans chaque version du test. Néanmoins, cette méthode impliquera une intervention plus étendue que nécessaire dans le design original du test, en particulier pour la version 1 où il n'y avait pas des fautes de frappe.

La figure 4-3 montrent les scores selon les différents alternatives pour les participants assignés à la version 3, qui est la version avec le plus de fautes, et où les choix de stratégie influent le plus. Les différentes méthodes de calcul de points ont le plus d'impact sur le résultat du participant 2. Ce participant a indiqué connaitre trois sur les quatre mots contenant des fautes de frappe en version 3 et pour son score il s'agit d'une différence de 830 points entre les alternatives 1 et 3. Pour les autres participants, l'écart entre les résultats n'est pas si considérable.

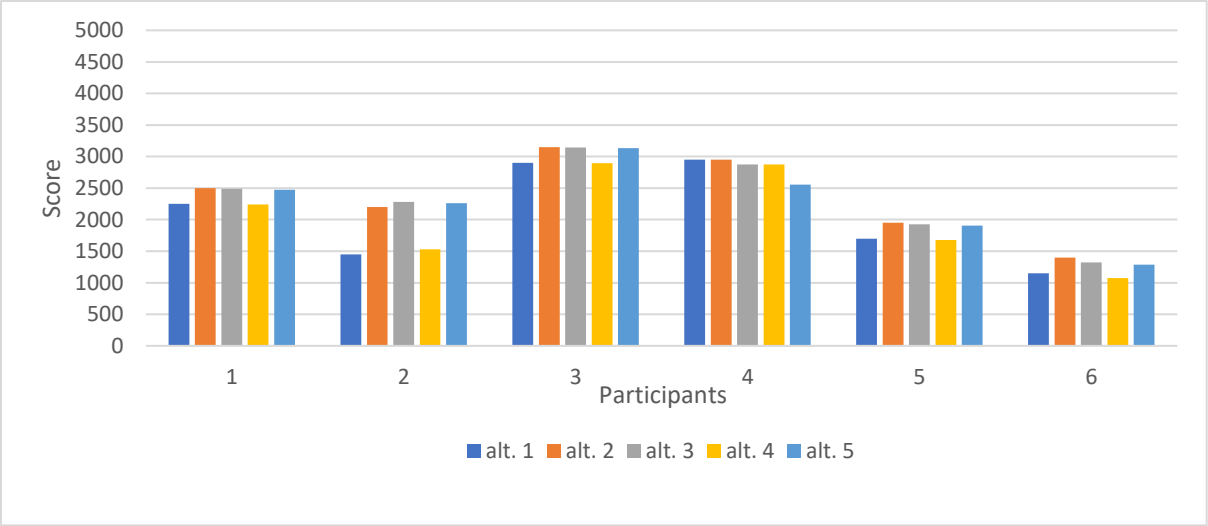


Figure 4-3 Les scores de chaque participant en version 3 selon les différentes méthodes de calcul de points

	alternatif 1	alternatif 2	alternatif 3	alternatif 4	alternatif 5
version 1	2487,5	2487,5	2487,5	2487,5	2516,4
version 2	2558,3	2558,3	2538,3	2538,3	2532,8
version 3	2066,6	2358,3	2340,3	2048,6	2269,8

Tableau 4-2 Le score moyen de chaque version du test selon l'alternative de calcul de points.

Après mûre réflexion, nous avons opté pour l'alternative 3. Nous pensons que cette méthode de calcul de points maintient suffisamment l'équilibre entre le risque de surestimation et de sous-estimation du score de chaque participant, sans pour autant trop altérer le design original du test. De ce fait, cette méthode nous permet de garder les résultats de la version 3.

Chapitre 5. Résultats

Dans ce chapitre, nous allons décrire et analyser les résultats du test X-Lex. D'abord nous allons analyser les résultats du test de vocabulaire. Ensuite, nous allons voir la distribution de bonnes réponses selon bande de fréquence. Finalement, nous allons considérer l'auto-évaluation de niveau de maîtrise de langue et lesquels parmi les mots inventés qui sont les plus cochés.

5. 1. Taille du vocabulaire

Après l'exclusion des participants ayant coché trop de mots inventés, (cf. David (2008) et Lindqvist (2017)), 16 participants sont inclus dans notre étude. Il nous reste quatre réponses de la version 1, et six de chacune des versions 2 et 3. Le tableau 5-1 montre les résultats des tests de la taille du vocabulaire pour les versions respectives ainsi que le moyen pour chaque version. Les résultats montrent que la connaissance varie de 1323 mots à 3600 mots et le score moyen est de 2451 mots. Nous avons aussi calculé le moyen sans les réponses de la version 3 étant donné qu'il y avait des fautes de frappe (voir 4.2.3).

Version	n	Minimum	Maximum	Moyen	Moyen total	Moyen sans version 3
1	4	1550	3600	2488	2451	2518 (n=10)
2	6	1879	3540	2538		
3	6	1323	3146	2340		-

Tableau 5-1 Résultats au test de taille du vocabulaire

La tableau 5-1 montre la variance des scores dans chaque version du test. Les croix dans les boites sont les moyennes et les barres horizontales centrales sont les médianes. Les moustaches montrent les scores minimums et maximums. Les limites inférieures et supérieures des boîtes sont les premier et troisième quartiles, respectivement. 50% des données centrales se trouvent entre ces deux limites, c'est-à-dire que 25% se trouvent au-dessus et 25% en-dessous.

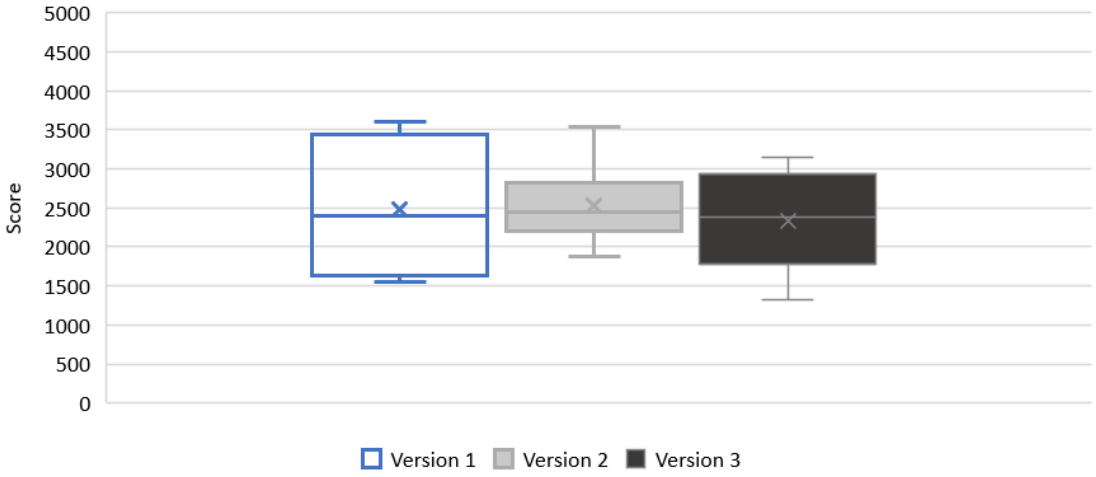


Figure 5-1 Maximum, minimum et score moyen pour chaque version du test

5. 2. Distribution selon bande de fréquence

Nous avons calculé le pourcentage de connaissance pour chaque bande de fréquence. Ces résultats, présentés dans la figure 5-2 montrent le pourcentage de mots qui ont été indiqués comme connus par les participants dans chaque bande de fréquence pour les trois versions du test combinées. Les nombres ont été corrigés des fautes de frappe dans la version 3. Ce profil correspond bien avec les courbes escomptées et présentées par Milton (2009, p. 27) : les apprenants connaissent typiquement le plus de mots dans la bande de fréquence 1K, et le moins dans la bande 5K. Nous voyons que les participants connaissent en moyenne 89% des 1000 mots les plus fréquents. Le pourcentage de connaissance diminue progressivement à

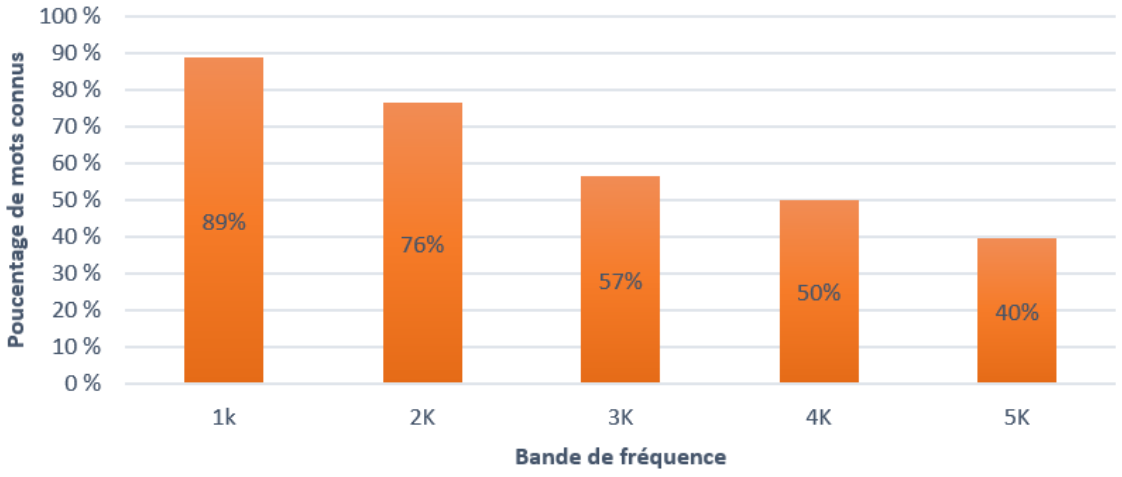


Figure 5-2 Connaissance moyenne des mots selon bande de fréquence.

mesure que le niveau de fréquence s'amoindrit. Dans la bande de fréquence 5K, les participants ont annoncé qu'ils connaissent en moyenne 40% des mots.

Nous avons aussi calculé la distribution de mots connus selon la bande de fréquence pour chacune des versions individuellement. Les résultats sont présentés dans la figure 5-3 ci-dessous. Nous voyons que les profils varient plus, mais nous pouvons constater que la tendance générale est la même ; les participants connaissent plus de mots dans les bandes les plus fréquentes. En analysant ces résultats, il faut se rappeler que cet écart pourrait notamment s'expliquer par la petite taille de notre échantillon.

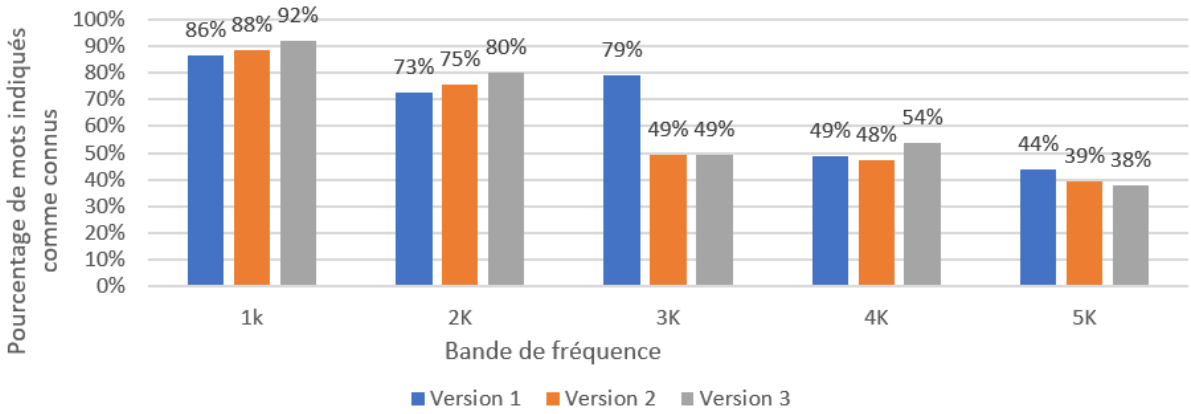


Figure 5-3 Connaissance des mots selon bande de fréquence pour chaque version du test

Pour vérifier si cette tendance continue quand nous regardons les résultats des participants individuels, nous avons également évalué selon la bande de fréquence les score des trois participants, notamment les participants ayant obtenu le score le plus élevé et le plus bas ainsi qu'un participant ayant obtenu un score proche du score moyen. Nous avons fait cela pour voir s'il y une différence entre les profils d'un participant ayant obtenu un score élevé et un participant ayant obtenu un score bas. La figure 5-4 montre le nombre de mots indiqués comme connus dans chaque bande de fréquence pour ces trois participants. Nous voyons que le participant ayant marqué le plus de points (3600) a indiqué connaître tous les mots dans les bandes de fréquence 1K et 3K, et 95% et 90% des mots dans les bandes de fréquence 2K et 4K respectivement. Même pour les mots les moins fréquents, c'est-à-dire dans la bande de fréquence 5K, ce participant a indiqué connaitre 75% des mots. Ce participant a donc une bonne connaissance des mots dans chaque bande de fréquence. La raison pour laquelle le score de ce participant n'est que 3600 points est qu'il a coché « oui » pour quatre mots inventés et ainsi il a perdu 1000 points.

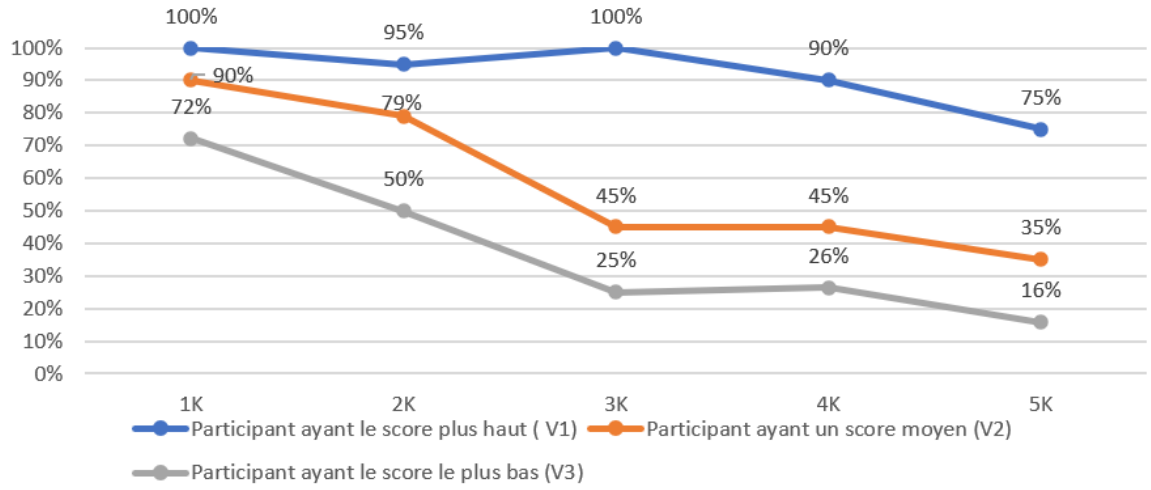


Figure 5-4 Les résultats du test de vocabulaire pour trois participants selon bande de fréquence

Le participant ayant obtenu le score le plus bas, 1323 points, a indiqué connaitre 72% des mots dans la bande de fréquence 1K, 50% en 2K, 25% en 3K, 26% en 4K et seulement 16% des mots dans la bande de fréquence 5K. Ce participant obtient un score significativement plus bas dans toutes les bandes de fréquences que le participant avec le score le plus élevé. Son résultat est aussi plus bas que la moyenne globale pour le test (cf. figure 5-2), surtout dans les trois bandes les moins fréquentes. En revanche, ce participant a indiqué connaître seulement deux mots inventés, ce qui menait à une déduction de 500 points.

Nous voyons que le participant ayant obtenu le score le plus élevé a une connaissance très élevée des mots dans toutes les quatre premières bandes de fréquence et il indique connaitre un nombre relativement haut dans la bande 5K aussi. Le participant moyen et le participant ayant obtenu le score le plus bas connaissent le plus de mots dans les bandes 1K et 2K, mais leurs connaissances dans les bandes moins fréquentes sont plus basses. Cela montre que c'est surtout dans les bandes 3K, 4K et 5K que ces participants ont un déficit de connaissance des mots par rapport au participant ayant obtenu le score le plus haut. Le participant moyen et le participant ayant obtenu le score le plus bas ont un profil de notation très similaire, leur connaissance des mots diminue avec chaque bande de fréquence, mais le participant qui a obtenu un score moyen a une meilleure connaissance globale dans toutes les bandes de fréquence par rapport au participant ayant obtenu le score le plus bas dans notre étude.

5. 3. Auto-évaluation de niveau de maîtrise de langue

Dans le questionnaire biographique, nous avons demandé aux participants d'évaluer leurs propres compétences en anglais et en français sur une échelle de cinq niveaux ; 1-basique, 2 - intermédiaire, 3 -indépendant, 4 – compétent, et 5 – courant [5]. Le figure 5-5 ci-dessus montre la relation entre le score du test X-Lex des participants et leur auto-évaluation de niveau de langue en français et en anglais. La plupart des étudiants ont estimé leur niveau de français intermédiaire. 11 des participants ont répondu qu'ils ont niveau anglais courant. Il y a une bonne concordance entre le score du test X-Lex et l'auto-évaluation de niveau de langue. Cela est conforme au fait que la taille du vocabulaire est largement corrélée au niveau de compétence linguistique. Le participant ayant obtenu le score le plus haut est le seul à estimer sa compétence en français à niveau 4 – compétent. Le participant ayant obtenu le score le plus bas a indiqué que son niveau de français est intermédiaire (2). Le score moyen d'auto-évaluation de niveau de langue est 2,38 en français et 4,56 en anglais.

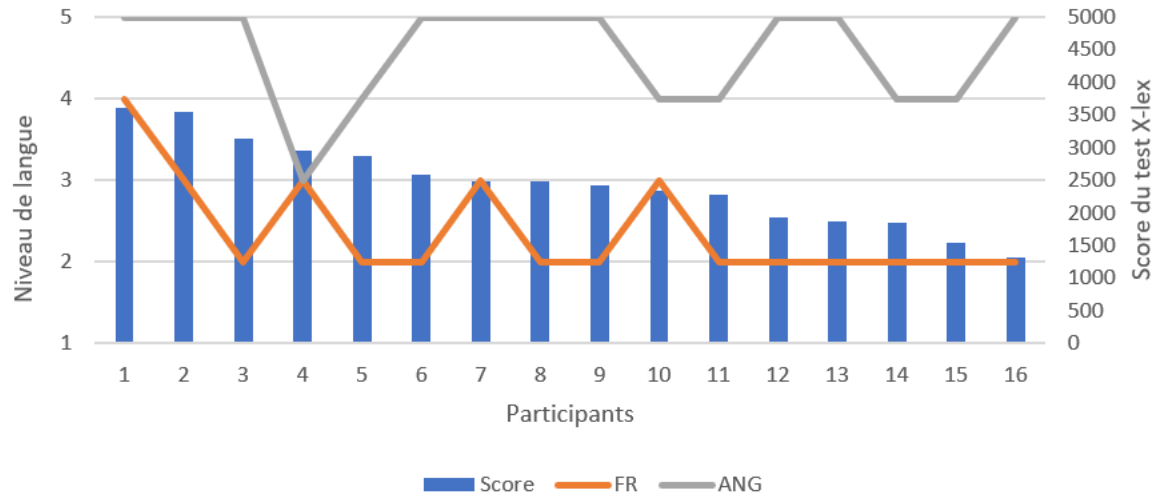


Figure 5-5 Corrélation entre les auto-évaluations de niveau de langue en français et en anglai et les scores du test X-Lex

[5] 1- Basique: Je peux reconnaître et utiliser des mots et des expressions simples. 2- Intermédiaire: Je peux utiliser des mots courants et des phrases simples. Je peux communiquer sur un niveau basique. 3- Indépendant: Je peux comprendre le contenu principal d'un discours standard et écrire/parler de manière cohérente sur des sujets familier. 4- Compétent: Je peux parler assez couramment et comprendre un discours cohérent sur des sujets difficiles et familiers. Je peux écrire en détail sur de nombreux sujets. 5 -Courant: je peux m'exprimer couramment, spontanément et précisément. Je comprends facilement tout ce que je lis et j'entends et je peux utiliser la langue efficacement dans des contextes sociaux et professionnels.

5. 4. Quels mots inventés ont été cochés ?

En section 3.3.2, nous avons mentionné les travaux de Lindqvist (2017, 2020) traitant de l'effet de *cognate advantage* sur les vrais mots français que les participants ont signalé connaître. Nous avons étudié la relation entre les mots inventés indiqués comme connus et leur ressemblance avec de vrais mots anglais. Le tableau 5-2 ci-dessous montre le nombre de participants ayant indiqué connaitre les différents mots inventés de chaque version du test et éventuellement leur mot «cognat » en anglais. Dans ce cas, il ne s'agit pas de vrais cognats car les mots inventés n'existent pas en français. Nous observons ici une tendance intéressante : les mots inventés ressemblant à de vrais mots anglais sont les plus cochés. Par exemple, tous les participants ont coché le mot **expecter* en version 1, forme très proche au verbe *to expect* en anglais. Les mots faux qui n'ont pas d'équivalent en anglais comme **nadoir*, **luvois* et **gillais*, semblent, en général, être moins cochés.

Version 1 (n=4)		**Version 2 (n=6)**		**Version 3 (n=6)**	
expecter (*to expect)*	4	ministeur (*minister)*	3	provocatif (*provocative)*	4
disabilité (*disability)*	3	liabilité (*liability)*	2	rescuer (*to rescue)*	4
entrance (*entrance)*	3	permissable (*permissible)*	2	ultimation (*ultimation)*	3
dour (*dour)*	2	fronter (*to front)*	1	manchir	2
défaulter (*defaulting*)	1	grasper (*to grasp)*	1	talenté (*talented)*	2
crétale	0	froise	1	reparlance	1
abjecter (*abject(ion))*	0	litéracie (*literacy)*	1	houroux	1
arguable (*arguable)*	0	naçon	0	jerette	1
euplain	0	garmente (*garment*)	0	vicinité (*vicinity)*	1
elstrisse	0	giste (*gist)*	0	vernique	0
formirique	0	piédeur	0	gillais	0
signard	0	outrir	0	diroir	0
écourt	0	triparoix	0	brigeable	0
gestide	0	joyance (*joyance* -archaïque)	0	porvent	0
précont	0	abtrâte	0	slendre (*slender)*	0
jerette	0	luvois	0	toutceul	0
diroir	0	malignant (*malignant)*	0	statutoire (*statutory)*	0
lifrer	0	manchir	0	aperne	0
nadoir	0	prévieux	0	introis	0
tirôt	0	soupaire	0	siéve (*sieve*)	0

Tableau 5-2 Le nombre de participants ayant indiqué connaître les mots inventés et leurs faux cognats en anglais

Nous avons aussi certains cas limites où les « cognats » sont moins évidents, nous pouvons par exemple penser que le mot faux **prévieux* peut ressembler au mot anglais *previous*. La fréquence d'occurrence du mot correspondant en anglais peut influencer le fait que les faux mots soient ou non cochés. Il aurait été intéressant de faire une analyse similaire des vrais mots cochés dans notre étude aussi, mais il n'y a pas de place pour cela dans ce mémoire.

Chapitre 6. Discussion

Dans cette partie, nous allons explorer la signification des découvertes présentées dans la section précédente. D'abord, nous répondrons à notre question de recherche en comparant nos découvertes avec les résultats des études précédentes. Ensuite, nous allons discuter les implications pédagogiques. Finalement, nous allons traiter des limites de notre recherche et donner des perspectives de recherches futures sur la taille d vocabulaire.

6. 1. Les résultats en lumière d'autres études

Notre première question de recherche était *Quelle est la taille du vocabulaire français chez les étudiants de première année à l'université en Norvège ?* Nous avons été inspirées par l'étude de Lindqvist (2017) à contribuer à combler la lacune de ce type d'études pour le français comme deuxième langue étrangère. Étant donné qu'aucune étude comparable n'a été effectuée auparavant en Norvège, nous n'avions pas d'hypothèse précise quant à la taille du vocabulaire de nos participants, mais nous estimions que le nombre de mots français que possèdent les étudiants en Norvège est comparable avec les résultats de Milton (2008) et David (2008). Nous avons aussi posé l'hypothèse que les étudiants norvégiens ont un vocabulaire plus grand que les élèves les plus âgés dans l'étude de Lindqvist (2017).

Comment se placent alors nos résultats par rapport aux chiffres des études antérieures ? Nous résultats ont montré que les étudiants de français en Norvège connaissent en moyenne 2451 mots au milieu du premier semestre. Le tableau 6-1 présente un aperçu sur les différents résultats des études mentionnées précédemment ainsi que les résultats de notre étude avec et sans la version 3 du test. Les participants le plus âgés (dernière année au collège) dans l'étude en Suède avaient une connaissance moyenne de 1150 mots. Quant aux recherches sur la taille du vocabulaire chez des étudiants au premier semestre au Royaume-Uni, l'étude de Milton (2008) a donné une moyenne de 1950 mots et celle de David (2008) a indiqué en moyenne 2524 mots. Les deux études effectuées par Pignot-Shahov (2014, 2018) ont indiqué une connaissance moyenne de 3233 mots et 2571 mots respectivement. Les études de Milton (2008) et Pignot-Shahov (2014) ont testé la taille du vocabulaire chez ces participants deux fois, une fois pendant le premier semestre et une fois pendant le deuxième semestre. Les résultats du deuxième semestre sont également inclus dans le tableau 6-1. Dans les études précédentes il y avait une grande écart entre les scores des participants et nous résultats démontrent également une hétérogénéité dans le groupe de participants ; les scores vont d'une connaissance de 1323 mots à une connaissance de 3600 mots.

Les résultats de notre étude concorde bien avec nos hypothèses initiales ; les étudiants en premier semestre des études françaises en Norvège connaissent plus de mots que les élèves les plus âgés dans l'étude de Lindqvist (2017) et environ le même nombre de mots que leurs pairs au Royaume-Uni. Il est important de souligner que toutes les études mentionnées ici sont assez petites et les chiffres sont par conséquent incertains. Cependant, le fait que notre enquête ait donné des chiffres semblables aux études précédentes, ajoute à la fiabilité de nos résultats.

	n	minimum	maximum	moyen	écart-type
Milton oct.	29	-	-	1950	678
Milton mai	29	-	-	2555	678
Pignot-Shahov pilot	6	2400	4750	3233	817
Pignot-Shahov nov.	12	1750	3700	2571	646,83
Pignot-Shahov mai	9	1900	4250	3044	763,39
David	120	750	4100	2524	589
Kjellmark	16	1323	3600	2451	659,8
Kjellmark, sans version 3	10	1550	3600	2518	686,8

Tableau 6-1 Aperçu sur la taille moyen du vocabulaire démontré dans notre étude et les études antérieures

En ce qui concerne la distribution selon les bandes de fréquence, le groupe de participants dans son ensemble démontre un profil de connaissance lexical typique, c'est-à-dire que le pourcentage de mots connus dans chaque bande diminue à mesure que la fréquence diminue. Ceci concorde avec l'hypothèse établie selon laquelle le nombre de mots connus diminue lorsqu'ils deviennent moins fréquents. Toutefois, il y a un écart notable entre les connaissances des participants. Le participant ayant obtenu le score les plus élevée semble avoir une bonne connaissance dans toutes les bandes de fréquence. D'autre part, le participant ayant obtenu le score le plus bas connaissait le plus de mots dans la bande 1K et 2K et sa connaissances de mots dans les bandes 3K, 4K et 5K était assez faible.

Dans la section 3.3.2, nous avons vu que Lindqvist (2017, 2020) affirme que les élèves suédois ont bénéficié de leur connaissance des mots cognats suédois/anglais/français et anglais/français dans une certaine mesure. Nous pouvons supposer que le nombre de mots cognats dans un test de vocabulaire pourrait donner aux participants un score exagéré. Nous pouvons, pour la même raison, supposer que les apprenants norvégiens peuvent reconnaitre certains mots dans chaque version du test X-Lex sans forcément les avoir rencontrés en français auparavant. Deux exemples sont les mots *fragment* et *observation* qui se trouvent dans la bande de fréquence 5K. L'orthographe de ces mots est très proche en norvégien et en anglais (*fragment/fragment* et *observasjon/observation*), et les significations sont les mêmes

aussi. En outre, les apprenants norvégiens peuvent probablement identifier quelques mots n'ayant pas de cognat en norvégien mais qui ressemblent aux mots qu'ils connaissent déjà en anglais. La plupart de ces mots, comme *source* sont des emprunts au français. Lorsque nous considérons ces facteurs, ils pourraient expliquer les scores relativement élevés dans les bandes de fréquence 4K et 5K (50% et 40% respectivement) par rapport à ce à quoi on aurait pu s'attendre, vu le niveau de maîtrise de langue générale.

D'autre part, comme nous l'avons présenté dans la section 5. 4. , il semble que les participants sont plus enclins à indiquer qu'ils connaissent les mots inventés ayant des équivalents anglais que les faux mots qui n'en ont pas. Nous avons aussi vu, lors de l'auto-évaluation présentée en section 5. 3. , que les étudiants norvégiens estiment que leur connaissance de l'anglais est bonne. Leur confiance en anglais pourrait ainsi contribuer à cette tendance. Cocher des mots inventés ressemblant aux vrais mots anglais conduirait à une réduction du score. Il est possible que cette tendance dans une certaine mesure contrebalance l'avantage présumé obtenu à partir des mots cognats anglais/français. Notons également la possibilité que les participants sachent que la racine d'un mot existe en français, mais pas si l'inflexion ou la dérivation particulière dans répertorié le test est correcte. Un exemple est le mot inventé **provocatif,* qui est un faux cognat au mot anglais *provocative,* mais qui est aussi très proche des formes *provocant, provocateur* et *provocation* en français. En plus, les instructions demandent aux participants d'indiquer les mots qu'ils reconnaissent ou comprennent. Nous pourrions alors affirmer que l'effet de mots cognats ne causent pas de surestimations de la connaissance lexicale précisément parce que les participants sont capables de reconnaitre et comprendre ces mots, peu importe qu'ils puissent le faire grâce à ces connaissance des autres langues.

L'effet des cognats parmi les vrais mots, mais aussi des faux cognats parmi les mots inventés devrait être exploré davantage. Il convient alors de poser la question si l'effet de mots cognats n'affecterait pas également les scores des participants britanniques, étant donné qu'il existe un bon nombre des mots cognats anglais/français. En outre, la recherche psycholinguistique (cf. la section 2.3.1) indique que de telles similarités entre des langues facilitent l'apprentissage de nouveaux mots. Comme nous l'avons vu, Allen (2019) et Szabo (2020) proposent assurer que la proportion de mots cognats dans la langue maternelle et la langue cible soient la même dans un test de vocabulaire afin d'améliorer la précision de ces tests et éviter une surestimation de la connaissance du vocabulaire.

6.1.1. Implications pédagogiques/pratique

Les programmes pour les langues étrangères suggèrent que les élèves devraient atteindre un niveau correspondant au niveau A2 au lycée (Utdanningdirektoratet, 2020a) et le niveau de connaissances préalable recommandées d'admission pour les études de français à l'université est le niveau A2 en français aussi (voir par exemple www.ntnu.no/studier/aafran/om). Le score moyen des participants à notre étude, 2451 mots, concorde avec la taille du vocabulaire associée au niveau du CECRL A2 (Meara et Milton, 2003; Milton, 2010). Comme nous l'avons vu en la section 3. 1. les apprenants de français L2 en Espagne et en Grèce ayant obtenu des scores de vocabulaire similaires à les étudiants norvégiens ont été placés aux niveaux B1 et B2 (Milton, 2010). Nous pourrions alors également affirmer que les étudiants norvégiens se situent à ces niveaux. Il semble donc qu'il y ait une bonne concordance entre le niveau réel des étudiants et les exigences formelles.

Nous avons vu en section 3.1.1. que Nation (2013) propose que le niveau de la compréhension minimale acceptable soit une couverture d'environ 95%. Plusieurs recherches concordantes ont démontré qu'au moins 98% de couverture est nécessaire pour arriver à une lecture plaisante et fluide (Nation, 2013; Schmitt, Jiang et Grabe, 2011). Il faut souligner que ces estimations ont été établies sur la base de l'anglais ; il existe très peu d'études sur ce sujet pour la langue française. Cependant, Ramnäs (2019) a trouvé qu'une connaissance d'entre 3000 et 4000 lemmes est nécessaire pour arriver à une couverture de 95% pour trois romans jugés suffisamment faciles pour la première année de français à l'université en Suède. Ceci implique que seulement trois des participants de notre étude possèdent les connaissances nécessaires pour une lecture acceptable de tels livres. L'un de ces romans, *L'étranger* d'Albert Camus faisait auparavant partie de la lecture obligatoire en première année à l'université où nos recherches ont été menées.

Dans la section précédente nous avons remarqué qu'il y a un écart notable entre le participant ayant obtenu le score le plus bas, 1323 mots (1550 sans la version 3) et le score le plus haut de 3600 mots. Les étudiants possédant le moins de mots auront probablement des difficultés à suivre les cours au niveau universitaire, ceci malgré le fait qu'ils atteignent le niveau de connaissances préalables recommandées. Nous ne savons pas dans quelle mesure les textes du syllabus actuel correspondent au niveau réel des étudiants, mais cela pourrait poser un problème si les étudiants sont censés lire et étudier des textes qui dépassent leur niveau de compétence. Le test X-Lex ne mesure pas la connaissance de mots au-delà de la bande de fréquence 5K. Les participants dans l'étude de Pignot-Shahov (2018) connaissaient en

moyenne environ 765 mots au total dans les bandes de fréquence 6K et 7K. Nous pouvons supposer que les étudiants norvégiens auront également une certaine connaissance de mots au-delà de la bande de fréquence, mais cela doit être exploré par des recherches supplémentaires. La connaissance au-delà du niveau 5K contribuera probablement à augmenter le niveau de couverture mais pas de manière considérable.

En Suède il a été élaboré une liste qui consiste en 4000 mots que les étudiants sont censés apprendre au cours de premier semestre d'études de français. Si les étudiants en Suède ont environ le même niveau de vocabulaire que les étudiants norvégiens dans notre étude, ils doivent apprendre environ 1600 mots en plus avant la fin du premier semestre et ensuite 2000 mots supplémentaires au cours du deuxième semestre. Les résultats de Milton (2008) et Pignot-Shahov (2018) montrent que l'augmentation de taille du vocabulaire entre le premier et le deuxième semestre d'études supérieures de français est 605 mots et 473 mots respectivement. Nous pouvons par conséquent nous demander s'il n'est pas irréaliste de demander aux étudiants d'élargir leur vocabulaire autant au cours d'un semestre. Cependant, l'étude de Ramnäs (2019) affirme qu'une connaissance de beaucoup plus de mots est nécessaire pour passer d'une couverture de 95% à une couverture de 98%. Une telle liste est un bon soutien à l'apprentissage des nouveaux mots pour les étudiants. De plus, à l'université de Göteborg ils ont réalisé un « programme » ciblant en particulier l'enseignement du vocabulaire en l'intégrant dans les cours de grammaire. Une liste de vocabulaire conseillé pourrait être très utile pour les étudiants norvégiens aussi. Nous proposerons également de mettre davantage l'accent sur l'enseignement du vocabulaire.

6. 2. Les limites de notre recherche

Nous avons déjà traité en détail les divers défis et limitations avec le test X-Lex (Meara et Milton, 2003), plus en détail en section 3.2.5 et le but de cette étude n'est pas d'évaluer la qualité du test X-Lex. Dans cette section, nous examinerons de plus près les défis spécifiquement liés à notre étude.

L'un des points faibles de notre étude est les fautes de frappe qui n'ont malheureusement pas été découvertes dans la version 3 (voir 4.2.3). Ce type de faute est critique dans un test qui prend comme point de départ la différence entre des mots réels et des mots fictifs, où justement l'orthographe sert à distinguer entre les deux. Cela dit, nous sommes convaincues

qu’avec les ajustements discutés en chapitre 4, nos résultats ne seront pas trop affectés par ces fautes, ce que montrent aussi les calculs dans le tableau 4-2 et la figure 4-3.

Tous les participants aux études mentionnés ici apprennent le français en tant que langue étrangère dans un contexte formel où ils reçoivent peu d’input de la langue cible dans la vie quotidienne. Cependant, il faut garder à l'esprit que les différences en langue maternelle des participants, des écarts en ce qui concerne l’objectif de recherche, la procédure et la méthodologie des différentes études pourraient avoir un impact sur les résultats des études. Cela rend la comparaison des résultats plus complexe. Ceci est la raison pour laquelle nous avons suivi au près la méthodologie utilisée par Lindqvist (2017) (voir sections 3. 3. et 4. 2.). Pour les participants dans l’étude de Lindqvist (2017) et dans la nôtre, le français est normalement leur L3. Le suédois et le norvégien sont également des langues très proches et elles partageront beaucoup des mêmes cognats du français et de l'anglais. Les systèmes scolaires en Suède et en Norvège sont assez similaires, et il n'est donc pas anormal de comparer avec les résultats suédois. Néanmoins, les variables sont toujours nombreuses et l’estimation de la taille du vocabulaire n’est pas une science exacte, mais nous espérons que cela ajoute à la force de notre étude qu’elle ait été effectuée avec des méthodes aussi similaires que possible.

Même si les instructions et le format du test sont assez simples, ils n’ont pas été présentés aux participants de manière identique dans toutes les études utilisant le test X-Lex. Dans la version originale numérique, (Meara et Milton, 2003) les participants doivent choisir entre deux émoticônes (😊 ou ☹) pour indiquer s’ils comprennent le mot en question. Dans d’autres études, le format était le même que celui présenté pour le EVST (cf. la section 3.2.3) où les participants devaient cocher les mots qu'ils connaissaient. Dans l’étude de Lindqvist (2017), sur laquelle nous reviendrons dans la section 3.3.1, le test avait le format oui/non, demandant aux participants de cocher « oui » s’ils reconnaissent le mot, et « non » s’ils ne le reconnaissent pas. Il y a par conséquent des nuances différentes dans la question principale du test. Cela peut amener les participants à donner des réponses différentes, en fonction de leur interprétation de la question. Si un participant est en doute, il peut cocher *oui* mais un autre peut opter pour *non* même s’il a une idée de la signification du mot. Plusieurs études ont démontré que les apprenants dans différents pays affichent des tendances différentes quant à l'évaluation de leurs propres connaissances. Il est difficile de dire si ces nuances pourraient ou non avoir un impact sur les résultats du test, mais pour éliminer cette possibilité, il serait peut-

être avantageux que les études futures s'en tiennent à un seul format aussi proche que possible afin de faciliter la comparaison de trouvailles.

L’obstacle le plus important pour une telle étude est peut-être le volume de participants. Notre échantillon de 16 participants est un nombre assez restreint et il peut être difficile d’obtenir des résultats qui sont scientifiquement valides. Il aurait été préférable d’avoir davantage de participants et ainsi de couvrir un échantillon plus représentatif des étudiants de français à l’université en Norvège mais nous espérons avoir mis en lumière certains points qui pourraient mener à une étude plus quantitative à l’avenir. En plus, il serait intéressant de comparer les résultats des études ayant utilisé le X-Lex avec les résultats d'un autre type de test visant à estimer la taille du vocabulaire, tel que le TTV qui mesure aussi la taille du vocabulaire au-delà de la bande de fréquence 5K.

Chapitre 7. Conclusion

La présente étude a montré que les étudiants en premier semestre d'études de français en Norvège connaissent en moyenne 2451 mots. Ce résultat correspond bien avec les trouvailles des études précédentes effectuées en Royaume-Uni (David, 2008; Milton, 2008; Pignot-Shahov, 2014, 2018). Il y a un écart notable parmi les participants quant à leurs scores individuels, mais en général les étudiants connaissent plus de mots dans les bandes des mots les plus fréquentes. Nos résultats nous conduisent à poser la question de savoir si les étudiants norvégiens ont un vocabulaire suffisamment large pour lire et comprendre les textes qui figurent dans le programme et de plus, s'il est nécessaire d'intégrer l'enseignement du vocabulaire dans les cours aux niveaux supérieurs aussi. En outre, de futures études pourraient porter sur les élèves au collège et au lycée. Ainsi, il serait possible de comparer les résultats de ces groupes en Norvège avec les résultats des groupes correspondants obtenus dans des autres pays.

Quant au rôle des mots cognats, plusieurs recherches récentes (voir Allen, 2019; Lindqvist, 2020; Szabo, 2020) ont traité l'effet de mots cognats dans les test de la taille du vocabulaire. Ces études se sont concentrées sur l'effet de cognats parmi les mots réels indiqués comme connus par les participants aux tests. Nous avons observé que les participants semblent indiquer plus souvent qu'ils connaissent les faux mots ressemblant aux mots anglais que les faux mots n'ayant pas de « cognat » en anglais. Nous ne savons pas dans quelle mesure cette tendance peut influencer le score des participants. Plus de recherches sur cet aspect du tests en vocabulaire pourrait nous éclairer à propos de cette question.

Nous avons constaté à plusieurs reprises qu'il y a beaucoup de recherches à faire qui peuvent améliorer la précision des tests de vocabulaire, par exemple dans les corpus sur lesquels ils sont basés, la proportion de mots cognats et la façon dont les faux mots sont construits. Cette étude est évidement trop restreinte pour permettre de tirer des conclusions générales par rapport à la taille du vocabulaire chez tous les étudiants de français en Norvège, mais nous pouvons toutefois tirer la même conclusion que Ramnäs (2019) ; le vocabulaire est un domaine où le lien entre recherche et enseignement mérite d'être développé. Nous espérons que cette étude pourrait servir de point de départ à des études futures plus approfondies, ce qui à son tour pourrait contribuer au développement de pratiques d'enseignement des langues étrangères en Norvège.

Bibliographie

Aitchison, J. (2012). *Words in the mind : an introduction to the mental lexicon*. (4. utgave.e éd.). Chichester, West Sussex: Wiley-Blackwell.

Allen, D. (2019). Cognate Frequency Predicts Accuracy in Tests of Lexical Knowledge. *Language assessment quarterly, 16*(3), 312-327. doi: 10.1080/15434303.2019.1635134

Ameel, E., Storms, G., Malt, B. C. et Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of memory and language, 52*(3), 309-329. doi: 10.1016/j.jml.2005.01.003

Anderson, R. C. et Freebody, P. (1981). Vocabulary knowledge. Dans J. T. Guthrie (dir.), *Comprehension and teaching: Research reviews* (p. 77-117). Newark, Delaware: International Reading Association.

Bardel, C., Gudmundson, A. et Lindqvist, C. (2012). Aspects of Lexical Sophistication in Advanced Learners' Oral Production. Vocabulary Acquisition and Use in L2 French and Italian. *Studies in Second Language Acquisition, 34*(2), 269-290. doi: 10.1017/S0272263112000058

Batista, R. et Horst, M. (2016). A New Receptive Vocabulary Size Test for French. *Canadian modern language review, 72*(2), 211-233. doi: 10.3138/cmlr.2820

Baudot, J. (1992). *Fréquences d'utilisation des mots en français écrit contemporain*. Presses de l'Université de Montréal.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M. et Velde, H. V. d. (2001). Examining the Yes/No vocabulary test: Some methodological issues in theory and practice. *Language testing, 18*(3), 235-274. doi: 10.1177/026553220101800301

Clahsen, H., Eisenbeiss, S., Hadler, M. et Sonnenstuhl, I. (2001). The Mental Representation of Inflected Words: An Experimental Study of Adjectives and Verbs in German. *Language, 77*(3), 510-543. doi: 10.1353/lan.2001.0140

Cobb, T. et Horst, M. (2004). Is there room for an academic word list in French? Dans P. Bogaards & B. Laufer-Dvorkin (dir.), *Vocabulary in a second language : selection, acquisition, and testing* (Vol. vol. 10). Amsterdam: John Benjamins.

Conseil de l'Europe. (2018). Cadre européen commun de référence pour les langues - Apprendre, Enseigner, Évaluer (CECRL). Repéré le 21.01.21à https://rm.coe.int/cecr-volume-complementaire-avec-de-nouveaux-descripteurs/16807875d5

Daller, H., Milton, J. et Treffers-Daller, J. (2007). Editor's introduction: conventions, terminology and an overview of the book. Dans H. Daller, J. Milton & J. Treffers-Daller (dir.), *Modelling and assessing vocabulary knowledge* (p. 1-32). Cambridge: Cambridge University Press.

David, A. (2008). Vocabulary breadth in French L2 learners. *Language learning journal, 36*(2), 167-180. doi: 10.1080/09571730802389991

de Bot, K. (2004). The Multilingual Lexicon: Modelling Selection and Control. *International Journal of Multilingualism, 1*(1), 17-32. doi: 10.1080/14790710408668176

Harsch, C. et Hartig, J. (2016). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language testing, 33*(4), 1-21. doi: 10.1177/0265532215594642

Krautz, A. E. (2020). The Multilingual Lexicon: Evidence from Primed Translation Lexical Decision. Dans P. Booth & J. Clenton (dir.), *First Language Influences on Multilingual Lexicons* (1e éd., p. 69-86). doi: 10.4324/9780429031410-7.

Kroll, J. F., Gullifer, J. W. et Rossi, E. (2013). The Multilingual Lexicon: The Cognitive and Neural Basis of Lexical Comprehension and Production in Two or More Languages. *Annual Review of Applied Linguistics, 33*, 102-127. doi: 10.1017/S0267190513000111

Laufer, B. et Ravenhorst-Kalovski, G. C. (2010). Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a foreign language, 22*(1), 15.
Lenoble, M. (1992). Répertoire des fréquences du français. *Revue Informatique Et Statistique Dans Les Sciences Humaines, 28*, 321.
Lindqvist, C. (2016). Tredjespråkets ordförråd. Dans C. Bardel, Y. Falk & C. Lindqvist (dir.), *Tredjespråksinlärning* (p. 59-75). Lund: Studentlitteratur.
Lindqvist, C. (2017). Le développement de la taille du vocabulaire en français L2 chez les élèves suédophones. *Synergies Pays Scandinaves*(11-12), 151-161.
Lindqvist, C. (2020). First and Second Language Cognate Effects in Third Language Vocabulary Size Estimates. Dans P. Booth & J. Clenton (dir.), *First Language Influences on Multilingual Lexicons* (1e éd.). New York: Routledge.
Lindqvist, C., Gudmundson, A. et Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. *Eurosla Monograph, Serie 2*, 109-126.
Lindqvist, C. et Ramnäs, M. (2016). L'enseignement du vocabulaire à l'université. *Synergies Pays Scandinaves*(11/12), 55-64.
Listhaug, K. F. (2015). *Spatial prepositions and second language acquisition : the acquisition of spatial prepositions in French by native speakers of Norwegian.* (PhD, Norwegian University of Science and Technology, Trondheim).
Lonsdale, D. et Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners*. New York: Routledge.
Meara, P. (1993). The Bilingual Lexicon and the Teaching of Vocabulary. Dans R. Schreuder & B. Weltens (dir.), *The Bilingual lexicon*. Amsterdam: John Benjamins Publishing Company.
Meara, P. M. et Milton, J. (2003). *X-lex: the Swansea levels test*. Express Publishing.
Milton, J. (2008). French vocabulary breadth among learners in the British school and university system: comparing knowledge over time. *Journal of French Language Studies, 18*(3), 333-348. doi: 10.1017/S0959269508003487
Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual matters.
Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe. *Eurosla Monograps Series, 1*, 211-231.
Milton, J. et Alexiou, T. (2020). Vocabulary Size Assessment: Assessing the Vocabulary Needs of Learners in Relation to Their CEFR Goals. Dans M. Dodigovic & M. P. Agustín-Llach (dir.), *Vocabulary in Curriculum Planning: Needs, Strategies and Tools* (p. 9-27). Cham: Palgrave Macmillan.
Milton, J., Wade, J. et Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. Dans D. R. Chacón-Beltrán, C. Abello-Contesse & M. d. M. Torreblanca-López (dir.), *Insights Into Non-native Vocabulary Teaching and Learning* (p. 83-98).
Modiano, P. (2001). *La Petite Bijou*. Paris: Gallimard.
Mot. (2016). Paris.
Nation, I. (1983). Testing and teaching vocabulary. Guidelines, 5 (1), 12-25 (p. 12-25).
Nation, I. S. P. (2001). *Learning vocabulary in another language*
Nation, I. S. P. (2006). How Large a Vocabulary is Needed For Reading and Listening? *The Canadian Modern Language Review, 63*(1), 59-82. doi: 10.3138/cmlr.63.1.59
Nation, I. S. P. (2013). *Learning vocabulary in another language*. (2nd ed.e éd.). Cambridge: Cambridge University Press.

Petitpas, T. (2010). Enseigner la variation lexicale en classe de FLE. *The French Review, 83*(4), 800-818.
Pignot-Shahov, V. (2014). Le développement lexical des apprenants de français langue étrangère d'une université britannique: une étude pilote. *Synergies Royaume-Uni et Irlande*(7), 123-131.
Pignot-Shahov, V. (2018). *L2 French lexical development of undergraduate students in a UK university.* (PhD, University of Southampton, Southampton). Repéré à https://eprints.soton.ac.uk/432082/
Ramnäs, M. (2019). Étendue du vocabulaire et compréhension écrite – le français à l'université en Suède. *Bergen language and linguistics studies, 10*(1), 12. doi: 10.15845/bells.v10i1.1430
Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? Dans P. Bogaards & B. Laufer-Dvorkin (dir.), *Vocabulary in a second language: Selection, acquisition, and testing*. Amsterdam: Benjamins.
Riegel, M., Pellat, J.-C. et Rioul, R. (2018). *Grammaire méthodique du français*. (7e édition.e éd.). Paris: Presses Universitaires de France.
Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.
Schmitt, N., Jiang, X. et Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *The Modern language journal (Boulder, Colo.), 95*(1), 26-43. doi: 10.1111/j.1540-4781.2011.01146.x
Schmitt, N., Schmitt, D. et Clapham, C. (2016). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language testing, 18*(1), 55-88. doi: 10.1177/026553220101800103
Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language learning journal, 36*(2), 139-152. doi: 10.1080/09571730802389975
Szabo, C. Z. (2020). The Reassesment of Vocabulary Tests Based on Cognate Distribution. Dans P. Booth & J. Clenton (dir.), *First language influences on multilingual lexicons* (1st Edition.e éd.).
Thornbury, S. (2002). *How to teach vocabulary*. Harlow: Longman.
Utdanningdirektoratet. (2020a, 26.06.2020). Kjennetegn på måloppnåelse - fremmedspråk nivå II. Repéré à https://www.udir.no/laring-og-trivsel/lareplanverket/kjennetegn/kjennetegn-pa-maloppnaelse-fremmedsprak-niva-ii/
Utdanningdirektoratet. (2020b, 26.06.2020). Kjennetegn på måloppnåelse – fremmedspråk nivå I. Repéré à https://www.udir.no/laring-og-trivsel/lareplanverket/kjennetegn/kjennetegn-pa-maloppnaelse--fremmedsprak-niva-i/
Vinet, A. (2011). *L'enseignement du vocabulaire en cours de français.* (L'Université de Stockholm, Stockholm).
Wlosowicz, T. M. (2010). Le transfert et les interférences entre L1, L2 et L3 dans la production des cognates aux terminaisons différentes. *Synergies Espagne*(3), 159-170.

Annexes

I Formulaire d'information et de consentement

Vil du delta i forskningsprosjektet Vokabularstørrelse i fransk som tredjespråk?
Dette er et spørsmål til deg om å delta i et forskningsprosjekt hvor formålet er å undersøke vokabularstørrelsen i fransk som tredjespråk. I det følgende gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg.

Formål
Dette er en masteroppgave der målet er å undersøke vokabularstørrelsen (hvor mange ord man kan på et språk) i fransk hos studenter som har startet på førsteåret med franskstudier på norske universiteter.
For å kunne forstå og å uttrykke seg godt muntlig og skriftlig på et språk, trenger vi et ganske stort ordforråd. Vi vil derfor undersøke hvor stort ordforråd studenter som begynner på franskstudier i Norge har. De aller fleste av disse studentene har lært fransk som fremmedspråk på ungdomsskolen og videregående. Resultatene fra testene vil bli analysert og sammenlignet med resultater fra liknende studier i andre land. Resultatene kan også bli brukt senere i forbindelse med en større studie som tester ordforrådet på fransk hos elever på ungdomsskolen og videregående.

Hvorfor får du spørsmål om å delta?
Du har blitt spurt om å delta fordi du er førsteårsstudent på fransk ved et norsk universitet.

Hva innebærer det for deg å delta?
Hvis du velger å delta, innebærer det at du først gjennomfører en vokabulartest. Du får se lister med franske ord, og blir bedt om å krysse av for om du kjenner igjen eller kan bruke disse ordene på fransk. Blant disse er det også blandet inn «falske» ord som er konstruert for å ligne på franske ord, men som ikke har noen betydning. Disse falske ordene fungerer derfor som en slags kontroll.
Deretter blir du best om å fylle ut et elektronisk spørreskjema med informasjon om deg og din språkbakgrunn. Spørreskjemaet inneholder spørsmål om din språkbakgrunn, hvor lenge du har lært fransk og andre faktorer som kan tenkes å ha innvirkning på språklæring.
Vi anslår at det vil ta deg omtrent 15-20 minutter totalt å fullføre undersøkelsen.
Det er frivillig å delta
Det er frivillig å delta i prosjektet. Hvis du velger å delta, kan du når som helst trekke samtykket tilbake uten å oppgi noen grunn. Alle dine personopplysninger vil da bli slettet. Kun fullstendig anonymiserte data vil bli beholdt. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg.

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger

NTNU har ansvar for prosjektet. Vi vil bare bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket.
Datamaterialet vil bli lagret på NTNUs servere. Det er kun prosjektansvarlig (veileder) og en masterstudent som vil ha tilgang til opplysningene som samles inn.
Du vil ikke kunne gjenkjennes i publikasjon av masteroppgaven.
Spørreskjemaet som brukes, er laget gjennom Nettskjema, som er en sikker løsning for datainnsamling via nett.

Hva skjer med opplysningene dine når vi avslutter forskningsprosjektet?

Opplysningene anonymiseres når prosjektet avsluttes/oppgaven er godkjent, noe som etter planen er 01.07.2021.

Dine rettigheter

Så lenge du kan identifiseres i datamaterialet, har du rett til:

- innsyn i hvilke personopplysninger som er registrert om deg, og å få utlevert en kopi av opplysningene,
- å få rettet personopplysninger om deg,
- å få slettet personopplysninger om deg, og
- å sende klage til Datatilsynet om behandlingen av dine personopplysninger.

Hva gir oss rett til å behandle personopplysninger om deg?

Vi behandler opplysninger om deg basert på ditt samtykke.
På oppdrag fra NTNU har NSD – Norsk senter for forskningsdata AS vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

Hvor kan jeg finne ut mer?

Hvis du har spørsmål til studien, eller ønsker å benytte deg av dine rettigheter, ta kontakt med prosjektansvarlig Kjersti Faldet Listhaug (kjersti.listhaug@ntnu.no) eller personvernombud ved NTNU,
Thomas Helgesen (thomas.helgesen@ntnu.no).
Du kan også ta kontakt med NSD – Norsk senter for forskningsdata AS, på epost (personverntjenester@nsd.no) eller telefon: 55 58 21 17.

Studien gjennomføres anonymt.
Dersom du skulle ønske å trekke deg underveis i undersøkelsen, kan du gå ut av skjemaet uten å fullføre det.
Hvis du trekker deg underveis, i studien vil dataene dine vil ikke bli registrert.
Når du fullfører undersøkelsen, vil dataene dine anonymiseres og du vil ikke kunne identifiseres.

Samtykke-erklæring

Jeg har mottatt og forstått informasjon om prosjektet *Vokabularstørrelse i fransk som tredjespråk*, og har fått anledning til å stille spørsmål. Jeg samtykker til:

☐ å delta i forskningsprosjektet.

Jeg har mottatt og forstått informasjon om prosjektet *Vokabularstørrelse i fransk som tredjespråk*, og har fått anledning til å stille spørsmål. Jeg samtykker til:

☐ at mine opplysninger kan behandles fram til prosjektet er avsluttet.

II Questionnaire biographique

Bakgrunnsinformasjon for forskningsprosjekt om ordforråd i fransk

For at resultatene fra prosjektet skal bli så nøyaktige som mulig, ber vi om at du nå fyller ut dette skjemaet, hvor vi spør om litt nødvendig informasjon om deg og din språkbakgrunn.

- Hvor gammel er du?

☐ 18-21 år
☐ 22-25 år
☐ 26-30 år
☐ Over 30 år

- Hva er morsmålet ditt? Lærte du to eller flere språk fra fødselen av, oppgir du begge/alle språkene.

- Har du familie/nære venner som du snakker et annet språk enn norsk med?

☐ Ja
☐ Nei

- Du svarte at du snakker et annet språk enn norsk med familie/nære venner. Vennligst oppgi hvem du snakker andre språk enn norsk med

Du må velge minst ett svaralternativ.

☐ Nær familie
☐ Storfamilie
☐ Nære venner
☐ Perifere venner / bekjente

- Du svarte at du snakker et annet språk enn norsk med familie/nære venner

Vennligst oppgi hvilke(t) språk dere bruker

- Hvor gammel var du da du begynte å lære fransk?

- Når begynte du med fransk på skolen? Oppgi hvilket klassetrinn. Hvis du ikke har hatt franskundervisning på skolen, eller noen annen form for formell opplæring i fransk, kan du bare skrive det.

- Hvor gammel var du da du begynte å lære engelsk?

- Hvilke språk kan du og hvordan vil du vurdere ferdighetene dine i disse språkene? Sett kryss i tabellen. Hvis du kan flere språk enn engelsk og fransk, fyll inn disse selv i neste spørsmål.

	Grunnleggende: Jeg kan kjenne igjen og bruke enkle ord og uttrykk.	Middels: Jeg kan bruke vanlige ord og enkle setninger, og kan kommunisere på grunnleggende nivå.	Selvstendig: Jeg kan forstå hovedinnholdet i standard tale og snakke/skrive sammenhengende om kjente tema.	Kompetent: Jeg kan snakke ganske flytende og forstå sammenhengende tale om vanskelige, kjente tema. Jeg kan skrive detaljert om mange tema.	Flytende: Jeg kan uttrykke meg flytende, spontant og presist. Jeg forstår uten problemer alt jeg leser og hører, og kan bruke språket effektivt i sosiale og faglige sammenhenger.
Engelsk					
Fransk					

- Hvis du kan andre språk enn fransk og engelsk, vennligst fyll inn i boksen under hvilke(t) og hvordan du selv vil vurdere ferdighetene dine i disse språkene.

Grunnleggende: Jeg kan kjenne igjen og bruke enkle ord og uttrykk

Middels: Jeg kan bruke vanlige ord og enkle setninger, og kan kommunisere på grunnleggende nivå.

Selvstendig: Jeg kan forstå hovedinnholdet i standard tale og snakke/skrive sammenhengende om kjente tema.

Kompetent: Jeg kan snakke ganske flytende og forstå sammenhengende tale om vanskelige, kjente tema. Jeg kan skrive detaljert om mange tema.

Flytende: Jeg kan uttrykke meg flytende, spontant og presist. Jeg forstår uten problemer alt jeg leser og hører, og kan bruke språket effektivt i sosiale og faglige sammenhenger.

Svareksempel: samisk- kompetent

- I løpet av det siste året, hvor ofte har du hørt fransk utenom skolen/universitetet?
  For eksempel gjennom film, YouTube, TV, dataspill, bøker, musikk eller lignende.

Du må velge minst ett svaralternativ.

☐ Aldri
☐ Av og til, men mindre enn 1 time i uka i gjennomsnitt
☐ 1-3 timer i uka i gjennomsnitt
☐ Mer enn 3 timer i uka i gjennomsnitt

- I løpet av det siste året, hvor ofte har du hørt engelsk utenom skolen/universitetet?
  For eksempel gjennom film, YouTube, TV, dataspill, bøker, musikk eller lignende

Du må velge minst ett svaralternativ.

☐ Aldri
☐ Av og til, men mindre enn 1 time i uka i gjennomsnitt
☐ 1-3 timer i uka i gjennomsnitt
☐ Mer enn 3 timer i uka i gjennomsnitt

- Har du bodd i utlandet i mer enn 3 måneder sammenhengende?
  For eksempel på utveksling

☐ Ja
☐ Nei

- Du oppga at du har bodd i utlandet i mer enn 3 måneder. Vennligst oppgi hvor du bodde, hvor lenge du bodde der, hvor gammel du var og hvilke(t) språk du brukte der.
  Eksempel: Tyskland, 4 måneder, 19 år, tysk og engelsk.

- Har du dysleksi? Vi spør kun om dette fordi det kan tenkes å påvirke resultatet av vokabulartesten.

☐ Ja

☐ Nei

Tusen takk for din deltakelse!

Husk å trykk på "Send inn" for å fullføre og sende inn skjemaet.

Med vennlig hilsen

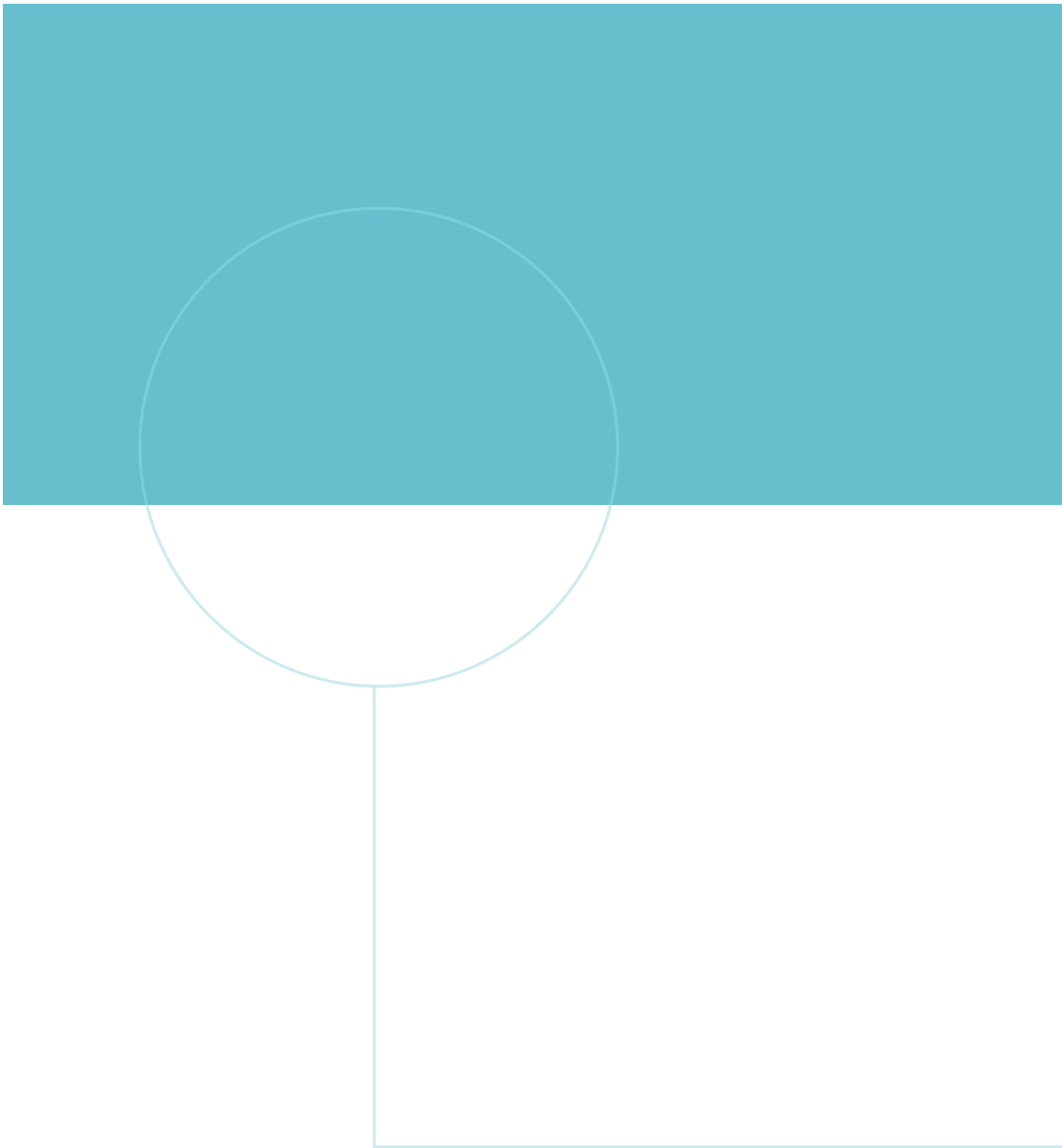
Kjersti Faldet Listhaug (Veileder)

Maren Langen Kjellmark (masterstudent)

III Relevans for lektorutdanningen

I denne masteroppgaven har jeg sett på hvor stort vokabular norske studenter på første semester av franskstudier har. Stadig mer forskning viser at det å kunne mange ord er en helt sentral del av det å tilegne seg et nytt språk, men det holder ikke å kun kjenne til de mest frekvente ordene på et språk. For å kommunisere på en effektiv måte, må man også kunne en god del mindre frekvente ord. Det å ikke finne ordene man trenger for å formilde det man vil si, kan være en frustrerende del av det å lære seg et nytt språk. Imidlertid er vokabularlæring ofte ikke sett på som en like viktig del av undervisningen som for eksempel grammatikk. Vokabular og strategier for å tilegne seg nye ord bør derfor bli en mer eksplisitt del av fremmedspråksopplæringen.

Arbeidet med denne oppgaven har lært meg flere ting jeg kan ta med meg inn i yrket som lektor. Jeg blitt fått et godt overblikk over forskning fra ulike felt som omhandler fremmedspråkopplæring. Dette vil gjøre det enklere å holde seg oppdater på nye funn på dette området senere. I tillegg har jeg blitt kjent med flere vokabulartester som kan brukes til å kartlegge elevenes vokabularstørrelse i fransk, samt det å bearbeide dataene en slik undersøkelse gir. Slike tester kan være fordelaktig for planlegging av undervisningen på flere måter, blant annet ved å avdekke eventuelle kunnskapshull eller for å gi en pekepinn på elevenes generelle språknivå uten å bruke tid på mer omfattende språktester. Å gjennomføre en vokabulartest i praksis, har tillegg gitt meg et godt grunnlag for å bedrive forskning på og utvikling av egen undervisningspraksis senere. Jeg håper dermed at det jeg har lært gjennom denne prosessen kan bidra til at elever får oppleve mestringsfølelse i franskundervisningen slik at de med egne øyne kan se verdien av å kunne et annet fremmedspråk i tillegg til engelsk.



NTNU

Kunnskap for en bedre verden