La richesse lexicale dans la production orale de l'apprenant avancé de français

Christina Lindqvist

The Canadian Modern Language Review / La revue canadienne des langues vivantes, Volume 66, Number 3, March / mars 2010, pp. 393-420 (Article)

Published by University of Toronto Press

La richesse lexicale dans la production orale de l'apprenant avancé de français

Christina Lindqvist

Résumé : La présente étude vise à comparer la richesse lexicale de la production orale des apprenants locuteurs natifs de suédois en français langue seconde à l'aide de la méthode *Lexical frequency profile* et du programme *Vocabprofile*, élaborés par Laufer et Nation (1995) pour l'anglais écrit. La version française de *Vocabprofile* a été adaptée pour l'écrit, tandis que la présente étude s'adresse aux données orales. L'analyse se concentre principalement sur l'emploi de mots rares, qui sont censés indiquer un vocabulaire évolué. La comparaison de deux groupes d'apprenants locuteurs natifs de suédois de niveau avancé en français et d'un groupe de locuteurs natifs francophones montre que la proportion de mots rares est en corrélation avec le niveau de compétence. De plus, le groupe le plus avancé présente un profil lexical similaire à celui des locuteurs natifs. Toutefois, l'emploi d'une base de données écrites pour l'examen de la langue parlée ne semble pas entièrement satisfaisant en raison des différences observées dans la fréquence de certains mots à l'oral et à l'écrit.

Mots clés : richesse lexicale, apprenant de niveau avancé, français parlé, fréquence

Abstract: The goal of this study is to compare the lexical richness of the oral production of Swedish learners of French as a second language using the Lexical Frequency Profile method and the Vocabprofile program, elaborated by Laufer and Nation (1995) for written English. The French version of Vocabprofile is designed for written language; however, the study used oral data. The analysis focuses mainly on the use of infrequent words, which is supposed to indicate an advanced vocabulary. The comparison of two groups of advanced Swedish learners of French with a group of native speakers of French shows that the proportion of infrequent words increases with proficiency level. Moreover, the most advanced group has a lexical profile similar to that of the native speakers. However, using a database of written language to analyze spoken language does not seem entirely reliable, because of differences in frequency of certain words in oral and written language.

 The Canadian Modern Language Review/La Revue canadienne des langues vivantes, 66, 3 (March/mars), 393–420 doi:10.3138/cmlr.66.3.393

Keywords: lexical richness, advanced learner, spoken French, frequency

La présente étude vise à comparer la richesse lexicale de la production orale de deux groupes d'apprenants de niveau supérieur en français langue seconde. Cette comparaison est effectuée à l'aide de la méthode *Lexical Frequency Profile* (LFP), élaborée à l'origine par Laufer et Nation (1995) pour l'anglais écrit, et la version française du programme *Vocabprofile*. Le second objectif évalue la méthode sélectionnée. On sait qu'il existe plusieurs méthodes pour mesurer et évaluer les connaissances lexicales dans le cadre des recherches sur l'acquisition d'une langue seconde (par ex., Bulté, Housen, Pierrard et Van Daele, 2008). Dans le cadre du LFP, la fréquence est l'aspect crucial de l'acquisition du lexique en langue seconde. Bon nombre de chercheurs ont insisté sur l'importance de la fréquence dans l'acquisition de mots nouveaux (par ex., Cobb et Horst, 2004, ou Vermeer, 2004) : l'apprenant assimilerait d'abord les mots les plus fréquents et par la suite les mots les moins fréquents de la langue cible. On se demande alors quels sont les mots les plus fréquents d'une langue. Et combien de mots sont fréquents. Comment faire la distinction entre mots fréquents d'un côté et mots rares de l'autre? On peut recourir aux listes de fréquence des mots d'une langue. La méthode LFP (Laufer et Nation, 1995) consiste à répartir les mots d'une production donnée en quatre zones de fréquence, établies à partir d'une importante base de données. La répartition, c'est-à-dire le pourcentage de mots dans les zones de fréquence respectives, constitue le profil lexical d'un apprenant. Plus les mots apparaissent fréquemment dans les zones de faible fréquence, plus le profil lexical est considéré comme avancé.

La méthode LFP a été développée pour l'anglais écrit, et les zones de fréquence sont fondées sur une base de données composée de textes écrits. Goodfellow, Lamy et Jones (2002) et Cobb et Horst (2004) ont élaboré la méthode pour le français (Profil de fréquence lexicale, désormais PFL). Comme dans la version anglaise, la base de données est composée de documents écrits. Dans la présente étude, nous nous interrogeons sur l'emploi d'une base de données écrites pour l'étude de la langue orale. Notre objection principale est qu'il est plausible que des mots à l'écrit et à l'oral présentent des différences de fréquence. Un mot donné peut notamment apparaître fréquemment dans la langue parlée, mais peu souvent dans la langue écrite (Nation, 2001, p. 126; 2006, p. 63). En effet, McCarthy (1998, p. 122) soutient que les listes de fréquence de la langue parlée se distinguent fondamentalement de celles de la langue écrite, et surtout lorsque ces listes sont fondées sur des textes journalistiques. Pour ce qui est du

français, Campione, Véronis et Deulofeu (2005) démontrent l'existence de différentes fréquences pour certaines classes de mots à l'écrit et à l'oral. En ce qui a trait au français, Blanche-Benveniste (1997, p. 9) fait aussi remarquer que les mots *nous*, *lorsque* et *car* à l'écrit sont plus fréquemment rendus à l'oral par *on*, *quand* et *parce que*. Ainsi, en classant les mots d'une production orale dans des zones de fréquence basées sur l'écrit, on risque de se tromper. Mais, même s'il existe en général des différences de fréquence à l'oral et à l'écrit, on ne peut affirmer que le profil lexical d'un individu ou d'un groupe d'individus change en fonction de la base de données à laquelle les productions sont comparées. Ovtcharov, Cobb et Halter (2006) ont mis la méthode à l'essai avec des données orales et ils affirment que la méthode est valide malgré la nature différente des données. Dans la présente étude, nous voulons d'une part vérifier si la méthode PFL est adaptée à l'analyse de données orales et nous voulons d'autre part savoir si la méthode PFL permet d'effectuer une distinction nette entre les profils lexicaux des apprenants de français L2 très avancés et moins avancés.

Nous présentons d'abord des moyens de détermination de la richesse lexicale. Nous traçons ensuite les grandes lignes de la méthode PFL de Laufer et Nation (1995) et de sa version française, et nous effectuons un survol des études antérieures qui ont appliqué cette méthode à la langue française. Puis, nous formulons les questions de recherche et les hypothèses, et précisons les données et la méthode. La section Résultats et analyse présente les profils lexicaux de trois groupes d'informants (deux groupes d'apprenants et un groupe de contrôle constitué de locuteurs natifs francophones). Nous poursuivons par des remarques générales sur la répartition dans les différentes zones de fréquence basse. L'analyse se concentre sur les zones de fréquentes basse, celles qui sont censées contenir des mots rares. Selon l'hypothèse de départ, une proportion élevée de ce type de mots indiquerait un vocabulaire riche et avancé. C'est pourquoi nous effectuons une analyse approfondie des mots classés dans la zone Mots hors listes (ci-dessous). Enfin, l'article propose quelques conclusions et pistes à suivre dans les recherches futures.

Mesure de la richesse lexicale

Il existe plusieurs instruments de mesure de la richesse lexicale, et tous présentent évidemment des avantages et des inconvénients. Le rapport de type-occurrence (*type/token ratio*), par exemple, qui a été abondamment utilisé, a souvent été critiqué parce qu'il dépend trop de la longueur du texte (Daller et Xue, 2007; Vermeer, 2004 ou Tidball et

Treffers-Daller, 2007). Ainsi, plus un texte est long, plus la courbe de type-occurrence décroît, en raison de la récurrence des mots les plus fréquents, comme les mots fonctionnels. Plusieurs transformations mathématiques du rapport de type-occurrence ont été proposées afin d'éliminer l'effet de la longueur du texte, par exemple, le *Guiraud's Index* ou la mesure D de la diversité lexicale (Malvern, Richards, Chipere et Durán, 2004). Selon Jarvis (2002), la mesure D fournit une courbe correcte. Toutefois, en ce qui concerne la longueur du texte, Jarvis (2002, p. 81) admet que les textes utilisés dans son étude comportaient au plus 500 mots, et qu'il serait intéressant d'étudier des textes plus longs lors de recherches futures. McCarthy et Jarvis (2007) estiment que la mesure D est la meilleure de celles qu'ils ont examinées, mais qu'elle reste néanmoins dépendante de la longueur du texte. Selon ces chercheurs, la mesure semble stable pour les textes de 100 à 400 mots, mais il serait souhaitable d'avoir une mesure stable pour des textes comportant jusqu'à 2 000 mots. McCarthy et Jarvis (2007, p. 481) soulignent toutefois que Malvern et coll. (2004) n'ont jamais travaillé avec des textes plus longs. La question des textes plus longs n'a donc pas été approfondie.

Le rapport de type-occurrence et ses différentes transformations et modélisations mathématiques calculent la richesse lexicale de façon purement quantitative à partir d'un texte donné, sans tenir compte des aspects qualitatifs. Une autre méthode d'évaluation de la richesse lexicale consiste à tenir compte de la fréquence des mots de la langue, comme le font les méthodes basées sur les listes de mots. En réalisant une distinction entre mots fréquents d'un côté et mots rares de l'autre dans une langue donnée, on ajoute une perspective qualitative de la richesse lexicale (Daller, Van Hout et Treffers-Daller, 2003).

Bon nombre de mesures de la richesse lexicale se basent sur les listes de fréquence. Vermeer (2004) propose la mesure MLR (*Measure of Lexical Richness*) pour le néerlandais parlé L1 et L2 des enfants et insiste sur l'importance de la fréquence et du degré de difficulté des mots dans les données d'entrée pour mesurer la richesse lexicale. Selon Vermeer (2004, p. 176), le problème principal des mesures basées sur le rapport type-occurrence est qu'elles ne tiennent pas compte de la difficulté du mot. Il y a pourtant d'après l'auteur un rapport évident entre le degré de difficulté d'un mot et l'ordre d'acquisition : les mots les plus fréquents sont aussi les plus faciles à acquérir, et vice versa. La mesure de la richesse lexicale (MLR) est calculée à partir d'une base de données composée de 2 millions de mots provenant des données d'entrée écrites et orales à l'école primaire des Pays-Bas. Vermeer (2004) distingue neuf catégories de fréquence, qui correspondent à neuf niveaux de difficulté. Après avoir effectué une

comparaison du rapport type-occurrence et de la MLR, Vermeer a pu constater que cette dernière fournit une meilleure mesure, car elle permet de distinguer des groupes d'informants ayant des niveaux différents de vocabulaire. D'après Vermeer (2004, p. 185), la MLR serait de plus indépendante de la longueur des textes. Noter toutefois que les textes utilisés sont relativement longs (environ 1 000 mots). Il n'est donc pas certain que la MLR fonctionne de façon satisfaisante pour des textes plus courts[1]. Toutefois, selon Vermeer (2004, p. 186), une des limites de l'étude est que la mesure a été développée en premier lieu pour des enfants de niveau primaire. Ainsi, elle pourrait ne pas permettre de trancher entre des apprenants des niveaux plus avancés.

D'autres approches à la richesse lexicale combinent les mesures basées sur la distinction type-occurrence et celles fondées sur les listes de fréquence, telles que les mesures *Advanced TTR* et *Guiraud Advanced* proposées par Daller et coll. (2003, aussi Tidball et Treffers-Daller, 2007, 2008). Si l'on se fie aux études de fréquence pour analyser la richesse lexicale, les mesures basées sur des listes de mots seraient plus fiables. À notre avis, il est préférable d'étudier la richesse lexicale en tenant compte de cette dimension qualitative. Si nous voulons nous prononcer sur la richesse lexicale des apprenants, il est primordial de procéder à une analyse détaillée de leur vocabulaire. Ce qui ne veut pas dire qu'on doive éviter les approches quantitatives, mais plutôt qu'il est nécessaire de les compléter par des analyses plus qualitatives afin de dégager une image plus étoffée de leur richesse lexicale. Mesurer la richesse lexicale au moyen de la fréquence semble constituer un bon point de départ qui permet de tenir compte de la difficulté du mot, qui est mesurée selon la fréquence du mot (Vermeer, 2004). De plus, après la répartition dans les zones de fréquence, il est possible d'approfondir l'analyse de façon plus qualitative.

En ce qui a trait à la richesse lexicale, nous adoptons dans la présente étude la conception d'Ovtcharov et coll. (2006, p. 110), qui insiste sur l'aspect qualitatif du vocabulaire : « la richesse d'un vocabulaire ne se résumant pas à la simple addition des mots qui le compose concerne plutôt leur aspect qualitatif, et pour mieux définir cette richesse, il faut avoir recours à la notion de rareté de mots ». En d'autres termes, les chercheurs font la distinction entre mots rares d'un côté et mots communs de l'autre. Leur idée principale est que la plupart des locuteurs d'une langue ont recours aux mots les plus fréquents, tandis que les locuteurs natifs et les apprenants très avancés ont accès dans une plus large mesure aux mots les moins fréquents. On pourrait alors dire que les approches quantitative et qualitative convergent, dans le sens où l'on fait référence à la fréquence lexicale tout en accordant de l'importance au caractère du mot. En mettant l'accent sur la rareté des mots,

Ovtcharov et coll. (2006, p. 110) définissent ainsi la richesse lexicale : « le nombre de mots plus spécialisés, plus rares, monosémiques, porteurs de sens connotatif minimal et dont l'emploi se limite à un usage soit professionnel ou strictement thématique ». Selon les chercheurs, dans le cadre de l'acquisition d'une langue seconde, ce genre de mots serait alors surtout relevé dans le lexique d'apprenants L2 très avancés.

Profil de fréquence lexicale (Laufer et Nation, 1995)

La méthode PFL a été élaborée à l'origine par Laufer et Nation (1995) afin de mesurer la richesse lexicale en anglais L2 écrit. L'idée principale de la méthode est de diviser une production donnée en quatre zones de fréquence : le premier millier contient le premier millier de familles de mots les plus fréquentes, le deuxième millier de familles de mots, les mots académiques et les mots hors listes (MHL), ceux qui ne se trouvent dans aucune des trois premières zones. Laufer et Nation ont également développé un logiciel permettant le tri des mots d'un texte dans les quatre zones mentionnées. La répartition dans les zones de fréquence constitue le profil lexical d'un apprenant. Plus la proportion de mots dans les zones de faible fréquence est élevée, plus le profil lexical sera considéré comme avancé. Selon Laufer et Nation (1995, p. 312–313), les avantages de la méthode PFL seraient son objectivité et son indépendance en ce qui a trait à la syntaxe et à la cohésion du texte. En outre, les auteurs soulignent que la méthode se concentre uniquement sur le lexique, ce qui la rend plus apte que d'autres méthodes à mesurer la richesse lexicale.

Laufer et Nation (1995, p. 313) ont émis deux hypothèses. D'abord, ils s'attendaient à ce que le profil lexical de divers échantillons issus d'une même étape d'apprentissage soit égal. Il serait alors indépendant du type de texte. Ensuite, ils prévoyaient que la méthode permettrait de trancher entre divers niveaux de compétence linguistique. Ces deux hypothèses ont été confirmées dans leur étude. Dans la présente étude, notre analyse porte sur deux groupes d'apprenants de niveaux d'acquisition différents. La méthode PFL devrait ainsi nous permettre de constater des différences de profils lexicaux chez ces deux groupes.

Version française du Profil de fréquence lexicale

La méthode PFL a été adaptée au français en plusieurs étapes (Goodfellow et coll., 2002; Cobb et Horst, 2004). La version française est basée sur le corpus de Verlinde et Selva (2001), qui rassemble des textes des journaux *Le Monde* (France) et *Le Soir* (Belgique) et compte quelque 50 000 000 de mots écrits. Le logiciel Vocabprofile, qui permet

de diviser une production en zones de fréquence, a été adapté pour le français par T. Cobb[2]. Le contenu des zones de fréquence de la version française est légèrement différent de celui de la version anglaise. Ainsi, la zone contenant les mots académiques en anglais n'a pas de zone correspondante en français, et a été remplacée par le troisième millier de mots dans la version française (Cobb et Horst, 2004).

On peut évidemment s'interroger sur la pertinence de l'emploi d'une base de données écrites pour l'examen de la langue orale. Il est évident que certains mots appartiennent spécifiquement à l'oral, et d'autres, à l'écrit (Tidball et Treffers-Daller, 2008; McCarthy, 1998). Ainsi, une analyse de la langue parlée se fondant sur une classification établie à partir de la langue écrite pourrait conduire à classer certains mots dans la « mauvaise zone ». Comme le souligne McCarthy (1998), certains mots présentent des différences de fréquence en fonction du registre (parlé/écrit). C'est donc l'hypothèse des différences de fréquence en langue parlée et écrite que nous voulons examiner dans cette étude. Comme nous l'avons déjà mentionné, Ovtcharov et coll. (2006, p. 121) défendent la pertinence de la méthode PFL pour les données orales, et affirment que « si l'on utilise un système de mesures et des unités de mesure identiques pour effectuer des déterminations sur des mesures de même nature, qu'on les mesure en pouces ou en centimètres, les mesures seront intégrales et exprimeront équitablement les valeurs des différences observées ». Nous pensons qu'ils ont raison dans la mesure où il est possible de comparer des groupes qui effectuent la même tâche. Toutefois, dans une perspective plus qualitative, nous nous interrogeons quant à la répartition dans les zones de fréquence, puisque cette dernière est basée sur la langue écrite. Il ne semble pas plausible que la correspondance soit exacte entre la fréquence lexicale en langue parlée et celle en langue écrite (Campione et coll., 2005, pour un survol des fréquences des classes de mots à l'écrit et à l'oral). Ainsi, le profil lexical d'une production orale pourrait se distinguer en fonction du type de base de données. Nous espérons que les résultats de la présente étude contribueront à éclairer la question de l'emploi d'une base de données écrites pour étudier la langue parlée. Il faut souligner que très peu d'études ont testé la méthode PFL sur la langue orale, ou sur le français (Ovtcharov et coll., 2006, p. 119). Ces rares études sont présentées dans la section suivante.

Études antérieures qui appliquent le PFL au français L2

L'étude de Goodfellow et coll. (2002) est la première à avoir tenté d'adapter la méthode PFL au français. L'étude a un objectif plutôt

pédagogique, en ce que les chercheurs comparent le profil lexical de compositions écrites par des apprenants anglophones dits intermédiaires faibles (*low intermediate*) et les notes des professeurs. Ces résultats contredisent ceux de Laufer et Nation (1995) : l'emploi des mots de faible fréquence ne semble pas être en corrélation avec à la richesse du vocabulaire, telle qu'elle est évaluée par le professeur. Ainsi, Goodfellow et coll. (2002, p. 139) constatent que les mots hors listes ne permettent pas de distinguer les informants. D'après les chercheurs, la tâche, un test de compréhension, limiterait les possibilités d'emploi de mots rares. En outre, les chercheurs émettent l'hypothèse que tous les informants étaient influencés par les données d'entrée dans la mesure où ils en bénéficiaient dans leurs réponses. Ainsi, même les apprenants les moins avancés arrivaient à reproduire des mots « avancés », c'est-à-dire des mots rares. Il était donc difficile de savoir si les apprenants connaissaient ou non ces mots avant l'exercice. Nous pouvons en déduire que le PFL n'a pas été validé dans cette première tentative d'adaptation, car, selon Goodfellow et coll. (2002), les profils lexicaux obtenus ne reflètent pas le niveau linguistique des informants.

À l'aide du PFL, Granfeldt (2006) fait une analyse de quarante textes écrits en français L2 par des apprenants locuteurs natifs de suédois. Selon lui, les apprenants se trouvent à deux niveaux différents en fonction de critères morphologiques : aux stades post-initial et avancé inférieur, soit aux stades 2 et 4 des six stades proposés par Bartning et Schlyter (2004, ci-dessous). Les résultats ne présentent pas de différences significatives entre les deux groupes en ce qui a trait à la répartition dans les zones de fréquence. Il peut y avoir plusieurs explications à ces résultats. D'abord, il se peut que les deux groupes ne se trouvent pas en réalité à deux niveaux d'acquisition distincts. Ou encore, les apprenants sont à deux niveaux différents, mais l'analyse PFL ne peut pas montrer cette différence. Ensuite, nous pouvons nous demander si la tâche utilisée – des textes écrits produits à partir d'une série d'images – est trop restreinte pour que le programme puisse en saisir les différences. Il est vrai que cette tâche requiert un certain type de connaissances et suscite un vocabulaire plus ou moins obligatoire, par exemple, la référence aux personnes et aux objets « clés ». Pour sa part, Granfeldt (2006) conclut que la tâche n'est pas tout à fait adaptée. De plus, il se peut que l'échantillon soit trop limité pour effectuer des calculs statistiquement fiables. Ici encore, la méthode PFL appliquée au français écrit n'a pas réussi à distinguer entre des apprenants de niveaux de compétence linguistique différents. Cela ne signifie pas toutefois que la méthode n'est pas valide. Il semble

que les deux études ont un point problématique en commun, le type de tâche utilisé. Ainsi, Granfeldt (2006) et Goodfellow et coll. (2002) admettent qu'il est possible que les tâches n'aient pas été entièrement adaptées à ce genre d'analyse. Dans les deux études, l'emploi de tests assez limités a pu mener à des résultats moins satisfaisants.

Ovtcharov et coll. (2006) veulent vérifier « l'intuition commune » selon laquelle les apprenants avancés ont un vocabulaire très riche, un vocabulaire contenant moins de « mots de base » et davantage de mots spécialisés et peu fréquents. Les auteurs partent du postulat que la plupart des locuteurs d'une langue ont recours aux « mots de base », aux mots les plus communs, tandis que les locuteurs natifs instruits et les apprenants très avancés font usage de mots spécialisés ou peu fréquents. Dans la terminologie du PFL, les apprenants avancés auraient donc une proportion élevée de mots des zones de faible fréquence K3 et MHL. Les données comprennent les productions orales de 48 participants, tous des locuteurs natifs de l'anglais apprenant le français. Les enregistrements ont été effectués en situation d'examen, notamment au moment d'une interview orale. Les thèmes discutés sont principalement la vie professionnelle, mais aussi la vie en dehors du travail. Les apprenants se situent à deux principaux niveaux différents de leur acquisition : intermédiaire et avancé[3]. Ensuite, une division supplémentaire a été faite pour répartir les apprenants en quatre groupes : intermédiaire fort/faible et avancé fort/faible. Les résultats affichent des différences significatives entre tous les groupes d'apprenants. Les apprenants les plus avancés ont produit plus de mots dans les zones K3 + MHL, c'est-à-dire dans les zones de faible fréquence, mais aussi dans la zone K2. Cela étant dit, les chercheurs estiment avoir montré la validité de la méthode PFL, étant donné que cette dernière permet de faire une distinction entre les apprenants à différents niveaux de compétence linguistique. Nous avons aussi constaté qu'il n'y a pas de différence statistiquement significative entre les apprenants les plus avancés et les locuteurs natifs du corpus de contrôle. Selon les chercheurs, cela indiquerait que la richesse lexicale des apprenants très avancés se rapproche de celle des locuteurs natifs.

Tidball et Treffers-Daller (2008) utilisent la méthode PFL (entre autres mesures de la richesse lexicale) pour comparer le vocabulaire employé dans des descriptions orales de bandes dessinées, faites par deux groupes d'apprenants anglophones de français. Les résultats indiquent que les catégories MHL et K1 permettent de distinguer les groupes, ce que ne fait pas la catégorie K3. En réalité, Tidball et Treffers-Daller (2008, p. 310) constatent qu'une détermination basée sur des évaluations

de professeurs du vocabulaire avancé permet mieux de distinguer entre deux niveaux de compétence que les mesures basées sur les fréquences. Par ailleurs, Tidball et Treffers-Daller (2008) soulèvent la question de la pertinence de l'emploi d'une base de données écrites pour l'analyse de données orales.

En résumé, nous pouvons constater que l'étude d'Ovtcharov et coll. (2006) sur le français parlé montre la pertinence de la méthode PFL, tandis que les études sur le français écrit n'ont pu la valider, puisqu'elle n'a pas permis de trancher entre des niveaux de compétence linguistique différents. Quant à elle, l'étude de Tidball et Treffers-Daller (2008) montre en partie la validité de la méthode, étant donné que la catégorie MHL permet d'y distinguer les groupes d'apprenants. Ce qui n'est pas le cas de la catégorie K3. Notre propre étude présente beaucoup de similarités avec celle d'Ovtcharov et coll. (2006). D'abord, elle traite du français parlé. Ensuite, le type de tâche est le même, soit une interview avec un locuteur natif de français. C'est-à-dire que les données sont composées de productions parlées, relativement spontanées. Enfin, elle concerne l'apprenant avancé de français. Étant donné ces ressemblances, il semble plausible que la méthode PFL soit applicable à notre propre étude. Certes, l'étude de Tidball et Treffers-Daller (2008) concerne aussi le français parlé. Toutefois, le type de tâche ressemble plus à celles utilisés par Goodfellow et coll. (2002) et Granfeldt (2006), qui n'ont pas démontré la validité de la méthode.

Questions de recherche et hypothèses

Sur la base des études antérieures, nous nous posons trois questions en matière de recherche. 1) La méthode PFL est-elle un outil d'analyse fiable pour le français parlé, même si la base de données sur laquelle elle se fonde est composée de documents écrits? Pour être fiable, la méthode doit répondre aux critères suivants : a) elle doit pouvoir trancher entre des niveaux de compétence linguistique de divers apprenants. En d'autres termes, il doit y avoir des différences entre les profils lexicaux des deux groupes d'apprenants ; b) la division en zones de fréquence doit être fiable pour la langue orale. Étant donné que la base de données est composée de mots provenant de la langue écrite, et que notre corpus contient des données de la langue parlée, on pourrait s'attendre à des différences dans la répartition des mots dans les différentes zones. Autrement dit, un mot peut être fréquent dans la langue écrite mais rare dans la langue parlée, et inversement. Toutefois, comme nous venons de le voir, Ovtcharov et coll.

(2006) constatent que la méthode PFL est applicable à la langue parlée. 2) Y a-t-il des différences de profils lexicaux entre les groupes d'apprenants? 3) S'il y a des différences, reflètent-elles le niveau de compétence linguistique des apprenants?

Nous émettons les hypothèses suivantes. 1) La méthode PFL est fiable et applicable aux données orales (Ovtcharov et coll., 2006). 2) Si nous supposons que les apprenants se trouvent à deux niveaux de compétence distincts, et s'il y a une corrélation entre le niveau linguistique et le nombre de mots rares, l'étude devrait montrer que les apprenants les plus avancés produisent proportionnellement plus de mots dans les zones K2, K3 et MHL que les apprenants les moins avancés (résultat d'Ovtcharov et coll., 2006). 3) Les apprenants les plus avancés devraient se rapprocher des locuteurs natifs quant à la proportion de mots dans les zones K3 + MHL (aussi le résultat d'Ovtcharov et coll., 2006). 4) L'étude de Laufer et Nation (1995) sur la production écrite d'apprenants d'anglais L2 a démontré qu'il existait une différence significative entre les informants des zones K3 et MHL. Toutefois, les deux études sur le français parlé évoquées plus haut (Ovtcharov et coll., 2006; Tidball et Treffers-Daller, 2008) en sont arrivées à des résultats divergents. Tandis que les premiers ont constaté des différences significatives entre les groupes d'apprenants des zones K2, K3 et MHL, les catégories K1 et MHL ont permis aux dernières de distinguer les groupes. Étant donné que le type de tâche utilisé dans la présente étude se rapproche davantage de celui d'Ovtcharov et coll. (2006), il semble plausible que nos résultats aillent dans le même sens.

Données et méthode

Données

Pour vérifier la validité de la méthode PFL, il est primordial d'effectuer des analyses de locuteurs de niveaux d'acquisition différents. Nous savons qu'il est difficile de se prononcer de façon définitive sur la compétence linguistique d'apprenants d'une langue seconde. Le nombre d'années d'acquisition peut être une indication approximative du niveau de compétence, mais d'autres facteurs devraient jouer un rôle important. Les informants de la présente étude ont été classés à différents stades d'acquisition selon les critères morphosyntaxiques de Bartning et Schlyter (2004), qui ont proposé, sur la base d'analyses du corpus *InterFra* de l'Université de Stockholm et du corpus de

l'Université de Lund, six stades de développement pour le locuteur natif de suédois apprenant le français L2 : 1) stade initial, 2) stade post-initial, 3) stade intermédiaire, 4) stade avancé inférieur, 5) stade avancé moyen, 6) stade avancé supérieur. Le stade dit avancé comporte trois stades (4, 5 et 6). Étant donné que nous nous intéressons à la richesse lexicale chez l'apprenant avancé, nous avons choisi de comparer des apprenants au stade 6 – le stade avancé supérieur (sept enregistrements) à des apprenants au stade 4 – le stade avancé inférieur (sept enregistrements). Les apprenants au stade avancé supérieur (trois hommes et quatre femmes de 25 à 32 ans) sont des étudiants et des doctorants à l'Université de Stockholm. Le nombre total de mots produits s'élève à 13 000. Les apprenants au stade avancé inférieur sont tous étudiants à l'Université de Stockholm. Sept enregistrements (sept femmes de 19 à 29 ans) sont analysés. Le nombre total de mots est d'environ 9 000. Nous avons aussi un groupe de contrôle composé de dix locuteurs natifs de français – tous étudiants en échange Erasmus à l'Université de Stockholm (huit femmes et deux hommes de 19 à 26 ans). Leurs productions comptent au total quelque 25 000 mots.

Tous les informants ont effectué la même tâche : une interview semi-guidée d'environ 15 minutes avec un locuteur natif de français. Les thèmes de l'interview sont notamment les études à l'université, les loisirs, et la situation familiale. Les enregistrements font partie du corpus InterFra (Bartning et Schlyter, 2004) et ont été classés par Sanell (2007), aux stades de développement mentionnés, en fonction d'une vingtaine de critères morphosyntaxiques. Nous sommes consciente du nombre restreint d'enregistrements des différents groupes d'apprenants, mais nous ne disposons malheureusement pas pour le moment d'un plus grand nombre d'informants dont les enregistrements ont été classés aux stades avancés inférieur et supérieur. Nous espérons résoudre ce problème plus tard. Nous sommes aussi consciente du fait que la longueur des interviews varie, et qu'elles sont en général relativement longues (environ 1 300 mots en moyenne au stade avancé inférieur, 1 900 mots en moyenne au stade avancé supérieur, et 2 500 mots en moyenne chez les locuteurs natifs). En réalité, la longueur idéale des textes à soumettre à l'analyse du PFL n'a pas encore été déterminée. Toutefois, Laufer (1995, p. 267) estime que le profil (anglais) est stable entre 200 et 400 mots, mais il n'y pas de conclusions définitives à ce sujet. On pourrait supposer que la longueur des textes peut avoir un effet sur les résultats, dans la mesure où les mots les plus fréquents (comme les articles ou les auxiliaires) puissent revenir plusieurs fois. Ainsi, il est probable que le taux de mots en K1 augmentera dans le cas de textes longs. Néanmoins, rien n'empêche à notre avis que les textes examinés soient relativement longs[4].

Préparation des données

Dans le traitement des textes, Vocabprofil effectue quelques changements automatiques. Il remplace les chiffres par le mot « nombre », pour les exclure du comptage. Pour ce qui est des mots contractés, Vocabprofil les sépare : *j'ai* > *je ai*, *l'* > *le/la*, *s'* > *se/si*. Outre ces changements, il a fallu préparer les interviews transcrites avant de les soumettre à Vocabprofil. Cette préparation a consisté à supprimer les noms propres, les énoncés de l'intervieweur, les pauses remplies (*mm*, *mhm*, *euh*, *eh*, etc.)[5], les commentaires extra-linguistiques (bruits, rires, etc.), les mots inachevés (*ét-* étudiant) et les mots appartenant à d'autres langues que le français.

Méthode

Après avoir préparé les fichiers de la façon décrite, nous les avons soumis à Vocabprofil, d'abord un fichier par informant, puis un fichier par groupe. Les données de sortie montrent la répartition des mots du texte dans les quatre zones de fréquence K1, K2, K3 et MHL[6]. Afin de voir si les différences entre les trois groupes sont statistiquement significatives (au seuil de 5 %, soit $p < 0{,}05$), nous avons utilisé le test ANOVA, qui permet de comparer les moyennes constatées dans plusieurs groupes. Le test ANOVA a été réalisé avec le test post-hoc de Tukey, qui montre quels groupes présentent des différences significatives.

Résultats et analyse

Les profils lexicaux des informants

Le tableau suivant montre le profil lexical des différents groupes d'informants d'après les données de sortie de Vocabprofil. La répartition dans les zones de fréquence K1, K2, K3 et MHL est exprimée en pourcentages. Étant donné que nous nous intéressons surtout à la proportion de mots rares, nous avons réuni dans la dernière colonne les mots de la zone K3 et les mots hors listes[7]. Après quelques remarques générales portant sur toutes les zones de fréquence, nous concentrerons notre analyse sur les zones K3 et MHL.

Dans l'ensemble, la proportion est élevée dans la zone K1, celle qui inclut les mots les plus fréquents de la langue française. En effet, tous les groupes présentent une proportion de plus de 90 % dans cette zone de fréquence. Toutefois, il ressort du tableau que la proportion de mots

TABLEAU 1
La répartition des mots dans les différentes zones de fréquence (%)

Groupes	K1	K2	K3	MHL	K3 + MHL
Locuteurs natifs (LN) ($n = 10$)	91.93	3.43	0.84	3.80	4.64
Apprenants au stade avancé supérieur (AAS) ($n = 7$)	91.69	3.47	0.99	3.85	4.84
Apprenants au stade avancé inférieur (AAI) ($n = 7$)	94.10	2.89	0.59	2.43	3.02

en K1 diminue en fonction de la compétence linguistique, puisque les apprenants au stade avancé supérieur (AAS) présentent une proportion plus faible que les apprenants au stade avancé inférieur (AAI). Il semble donc qu'une grande proportion des mots les plus fréquents, ceux de la zone K1, indique un niveau de compétence linguistique plus bas. Pour ce qui est des autres zones – K2, K3 et MHL – le tableau indique que les AAS ont une proportion plus élevée que les AAI, ce qui est conforme aux résultats de deux études antérieures sur le français (Goodfellow et coll., 2002 et Ovtcharov et coll., 2006), c'est-à-dire que les apprenants les plus avancés ne se distinguent pas des moins avancés uniquement par les zones de fréquence K3 et MHL, mais aussi par la zone K2.

Il semble donc que les apprenants les moins avancés (AAI) se distinguent des autres groupes dans toutes les zones de fréquence, comme nous l'avions prévu. Ils ont une proportion plus élevée de mots dans la zone K1 et une proportion moins élevée de mots dans les zones K2, K3 et MHL. Toutefois, les différences entre les locuteurs natifs et les apprenants plus avancés (AAS) semblent très faibles. Afin de pouvoir nous prononcer sur la question des différences et de leur éventuelle signification statistique, nous avons effectué un test ANOVA pour comparer les moyennes dans les trois groupes (Tableau 2). Ce test confirme l'existence d'une différence significative, $p < 0{,}05$, quant au nombre de mots produits dans les toutes les zones de fréquence : $F(2,$

TABLEAU 2
La répartition des mots dans les différentes zones de fréquence en moyenne (%)

Groupes	K1	K2	K3	MHL	K3 + MHL
Locuteurs natifs (LN) ($n = 10$)	91.99	3.46	0.84	3.71	4.62
Apprenants au stade avancé supérieur (AAS) ($n = 7$)	91.86	3.48	0.98	3.67	4.65
Apprenants au stade avancé inférieur (AAI) ($n = 7$)	94.14	2.84	0.55	2.46	3.09

TABLEAU 3
Résultats des tests post-hoc de Tukey (valeurs *p*)

Groupes comparés	K1	K2	K3	MHL	K3 + MHL
AAI vs. AAS	0.438	0.082	0.078	0.130	0.082
AAI vs. LN	0.009*	0.001*	0.014*	0.006*	0.003*
AAS vs. LN	0.144	0.139	0.809	0.411	0.423

*La différence est significative au seuil de 5 % ($p < 0{,}05$).

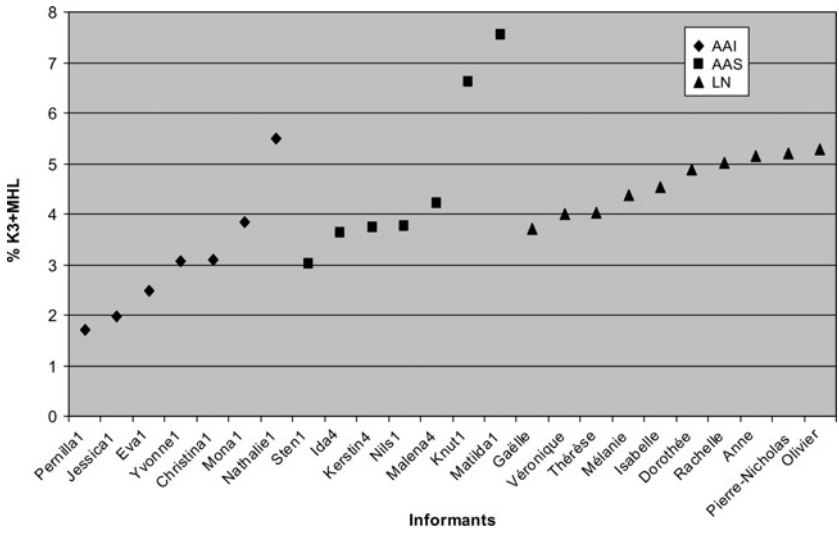
21) = 5,757, $p = 0{,}01$ pour la zone K1; $F(2{,}21) = 9{,}958$, $p = 0{,}001$ pour la zone K2; $F(2{,}21) = 5{,}13$, $p = 0$, 015 pour la zone K3; $F(2{,}21) = 6{,}134$, $p =$ 0,008 pour la zone MHL; et, finalement pour les zones de faible fréquence K3 + MHL réunies : $F(2{,}21) = 7{,}039$, $p = 0{,}005$.

Le test ANOVA a été suivi du test post-hoc de Tukey, qui permet de préciser où se trouvent exactement les différences significatives (Tableau 3). Nous pouvons constater qu'il y a une telle différence entre les groupes AAI et LN par rapport à toutes les zones de fréquence, ce qui indiquerait une différence générale dans les profils lexicaux de ces groupes. Les apprenants moins avancés auraient donc un vocabulaire moins riche que les LN, ce qui est conforme à notre hypothèse de départ. Par contre, nous n'avons constaté aucune différence significative du point de vue statistique entre les LN et les AAS, ni entre les AAI et les AAS. L'absence de différence significative entre les LN et les AAS confirme notre hypothèse et pourrait indiquer que le profil lexical des apprenants les plus avancés est similaire à celui des locuteurs natifs. Ce résultat confirme aussi notre hypothèse selon laquelle les apprenants les plus avancés auraient un vocabulaire aussi riche que les locuteurs natifs (Ovtcharov et coll., 2006). Toutefois, nous admettons que le nombre d'informants est restreint et qu'il faut traiter ces résultats avec prudence. Par contre, les tests statistiques ne nous permettent pas de trancher de façon claire et nette entre les deux groupes d'apprenants. Ce résultat n'est pas conforme à notre hypothèse de départ, et il se distingue des résultats d'Ovtcharov et coll. (2006). Il peut y avoir plusieurs explications. D'abord, il est possible que le nombre total d'enregistrements (24) soit trop restreint pour que le calcul statistique soit entièrement fiable. Ensuite, il est à noter que la proportion des mots dans les zones K3 + MHL, qui devraient en premier lieu permettre de trancher entre les groupes d'apprenants, est relativement faible dans tous les groupes, puisqu'elle varie de 3,02% à 4,84%. Certes, Nation (2006, p. 79) soutient que le taux de mots de haute fréquence est généralement plus élevé dans la langue parlée que dans la langue écrite. Le taux de

mots peu fréquents serait ainsi moins élevé dans la langue parlée. Toutefois, nous pouvons mentionner que les chiffres correspondants dans l'étude d'Ovtcharov et coll. (2006) varient de 4,79 % à 12,07 % (48 enregistrements). Nous pouvons en déduire que la production lexicalement plus riche dans nos données est presque égale à la production la moins riche dans Ovtcharov et coll. (2006)[8]. Il est difficile d'en expliquer la raison. Nous pourrions penser que le fait que nos informants ont une large proportion de mots dans la zone K1 et ainsi une proportion relativement faible dans les zones K3 + MHL, soit l'effet de la longueur des textes, dans la mesure où les mots les plus fréquents ont tendance à être répétés plusieurs fois. En effet, les interviews comptent environ 1 900 mots en moyenne, ce qui est un chiffre assez élevé par rapport à d'autres études[9]. On pourrait croire que le pourcentage de mots dans la zone K1 serait moins élevé dans un texte plus court. Pourtant, il semble que les interviews dans l'étude d'Ovtcharov et coll. (2006) comptent environ 2 400 mots en moyenne. Par conséquent, la longueur des textes ne semble pas expliquer les différences de proportions en K3 + MHL dans leur étude, ni dans la nôtre. Peut-être aussi que les apprenants dans l'étude d'Ovtcharov et coll. (2006) se trouvent à un niveau de compétence linguistique plus avancé en général, et que cette situation se reflète dans leur richesse lexicale. Toutefois, les locuteurs natifs de notre étude ont, eux aussi, une proportion très faible de mots rares, ce qui nous conduit à penser que les thèmes des interviews pourraient donner lieu à un vocabulaire plus ou moins avancé. Bien que les interviews dans Ovtcharov et coll. (2006) semblent être globalement structurées de la même manière que les nôtres, il pourrait exister des différences en ce qui a trait à certains thèmes faisant l'objet de discussion. Nous pouvons par exemple supposer que le thème du travail soit plus amplement discuté par les apprenants dans l'étude d'Ovtcharov et coll. (2006), étant donné qu'ils ont déjà l'expérience du marché du travail et que les interviews ont pour thème principal la situation du travail. Il est plausible que ce thème suscite l'emploi d'un vocabulaire spécialisé, ce qui devrait augmenter le nombre de mots dans les zones de faible fréquence. Par contre, les étudiants de notre étude n'ont pas, en général, beaucoup d'expérience professionnelle, et le thème du travail n'est que brièvement discuté. Par conséquent, ils n'auraient pas l'occasion de se servir au même degré de mots spécialisés[10].

Pour revenir à l'absence de différences statistiquement significatives entre les AAI et les AAS, il est possible qu'il y ait des différences individuelles au sein des groupes, ce qui indiquerait que nous avons affaire à des groupes hétérogènes. Regardons à ce propos la figure ci-dessous,

FIGURE 1
Le pourcentage de mots en K3 + MHL chez les informants



où nous nous concentrons sur les zones de faible fréquence, car ce sont surtout ces zones qui sont censées montrer des différences de niveau de compétence linguistique général.

Il est clair que les deux groupes d'apprenants sont hétérogènes, tandis que la variation intra-groupe n'est pas aussi prononcée chez les LN. En effet, deux des informants du groupe AAS ont des proportions de mots des zones K3 + MHL beaucoup plus élevées que les autres membres du groupe, et même supérieures à celles des LN. Nous rappelons que les apprenants ont été classés aux stades de développement selon des critères morphosyntaxiques. D'après ces critères, tous les AAS se trouvent donc au même stade développemental. Pourtant, comme nous venons de le voir, deux des AAS se distinguent des autres. Cela nous mène à penser que le développement de la morphosyntaxe ne coïncide pas toujours avec celui du vocabulaire.

Néanmoins, nous pouvons distinguer un développement relativement systématique de la proportion de mots des zones K3 + MHL, dans la mesure où cette proportion augmente selon le niveau de compétence. Ainsi, il semble que les résultats reflètent en gros la compétence linguistique des informants. Toutefois, en ce qui a trait à la fiabilité de la méthode PFL, nous avons proposé qu'elle doit permettre de trancher entre des niveaux de compétence linguistique différents des apprenants. Nous pouvons constater que ce critère n'est pas

rempli, car le test statistique ne nous permet pas d'établir l'existence de différences significatives entre les deux groupes d'apprenants.

L'autre critère de fiabilité de la méthode était que notre analyse se base sur la langue orale, tandis que la division en zones de fréquence se fonde sur des documents écrits. Nous aurions alors intérêt à nous interroger sur la nature des mots classés dans les zones K3 + MHL. Que nous apprennent-ils sur la compétence lexicale? En examinant de façon plus approfondie les données de sortie de Vocabprofil, nous avons découvert que, si la zone K3 semble en effet contenir des mots « avancés » (*flux*, *promouvoir*, *apercevoir*, *préoccupation*, *significatif*, *incapable*), les mots hors listes semblent de nature beaucoup plus variée. La question de la nature des mots hors listes est abordée dans la prochaine section.

Quels sont les mots hors listes?

De quelle nature sont les mots hors listes, les mots qui ne sont pas inclus dans les 3 000 familles de mots les plus fréquentes du corpus de Verlinde et Selva (2001)? En d'autres termes, que signifie une proportion élevée de mots hors listes? Selon notre hypothèse de départ, une proportion élevée de mots rares, mots dans les zones K3 + MHL, serait l'indice d'un vocabulaire riche et évolué. Étant donné que nous avons constaté une proportion relativement faible de ces zones de fréquence dans tous nos groupes de informants (Ovtcharov et coll., 2006), il nous semble intéressant de les analyser plus en détail. Tout en admettant que les apprenants moins avancés en produisent une faible proportion, nous pourrions néanmoins nous attendre à un taux plus élevé chez les apprenants plus avancés, et surtout chez les locuteurs natifs. Nous nous concentrerons maintenant sur la zone MHL, qui semble contenir des mots de caractère divers, et pas seulement des mots rares et avancés.

Les mots de loin les plus récurrents parmi les mots hors listes, tous groupes réunis, sont *ben* (125 occurrences au total), *ouais* (117) et *suédois* (114). Du point de vue quantitatif, ces mots se distinguent de façon remarquable des autres mots de la catégorie MHL. La plupart des autres mots ne reviennent guère plus de 10 fois. *Ben*, *ouais* et *suédois* ne figurent donc pas parmi les 3 000 mots les plus fréquents dans le corpus de Verlinde et Selva (2001). Il n'est pas difficile d'en comprendre la raison. Leur corpus est constitué d'articles de journaux francophones. Les mots *ben* et *ouais* sont des marqueurs de discours typiques de l'oral, que l'on ne retrouve pas souvent dans la langue écrite (McCarthy, 1998; aussi Tidball et Treffers-Daller, 2008, p. 303, selon

lesquelles la catégorie MHL contient non seulement des mots avancés, mais aussi des mots fréquents à l'oral mais non à l'écrit). Le mot *suédois* (l'adjectif ou le substantif) n'est apparemment pas souvent utilisé dans un contexte francophone. Il est naturel que notre corpus contienne le mot *suédois* en grande quantité, puisqu'il a été enregistré en Suède. Outre ces mots fréquents, nous trouvons une grande variété de mots qui ne semblent pas toujours indiquer un vocabulaire riche et avancé. En voici quelques exemples.

Mot hors listes = mot de la langue orale?

Aux mots *ben* et *ouais* s'ajoutent d'autres marqueurs de discours dans nos données : *aha*, *ouf*, *ok*, ainsi que d'autres mots typiques de la langue parlée : *rigolo*, *prof*, *extra*, *sympa*, *truc*, *boulot*, *cool*, *super*. Nous pouvons nous demander dans quelle mesure tous ces mots, qui appartiennent à la langue orale, doivent être considérés comme des indices d'un vocabulaire avancé ou riche. D'abord, ces mots ont été classés dans la liste MHL parce qu'ils ne se trouvent pas parmi les 3 000 familles de mots les plus fréquentes du corpus journalistique de Verlinde et Selva (2001). Il est évident que ces mots ne sont pas souvent employés dans ce genre de texte. Nous ne savons pas comment ces mots auraient été classés si nous avions disposé d'une base de données orales[11]. Nous pouvons quand même constater que notre groupe de contrôle de locuteurs natifs utilise fréquemment le mot *ben* (76 occurrences). En examinant la liste K1 de ce même groupe, nous trouvons des mots présentant à peu près la même fréquence que *ben* : *beaucoup* (80), *bien* (83), *comme* (68). Selon nous, cela indique que tous ces mots ont une fréquence relativement élevée dans la langue parlée des locuteurs natifs de français. Il est donc probable qu'en appliquant la même méthode à partir d'une base de données orales, les mots *ben*, *beaucoup*, *bien* et *comme* entrent dans la même zone de fréquence. À ce sujet, nous avons consulté la

TABLEAU 4
La fréquence des mots *ben*, *ouais*, *beaucoup*, *bien* et *comme* dans le corpus Corpaix

Mot	Fréquence	Classement
ben	2 936	66
ouais	3 100	65
beaucoup	1 861	87
bien	3 708	55
comme	3 974	49

liste de fréquence conçue par J. Véronis à l'université de Provence[12]. Cette liste est basée sur un million de mots des productions orales de locuteurs natifs de français issues du corpus Corpaix. Il est à noter que la liste n'est pas lemmatisée : elle montre toutes les formes produites dans le corpus. Tableau 4 montre les fréquences et le classement dans ce corpus des mots évoqués ci-dessus.

Nous pouvons noter que tous les mots se placent parmi les 90 mots les plus fréquents. En effet, il semble que *ben* et *ouais*, classés dans la zone MHL dans notre analyse, se retrouvent dans la même gamme de fréquence que les autres mots, qui avaient été classés dans la zone K1 dans notre analyse. Par ailleurs, il est intéressant de noter que les AAS utilisent *ben* 48 fois et que les AAI ne l'utilisent qu'une fois. Cela indique que les apprenants les plus avancés se rapprochent de l'emploi natif, alors que les apprenants les moins avancés en sont assez éloignés. De plus, nous retrouvons des mots typiques de l'oral chez les AAS et les LN, tels que *rigolo*, *prof* et *truc*, qui sont quasi inexistants chez les AAI (sauf une occurrence du mot *prof*). Certes, l'emploi de ces mots n'est pas exhaustif, mais le fait que les AAS les utilisent renforce l'impression d'un emploi plus natif chez les apprenants les plus avancés.

Nous rappelons qu'Ovtcharov et coll. (2006, p. 121) se font les défenseurs de l'emploi d'une base de données écrites dans l'examen de la langue orale. Tout en admettant qu'il existe un problème lié aux spécificités de la langue écrite et de la langue orale, les chercheurs soutiennent qu'il est approprié d'effectuer des comparaisons tant qu'on emploie la même mesure sur le même genre d'entités. Nous partageons ce point de vue. Ainsi, dans notre étude, il est possible de comparer les résultats des groupes puisqu'on mesure les mêmes entités. Par contre, dans une perspective plus qualitative, nous ne sommes pas convaincues que les résultats donnent une image correcte de la richesse lexicale telle qu'elle se manifeste à l'oral. Tidball et Treffers-Daller (2008, p. 303) font aussi remarquer qu'il est probable que le taux relativement élevé de mots rares dans l'étude d'Ovtcharov et coll. (2006) s'explique par le fait que des mots typiques à l'oral ont été classés dans les zones de faible fréquence, puisque la division en zones de fréquences est fondée sur l'écrit.

Mot hors listes = mot avancé?

Évidemment, la liste MHL contient aussi des mots rares de la langue parlée. Elle comprend les mots particuliers, thématiques, professionnels ou monosémiques, qui, selon notre hypothèse, sont l'indice d'un vocabulaire avancé (ce genre de mots est d'ailleurs fréquent dans la

liste K3). En voici quelques exemples tirés de la production des AAS : *théoriciens*, *programmation*, *déconstructivisme*, *romanistique*, *structuralisme*. Toutefois, aucun de ces mots ne figure dans la liste de fréquence de Véronis, ce qui confirme qu'ils sont peu fréquents en français parlé.

Mot hors listes = mot commun?

Il est surprenant de trouver des mots comme *bière*, *pharmacie*, *électricité*, *métro*, *nettoyer*, *couteau* et *oreille* parmi les MHL. Ces mots nous semblent relativement courants dans l'usage quotidien et, pour cette raison, ils devraient normalement être accessibles à la plupart des locuteurs. Ils sont néanmoins classés parmi les mots les moins fréquents (Tidball et Treffers-Daller, 2008, p. 307, pour une remarque similaire). À ce propos, il est intéressant de noter ce que constatent Verlinde et Selva (2001, p. 595) après avoir effectué une délimitation de la liste de fréquence en ne comptant que les lemmes qui avaient une fréquence de plus de 100 : « It is surprising to see that this limited list contains a large number of words that are very common in spoken language : maman, papa, job, sympa, bosser, for example ». Les MHL sont censés indiquer un lexique riche et avancé. Si une grande partie des MHL relève du vocabulaire commun, on peut s'interroger sur l'hypothèse selon laquelle un vocabulaire riche signifie une proportion élevée de mots K3 et MHL liés à un usage professionnel ou thématique. Afin d'examiner la fréquence des mots du français parlé mentionnés au tout début de la présente section, nous avons de nouveau consulté la liste de fréquence réalisée par J. Véronis. Nous rappelons qu'elle est basée sur un corpus d'un million de mots provenant de la langue parlée. L'analyse de cette liste de fréquence a donné les résultats suivants :

TABLEAU 5
La fréquence des « mots communs » dans le corpus Corpaix

Mot	Fréquence	Classement
bière	21	2 474
pharmacie	33	1 755
électricité	25	2 162
métro	22	2 418
nettoyer	17	2 963
couteau	14	3 414
oreille	19	2 720

Le tableau montre que tous ces mots sauf *couteau* figurent parmi les 3 000 mots les plus fréquents de la langue orale des locuteurs natifs de français. La liste de Véronis étant composée de formes et non pas de lemmes ou de familles de mots, elle contient moins d'unités que la liste de fréquence de Verlinde et Selva (2001). Autrement dit, les 3 000 mots les plus fréquents de la liste de Véronis sont 3 000 formes différentes. Dans la liste de Verlinde et Selva, les trois zones de fréquence K1, K2 et K3 comprennent les 3 000 familles de mots les plus fréquentes, chacune étant composée de plusieurs « membres ». Ainsi, les trois zones de fréquence comptent un plus grand nombre de formes différentes. Il semble que le type de mots évoqués dans cette section constitue une catégorie particulière. En fait, Gougenheim, Michéa, Rivenc et Sauvageot (1967, p. 145) proposent la catégorie *vocabulaire disponible* pour les mots « d'une fréquence faible et peu stable, qui sont toutefois des mots usuels et utiles ». De même, en ce qui concerne l'italien, De Mauro (2005) propose la catégorie de *parole di alta disponibilità* pour les mots connus de tous les locuteurs natifs, mais peu souvent utilisés.

Ces constatations témoignent de la difficulté de l'interprétation des résultats, puisqu'elles remettent en question l'hypothèse selon laquelle une proportion de mots rares est l'indice d'un vocabulaire avancé. Witalisz (2007), qui a testé la méthode PFL sur l'anglais écrit d'apprenants polonais, met en évidence des problèmes méthodologiques similaires. À propos du classement des mots dans les zones de fréquence de la version anglaise du PFL, Witalisz (2007, p. 110) affirme que : « the classification simply did not coincide with the experienced teacher's intuition as to which vocabulary is more advanced ». Afin de vérifier cette hypothèse, Witalisz (2007) a comparé le classement des mots de ses données dans le PFL avec celle du *Collins COBUILD English Dictionary*. Ainsi, l'auteur a pu noter des divergences dans la mesure où des mots hors listes selon PFL sont classés parmi les 2 000 les plus fréquents du dictionnaire.

Deux cas extrêmes

Nous avons examiné de plus près les deux apprenants au stade avancé supérieur, qui ont les taux les plus élevés de mots K3 + MHL, soit 7,54 % et 6,62 %, respectivement (Figure 1). Pourquoi ces apprenants se distinguent-ils des autres membres de leur groupe (AAS) et pourquoi présentent-ils même des taux plus élevés que ceux des locuteurs natifs? Quel genre de mots ces apprenants produisent-ils? En évaluant d'abord l'apprenante (Matilda) qui présente la proportion la plus

élevée dans les zones K3 + MHL, nous constatons que les mots en K3 sont relativement peu fréquents (1,08 % du total). Par contre, la proportion des mots hors listes s'élève à 6,46 % (le taux le plus élevé chez les LN est de 4,47 %). Il y a chez Matilda 67 types différents et 161 occurrences. La plupart des types ne sont répétés qu'une ou deux fois. Deux types se distinguent des autres : *ben* (46 occurrences) et *ouais* (23 occurrences). Nous pourrions alors nous demander dans quelle mesure cette apprenante a un vocabulaire riche, et cela pour deux raisons. La première est liée de nouveau à la nature des mots (typiques du discours oral, discussion plus haut). La deuxième est liée au problème du comptage. L'unité de comparaison, le nombre d'occurrences, favorise les informants qui répètent un certain type à plusieurs reprises. Par contre, il n'est pas apparent qu'un nombre élevé d'occurrences soit l'indice d'un vocabulaire avancé. Nous pourrions alors nous demander si, dans un examen plus approfondi des mots rares, il serait plus juste de comparer le nombre des types au lieu du nombre d'occurrences. Ce genre d'analyse ne fait pas partie du cadre de la présente étude, mais un premier calcul montre que trois des locuteurs natifs produisent plus de types de MHL (97, 96 et 85) que Matilda (67), mais que Matilda produit plus d'occurrences MHL (161) que ces trois locuteurs natifs (137, 153 et 124). Ces chiffres indiquent que ces locuteurs natifs se servent d'un plus grand nombre de *types* rares que Matilda. En conséquence, le vocabulaire de ces locuteurs natifs devrait peut-être être qualifié de plus riche et plus varié que celui de Matilda. Voir aussi Horst et Collins (2006) pour le comptage des types dans le cadre du PFL.

L'autre apprenant (Knut) présente quant à lui une proportion élevée de mots K3 + MHL. Chez lui aussi, les mots K3 sont peu fréquents (0,65 % du total), tandis que les MHL représentent 5,97 % du total. En effet, le mot le plus fréquent chez Knut est *ouais* (21 occurrences). Contrairement à Matilda, il n'emploie jamais *ben*. Un autre mot de fréquence élevée parmi les MHL est *manuscrit* (17 occurrences), qui relève du domaine de recherche de cet apprenant et qui peut être considéré comme avancé, dans la mesure où il appartient à un vocabulaire spécialisé.

Conclusions et perspectives

Dans cette étude, nous avons analysé la richesse lexicale de la production orale de l'apprenant avancé de français L2. Nous avons utilisé la méthode Profil de fréquence lexicale, qui permet de diviser une production en quatre zones de fréquence. De manière générale, nous avons pu constater

des différences entre les profils lexicaux des apprenants les moins avancés et de ceux des locuteurs natifs. De plus, il semble que les profils lexicaux des apprenants les plus avancés et des locuteurs natifs soient similaires. Globalement, ces résultats semblent refléter le niveau de compétence linguistique des apprenants. Pourtant, les tests statistiques ne nous ont pas permis de constater une différence significative entre les groupes d'apprenants. Ainsi, nous n'avons pas pu confirmer l'hypothèse selon laquelle la méthode permettrait de trancher entre deux niveaux de compétence linguistique. Par contre, l'absence de différences significatives entre les apprenants les plus avancés et les locuteurs natifs est conforme à notre hypothèse, et pourrait indiquer que ces apprenants se rapprochent du niveau natif en ce qui a trait à la richesse lexicale.

Notre étude a montré à la fois les points forts et les limites de la méthode PFL. Si la méthode semble pouvoir refléter du moins partiellement les niveaux de compétence linguistique des informants, elle n'est toutefois pas entièrement fiable comme outil d'analyse de la langue parlée. Notre analyse des mots classés dans la liste MHL a fait apparaître la nature diverse de ces mots. D'une part, nous y avons repéré des mots qui nous semblent « avancés », et qui sont donc classés à juste titre dans cette zone. D'autre part, nous avons pu constater que la liste MHL contient des mots 1) qui sont particulièrement fréquents dans la langue parlée, et 2) qui semblent relativement courants dans l'usage quotidien du français parlé, et auxquels la plupart des locuteurs devraient avoir accès. Afin de vérifier si ces mots sont fréquents en français parlé, nous nous sommes référée à une liste de fréquence du français oral, établie par J. Véronis. Cette liste nous a permis de confirmer notre hypothèse selon laquelle certains mots de fréquence élevée dans la langue parlée ne le sont pas dans la langue écrite, ce qui a déjà notamment été signalé par McCarthy (1998) et Campione et coll. (2005). C'est pourquoi une base de données écrites nous semble mal adaptée à l'analyse de la langue orale. Dans la suite de nos recherches, nous avons l'intention de garder la proposition de la division des mots en zones de fréquence, mais nous utiliserons une liste de fréquence basée sur la production orale de locuteurs natifs de français (Lindqvist et coll., soumis). Ainsi, nous disposerons de données plus comparables, dans lesquelles la division en zones de fréquence reflète l'usage lexical dans la langue parlée de locuteurs natifs de français. Il faudrait aussi ajouter plus de données afin de confirmer les résultats dans une documentation plus riche. Cela étant dit, nous sommes d'avis que l'analyse de profils lexicaux constitue un bon point de départ de l'analyse de la richesse lexicale, dans la mesure où elle permet aussi une analyse plus qualitative, basée sur la classification des mots dans différentes catégories. Par contre, cela n'exclut pas

que d'autres mesures de la richesse lexicale puissent apporter un complément à l'analyse du PFL.

La correspondence devrait être adressée à **Christina Lindqvist**, Département de français, d'italien et de langues classiques, Université de Stockholm, SE-106 91 Stockholm, Suède. Courriel : christina.lindqvist@fraita.su.se

Remerciements

Cet article a été élaboré dans le cadre du programme de recherche *High-level Proficiency in Second Language Use* à l'Université de Stockholm, et financé par *The Bank of Sweden Tercentenary Foundation*. Nous tenons à remercier nos collègues du sous-projet *Aspects of the advanced learner's lexicon*, Camilla Bardel et Anna Gudmundson, pour leur lecture attentive d'une version antérieure de cet article. Nous remercions aussi les trois évaluateurs anonymes de la RCLV pour leurs remarques pertinentes. Enfin, nous souhaitons remercier Hugues Engel pour ses suggestions visant les améliorations de la langue.

Notes

1 Nous remercions un évaluateur anonyme de cette remarque en ce qui a trait à la longueur des textes.
2 Voir www.lextutor.ca
3 Le but de l'entrevue, qui faisait partie d'un examen, était de placer les répondants à trois niveaux d'acquisition préétablis : novice, intermédiaire et avancé.
4 Après suggestion par un évaluateur anonyme, nous avons présenté à Vocabprofil des textes de longueurs différentes, mais issus d'une même production d'apprenant. En effet, le profil lexical n'a guère changé, que le texte soit de 200, 300, 400, 500, 600, 700, 800, 900 ou 1 000 mots. Le rôle de la longueur du texte fait l'objet d'un examen plus approfondi dans Bardel et Lindqvist, 2009).
5 Ces mots ne sont pas jugés faire partie du vocabulaire. Omettre les pauses remplies s'est révélé un choix difficile. Nous avons choisi de garder les mots considérés par le Trésor de la langue française informatisé (accessible sur http://atilf.atilf.fr/) comme des interjections, par exemple, *ouf*, *ah*, *aha*.
6 Noter que l'analyse du PFL ne tient pas compte du contexte. Nous ne pouvons donc pas savoir si les mots sont employés correctement du point de vue sémantique ou pragmatique.

7 Voir Laufer (1995), qui propose de réunir les zones de faible fréquence en une seule zone, *Beyond 2000*, et de l'opposer aux zones de haute fréquence, créant ainsi un « profil condensé ».
8 Nous parlons ici des valeurs moyennes au niveau des groupes. Il importe de mentionner que, dans notre étude, la valeur la plus élevée dans les zones K3 + MHL au niveau de l'individu est de 7,56 % (Figure 1).
9 Rappelons que Laufer (1995) émet l'hypothèse que le profil est stable entre 200 et 400 mots pour l'anglais. Toutefois, la question de la longueur des textes n'a pas encore été résolue.
10 Un évaluateur anonyme soulève la question des mots apparentés. Il est vrai que la L1 pourrait avoir une incidence sur la proportion de mots rares. En fait, les apprenants anglophones dans l'étude d'Ovtcharov, Cobb et Halter (2006) ont recours à un nombre important de mots apparentés anglais-français, qui pourraient faire augmenter la proportion de mots rares. Les suédois, pour leur part, n'ont peut-être pas l'occasion d'intégrer des mots apparentés au même degré, même s'ils possèdent tous une connaissance de l'anglais. La question des mots apparentés fait l'objet d'un examen plus minutieux dans Lindqvist, Bardel et Gudmundson (soumis) et Bardel et Lindqvist (2009). Voir aussi Horst et Collins (2006) et Meara, Lightbown et Halter (1997).
11 Un évaluateur anonyme nous signale que des mots comme *oh*, *yeah*, *okay* et *hello* sont inclus dans les 1 000 premières familles de mots dans une nouvelle version expérimentale du Vocabprofile anglais : BNC-20. Ainsi, il est possible que la version française soit elle aussi révisée, de sorte que ce genre de mots appartienne à la zone K1.
12 La liste est accessible sur http://sites.univ-provence.fr/veronis/data/freq-oral.txt.

Références

Bardel, C. et Lindqvist, C. (2009, septembre). Quantitative and qualitative aspects of the lexical profiles of Swedish learners' spoken French and Italian L2. Communication présentée à *Eurosla 19*, Cork (Irlande).

Bartning, I. et Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies*, *14*, 1–19.

Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*. Paris: Ophrys.

Bulté, B., Housen, A., Pierrard, M. et Van Daele, S. (2008). Investigating lexical proficiency development over time – the case of Dutch-speaking learners of French in Brussels. *Journal of French Language Studies*, *18*, 277–298.

Campione, E., Véronis, J. et Deulofeu, J. (2005). The French corpus. Dans E. Cresti et M. Moneglia (dir.), *C-ORAL-ROM, Integrated reference corpora for spoken Romance languages* (p. 111–133). Amsterdam : John Benjamins.

Cobb, T. et Horst, M. (2004). Is there room for an academic wordlist in French?Dans P. Bogaards et B. Laufer (dir.), *Vocabulary in a second language: Selection, acquisition, and testing* (p. 25–38). Amsterdam : John Benjamins.

Daller, H., Van Hout, R. et Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24*(2), 197–222.

Daller, H. et Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students: A comparison of different measures. Dans H. Daller, J. Milton et J. Treffers-Daller (dir.), *Modelling and assessing vocabulary knowledge* (p. 150–164). Cambridge : Cambridge University Press.

De Mauro, T. (2005). *La fabbrica delle parole. Il lessico e problemi di lessicologia*. Torino : UTET.

Goodfellow, R., Lamy, M.-N. et Jones, G. (2002). Assessing learners' texts using the Lexical Frequency Profile. *ReCall*, *14*(1), 133–145.

Gougenheim, G., Michéa, R., Rivenc, P. et Sauvageot, A. (1967). *L'élaboration du français fondamental (1er degré)*. Paris : Didier.

Granfeldt, J. (2006). Profils grammaticaux et lexicaux – à la recherche d'un rapport. Dans *Actes du XVIe Congrès des romanistes scandinaves*, Université Roskilde, Danemark. Téléchargé à http://www.ruc.dk/cuid/publikationer/publikationer/XVI-SRK-Pub/TVI/TVI15-Granfeldt/

Horst, M. et Collins, L. (2006). From *faible* to strong: How does their vocabulary grow? *La Revue canadienne des langues vivantes*, *63*, 83–106.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84.

Laufer, B. (1995). Beyond 2000: A measure of productive lexicon in a second language. Dans L. Eubank, L. Selinker et M. Sharwood Smith (dir.), *The current state of interlanguage* (p. 265–272). Amsterdam : John Benjamins.

Laufer, B. et Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322.

Le Trésor de la Langue Française informatisé. Téléchargé à http://atilf.atilf.fr/.

Lindqvist, C., Bardel, C. et Gudmundson, A. Lexical richness in the advanced learner's oral production of French and Italian L2. (soumis)

Malvern, D., Richards, B., Chipere, N. et Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. New York : Palgrave Macmillan.

McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge : Cambridge University Press.

McCarthy, P. et Jarvis, S. (2007). *vocD*: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488.

Meara, P., Lightbown, P.M. et Halter, R.H. (1997). Classrooms as lexical environments. *Language Teaching Research*, *1*(1), 28–47.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge : Cambridge University Press.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *La Revue canadienne des langues vivantes*, *63*(1), 59–82.

Ovtcharov, V., Cobb, T. et Halter, R. (2006). La richesse lexicale des productions orales: mesure fiable du niveau de compétence langagière. *La Revue canadienne des langues vivantes*, *63*(1), 107–125.

Sanell, A. (2007). *Parcours acquisitionnel de la négation et de quelques particules de portée en français L2*. Thèse de doctorat inédite, Université de Stockholm.

Tidball, F. et Treffers-Daller, J. (2007). Exploring measures of vocabulary richness in semi-spontaneous French speech: A quest for the Holy Grail? Dans H. Daller, J. Milton et J. Treffers-Daller (dir.), *Modelling and assessing vocabulary knowledge* (p. 133–149). Cambridge : Cambridge University Press.

Tidball, F. et Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: What frequency lists and teacher judgements can tell us about basic and advanced words. *Journal of French Language Studies*, *18*, 299–313.

Verlinde, S. et Selva, T. (2001). Corpus-based versus intuition-based lexicography: Defining a word list for a French learners' dictionary. Dans P. Rayson, A. Wilson, T. McEnery, A. Hardy et S. Khoja (dir.), *Actes de la Corpus Linguistics 2001 conference* (*Technical Papers*, 13, 594–598). Téléchargé le 21 janvier 2008 à http://www.kuleuven.be/grelep/publicat/verlinde.pdf

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. Dans P. Bogaards et B. Laufer (dir.), *Vocabulary in a second language: Selection, acquisition, and testing* (p. 173–189). Amsterdam : John Benjamins.

Véronis, J. Fréquences des mots en français parlé. Téléchargé le 17 décembre 2007 à http://sites.univ-provence.fr/veronis/data/freq-oral.txt

Witalisz, E. (2007). Vocabulary assessment in writing: lexical statistics. Dans Z. Lengyel et J. Navracsics (dir.), *Second language lexical processes : Applied linguistic and psycholinguistic perspectives* (p. 101–116). Clevedon : Multilingual Matters.