PROJECT MUSE®

Factors Predicting Low-Intermediate French Learners' Vocabulary Use in Speaking Tasks

Ann-Sophie Noreillie, Piet Desmet, Elke Peters

The Canadian Modern Language Review / La revue canadienne des langues
vivantes, Advance Online, (Article)

Published by University of Toronto Press

This is a preprint article. When the final version of this article launches,
this URL will be automatically redirected.

➔ For additional information about this preprint article

<https://muse.jhu.edu/article/762114/summary>

Factors Predicting Low-Intermediate French Learners' Vocabulary Use in Speaking Tasks

Ann-Sophie Noreillie¹, Piet Desmet², and Elke Peters¹

Résumé : Les auteurs explorent la production orale en français d'apprenants natifs et non natifs dans deux dialogues. L'étude vise plus précisément trois objectifs. Premièrement, celui d'étudier la relation entre la taille du vocabulaire réceptif et productif d'apprenants de français d'une part et le vocabulaire utilisé dans deux dialogues d'autre part. Deuxièmement, les auteurs veulent comparer le vocabulaire produit (c'est-à-dire le nombre de mots utilisés, le profil lexical de fréquence et la diversité lexicale mesurée à l'aide de D et HD-D) par des locuteurs non natifs avec celui des locuteurs natifs. Troisièmement, les auteurs veulent déterminer lesquelles des mesures mentionnées ci-dessus permettent de prédire l'évaluation holistique des connaissances lexicales d'apprenants de français par des experts. Cinquante apprenants de français de 17 ans ont rempli un test de vocabulaire réceptif et productif et ont participé à deux dialogues. Vingt-sept francophones du même âge ont produit les mêmes dialogues. Les résultats indiquent qu'il existe une corrélation positive et modérée entre le vocabulaire réceptif et les scores holistiques, mais pas entre le vocabulaire productif et les scores holistiques. Qui plus est, lorsqu'ils s'expriment, les locuteurs natifs produisent plus de mots et moins de mots à usage fréquent. La diversité lexicale de leur production est également plus élevée que celle des apprenants de français. Enfin, il semble qu'une combinaison de la diversité lexicale (HD-D) et du nombre de *types* utilisés par les locuteurs ainsi que la fréquence de ces mots expliquent une grande partie de la variance dans les scores holistiques. Les résultats de la présente étude jettent un nouvel éclairage sur le rôle de l'emploi du vocabulaire dans la production orale de locuteurs natifs et non natifs dans une langue autre que l'anglais, à savoir le français, et pour une population autre que des étudiants universitaires, c'est-à-dire des apprenants pré-intermédiaires.

Mots clés : connaissances lexicales réceptives et productives, diversité lexicale, emploi du vocabulaire dans la production orale, locuteurs natifs et non natifs de français, profil lexical de fréquence, scores lexicaux holistiques

Abstract: The present study aims to explore both native French speakers and French as a foreign language learners' spoken output in two dialogic speaking tasks. More specifically, the aim of this study was threefold. First, the researchers focused on the relationship between L2 learners' receptive and productive vocabulary knowledge and their vocabulary use in two dialogic speaking tasks. Second, the study set out to compare L2 speakers' vocabulary use (i.e., number of words produced, lexical frequency profile, and lexical diversity as measured by D and HD-D) with L1 speakers' vocabulary use in the same speaking tasks. Third, the researchers wanted to determine which of the aforementioned vocabulary measures could predict experts' holistic

¹ KU Leuven, Antwerp, Belgium

² KU Leuven Campus Kulak, Kortrijk, West-Flanders, Belgium

Correspondence should be addressed to Ann-Sophie Noreillie, Faculty of Arts, KU Leuven, Sint-Andriesstraat 2, Antwerp, Belgium; email: annsophie.noreillie@kuleuven.be.

ratings of lexical proficiency. Fifty 17-year-old L2 learners took a receptive and a productive vocabulary test and performed two dialogic speaking tasks, and 27 French L1 speakers performed the same two dialogic speaking tasks. The results indicate that there is a positive, moderate correlation between receptive vocabulary knowledge and holistic ratings but not between productive vocabulary knowledge and the holistic rating scores. Moreover, L1 speakers produce more tokens, types, and lemmas and fewer high-frequency words when they speak. The lexical diversity of their output is also higher than that of L2 speakers. Finally, it seems that lexical diversity (i.e., HD-D), combined with the number of types that speakers use and the frequency of those words, explains a large part of the variance in holistic rating scores. The findings of the present study shed more light on the role of L1 and L2 speakers' vocabulary use in spoken output for another language than English (i.e., French) and for another population than university students (i.e., low-intermediate learners).

Keywords: French L1 and L2 speakers, holistic lexical ratings, lexical diversity, lexical frequency profile, receptive and productive vocabulary knowledge, vocabulary use in speaking

While previous research has investigated the relationship between second language (L2) vocabulary knowledge and reading, listening, and writing (e.g., Crossley, Salsbury, McNamara, & Jarvis, 2011b; Laufer & Ravenhorst-Kalovski, 2010; Stæhr, 2008, 2009), less is known about the relationship between vocabulary knowledge and speaking (see de Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012, for an exception). Although the number of studies on L2 speech is definitely growing, most studies have investigated English L2 learners performing monologic tasks (e.g., de Jong et al., 2012; Uchiyama & Clenton, 2018). There is now some evidence that learners with a larger vocabulary tend to perform better in speaking tasks and that speech that contains more words, more low-frequency words, and a larger variety of words is rated more highly (e.g., Crossley & McNamara, 2012). Yet fewer studies have focused on learners of French (for a few exceptions, see Lindqvist, 2010; Malvern and Richards, 2002; Treffers-Daller, 2013) and on dialogic tasks (Lorenzo-Dus & Meara, 2005), which makes it difficult to generalize the current research findings.

The aim of the present study is to extend previous research into speaking by focusing on non-native French speakers' vocabulary use in dialogic tasks in three Flemish secondary school classrooms. Specifically, we wanted to determine which lexical indices could predict experts' holistic ratings of learners' vocabulary use. Additionally, we aimed to compare learners' vocabulary use with that of native speakers. Our final aim was to explore the relationship between learners' vocabulary knowledge and their vocabulary use during speaking.

Background

Vocabulary use in learners' output can be studied in different ways, for example, by counting the total words used or by analyzing the frequency of the words used. Before we discuss such studies, it might be useful to define the terms *token*, *type*, and *lemma*. The number of tokens in a text refers to all the words in a text, whereas the number of types refers to the number of unique words in a text. A lemma refers to a headword and its inflected and reduced forms, which usually belong to the same part of speech (see also Kyle, 2020). We will start with studies addressing the role of the number frequency of the words used in learner output before moving on to studies that have focused on lexical diversity, differences in output between L1 and L2 speakers, and finally the relationship between learners' knowledge of words and their use of words.

Relationship between number of words and experts' ratings of speaking tasks

Studies investigating the relationship between speaking and vocabulary use by English as a foreign language (EFL) learners have shown that there is a positive relationship between the

number of words learners used and the ratings they received (Crossley & McNamara, 2013; Iwashita, Brown, McNamara, & O'Hagan, 2008; Lorenzo-Dus & Meara, 2005). A study by Iwashita et al. (2008) investigated the relationship between EFL learners' spoken English in two monologic tasks and holistic scores. The results indicated that learners of distinct proficiency levels differ in the number of words (i.e., tokens) and the number of unique words (i.e., types) they use when their level is higher and experts' ratings are better. Similar results were reported in a study by Lorenzo-Dus and Meara (2005), who looked at the relationship between learners' vocabulary use and vocabulary grades. It is one of the few studies that used a dialogic task and collected data with learners in secondary schools. Twenty-nine British secondary school students took Spanish oral proficiency tests that consisted of two parts: a general conversation with the examiner/rater about their personal interests, studies, and plans for the future and a conversation about a prepared topic. The number of types was shown to be a distinctive factor for the grade; that is, the higher the number of types used, the higher the grade awarded. Further, Crossley and McNamara (2013) investigated the potential of the use of automated indices for predicting human judgments of speaking proficiency with a corpus of TOEFL speaking data, which consisted of a monologic speaking task in which speakers had to talk about a familiar topic and their opinions and experiences. The number of types used by speakers was predictive of the ratings given by experts, with speakers producing a larger number of unique words being judged to have greater speaking proficiency.

Finally, research has also shown that the number of tokens and types used is one important difference between L1 and L2 speakers' vocabulary use in speaking tasks (Treffers-Daller, 2013; Vermeer, 2000).

Relationship between lexical frequency profile and experts' ratings of speaking tasks

In addition to counting the number of (unique) words in a text, learners' vocabulary use can be measured by investigating the frequency of the words that are used. Lexical sophistication is the term used to refer to the proportion of low-frequency words in learners' output (Kyle, 2020). This can, for instance, be measured using the lexical frequency profile (LFP; Laufer & Nation, 1995), a tool that was originally developed for English written production and that measures the percentage of high- and low-frequency words in the output. The principle underlying the LFP is that the more low-frequency words a learner uses, the higher their proficiency is, since high-frequency words tend to be learned before low-frequency words. A more recent tool is TAALES (Kyle & Crossley, 2015), which automatically computes over 400 indices of lexical sophistication. TAALES uses an average frequency score to determine lexical sophistication. Except for one study (Uchihara & Clenton, 2018), most studies seem to provide evidence that speech containing more low-frequency words is rated as more lexically sophisticated.

Crossley and colleagues (Crossley & McNamara, 2013; Kyle & Crossley, 2015) found that frequency indices could account for part of the variance in holistic judgments of speaking proficiency. Their results showed that there was a positive relation between well-rated samples and the use of more academic words and more low-frequency words. However, Uchihara and Clenton (2018) did not find a significant correlation between the number of high- or low-frequency words used by ESL learners and experts' ratings in their study with ESL learners who performed a picture description task. Both the LFP and TAALES were used to analyze the frequency of the words learners produced. They concluded that learners who were lexically proficient according to experts' ratings "did not necessarily produce lexically sophisticated L2 words in spontaneous speech" (Uchihara & Clenton, 2018, p. 10). It

is difficult to pinpoint exactly why the research findings differ, as there were differences in methodology, namely, the rating scales that were used, the different lexical measures, and the different topics of the speaking tasks.

Two studies are important for the present study because they both focus on French and use dialogic tasks. Ovtcharov, Cobb, and Halter (2006) examined the vocabulary use of advanced learners of French in speaking tasks. They found that more advanced speakers used more mid- and low-frequency words. Similarly, Lindqvist (2010), who analyzed the vocabulary use of advanced learners of French in a semi-guided interview and compared it with native speakers, showed that higher-proficiency learners used more low-frequency words. Further, the most advanced students had a lexical profile very similar to that of native speakers. Finally, it should be noted that learners' vocabulary consisted mainly of words belonging to the 1,000 most frequent words of French. Lindqvist explains these findings in terms of the topic, as the learners were interviewed about family, hobbies, and studies. However, it might also be due to the fact that the number of high-frequency words is usually higher in spoken language than in written language (Nation, 2006, p. 79). Although most research points to the relationship between frequency and lexical proficiency, it remains unclear whether these findings would hold for language learning with a lower proficiency level.

Relationship between lexical diversity and experts' ratings of speaking tasks

In addition to frequency, learners' production has also been analyzed in terms of lexical diversity (LD). In the present study, LD will be defined according to Kyle's (2020, p. 458) definition: "the variety of lexical items in a text ... Because more proficient speakers and writers have more lexical items at their disposal, they are able to use a wide variety of lexical items to accomplish a task, while less proficient speakers and writers will tend to repeat a smaller number of lexical items."¹ The variety of the vocabulary that is used in spoken or written output is in contrast with the potential vocabulary a learner might have available but is not using. There are different ways to operationalize LD (see Kyle, 2020, for an overview); examples include Type-Token Ratio (TTR) or its transformations – for instance, Root TTR (Guiraud, 1960), Corrected TTR (Carroll, 1964), Advanced TTR (Daller, van Hout, & Treffers-Daller, 2003)), D (Malvern, Richards, Chipere, & Durán, 2004), the hypergeometric distribution HD-D (McCarthy & Jarvis, 2007; Wu, 1993), and the measure of textual and lexical diversity MTLTLD (McCarthy, 2005; McCarthy & Jarvis, 2010). Research has shown, however, that the TTR measure and its transformations are not appropriate for calculating LD because they are dependent on text length. Indeed, as the number of tokens in a text steadily increases, the relative number of types decreases, since tokens tend to be repeated after a relatively short amount of text. Therefore, researchers have looked for more robust approaches to testing LD, such as HD-D, D, and MTLTLD. These have also been validated for English by McCarthy and Jarvis (2010) and for French by Treffers-Daller (2013). The assumption behind LD indices is that more diverse texts are indicative of larger and more proficient lexicons and thus of a higher proficiency level.

The studies that have investigated the relationship between LD and experts' ratings of learners' oral output (e.g., Appel, Trofimovich, Saito, Isaacs, & Webb, 2019; Crossley & McNamara, 2013; Crossley, Salsbury, & McNamara, 2015; Crossley, Salsbury, McNamara, & Jarvis, 2011a; Iwashita, 2010; Lorenzo-Dus & Meara, 2005; Lu, 2012; Malvern & Richards, 2002; Yu, 2010) display conflicting evidence. Most studies have focused on English, and those found a positive relationship between lexical diversity and ratings. For instance, Crossley and his colleagues (Crossley et al., 2011a, 2015; Crossley & McNamara, 2013)

found that LD, measured as D or MTL D, explained a significant percentage of the variance in human ratings of speaking, which means that speakers who produce speech that is more lexically diverse is better rated. These findings were corroborated by Lu (2012) and Yu (2010). Looking at the comprehensibility and nativeness of L2 English speech samples, Appel et al. (2019) found that lexical diversity was a key factor. Finally, Iwashita (2010) showed that lexical diversity could distinguish between low- and high-proficiency learners. In other words, the more diverse the output, the more proficient the learner is.

In spite of the positive evidence for LD as a predictor of speaking, two studies focusing respectively on French and Spanish did not find statistically significant correlations. A study by Malvern and Richards (2002) that analyzed oral proficiency with British learners of French did not show statistically significant correlations between LD (measured by the diversity index D) and rated range of vocabulary. However, according to the authors, the absence of a correlation might be due to teachers' difficulty in assessing LD independently of other factors such as fluency, structural complexity, and pronunciation. In their study with British secondary school students learning Spanish, Lorenzo-Dus and Meara (2005) tested whether learners' vocabulary output and grade ratings were related. They found that the LD (D) of learners' output did not discriminate between the different groups that were created after the experts' ratings were attributed. These last two studies indicate that there might be a discrepancy between EFL and other foreign languages. It seems that the findings cannot be generalized from one language to the other. Consequently, more research is needed into other languages than English. The present study aims to address this gap by focusing on French language learning among Dutch speakers.

Relationship between learners' receptive and productive vocabulary knowledge and experts' ratings of speaking tasks

Vocabulary knowledge can be both receptive and productive. Receptive vocabulary knowledge, being necessary for listening and reading, refers to knowing the meaning of words, whereas productive vocabulary knowledge refers to knowing the form for a given meaning. Productive vocabulary knowledge is needed for speaking and writing (Nation, 2020). Previous research has shown that productive vocabulary knowledge (sentence completion, L1–L2 translation) is a good indicator of learners' speaking proficiency (e.g., de Jong et al., 2012; Koizumi & In'nami, 2013) and that, according to expert raters, vocabulary also constitutes one of the salient features for assessing speaking (e.g., Brown, Iwashita, & McNamara, 2005).

A study by de Jong et al. (2012), in which 181 L2 learners of Dutch and 54 native speakers of Dutch performed eight monologic speaking tasks and took nine tests of linguistic skills, indicated that there was a strong correlation between vocabulary knowledge (as measured by productive fill-in-the-gaps tests) and overall speaking proficiency ($r = 0.79$). Further, in a study with Japanese EFL students, Uchihara and Saito (2019) showed that there is a correlation between productive vocabulary (as measured using the Lex30 word association task) and the rated output of a picture description task. It should be noted that there was only a moderate correlation between the raw results on the vocabulary test and fluency measures, not between the Lex30 results and the global components of L2 speech that were measured (comprehensibility and accentedness). However, no correlations were found between the percentage scores on the vocabulary test and any of the oral measures.

Research has also focused on the relationship between receptive vocabulary knowledge and speaking. For instance, Uchihara and Clenton (2018) investigated the relationship

between receptive vocabulary knowledge and experts' ratings of EFL speakers' output (in a monologic picture description task). Learners' output was rated on its vocabulary component by means of the IELTS speaking descriptors. The results showed that the receptive vocabulary test correlated positively ($r = 0.55$) with the vocabulary ratings.

Research questions

Although a growing number of studies have explored the relationship between vocabulary use and speaking, research has revealed mixed findings regarding the relationship between lexical indices and lexical proficiency, on the one hand, and learners' vocabulary knowledge and lexical proficiency, on the other. Furthermore, little is yet known about learners' vocabulary use in languages other than English, and most studies have collected data from university students, which makes it difficult to generalize findings to other populations. Finally, most research has used monologic speaking tasks (see Lindqvist, 2010, for an exception), so it remains to be seen whether other task types result in similar outcomes. The present study aims to fill these gaps by focusing on the vocabulary use of low-intermediate learners of French in two dialogic tasks and by comparing L1 speakers' output to that of L2 speakers.

We addressed the following research questions:

1. Which lexical factors predict experts' holistic rating scores of L1 and L2 speakers' oral output in two semi-structured dialogic speaking tasks?
2. How does L1 and low-intermediate L2 speakers' output in two semi-structured dialogic tasks differ with respect to (a) the number of tokens, types, and lemmas, (b) the lexical frequency profile, and (c) lexical diversity?
3. What is the relationship between low-intermediate L2 learners' receptive and productive vocabulary knowledge and raters' judgement of holistic ratings of their spoken output?

Method

Participants

Data in this study were collected from French L1 and L2 speakers. The latter were 51 Flemish low-intermediate learners of French in two secondary schools in Flanders. It has to be stressed that, even though French is an official language in Belgium, according to the principle of territoriality (van der Jeught, 2017) the official language in Flanders is Dutch. Most of the participants in this study speak Dutch at home and do not encounter French a lot in their daily lives. The situation would probably be different for learners living in the border regions with Wallonia or France, where they would more often encounter French. French could therefore, in that sense, be considered a foreign language.² All the participants in this study were supposed to have reached the B1 level of the Common European Framework of Reference (CEFR) for speaking as well as for the other skills at the end of the year, that is, after eight years of French instruction. They were in their last year of secondary education (mean age = 17) and had three ($n = 22$) or four ($n = 29$) fifty-minute classes of French a week, depending on their curriculum. All but two learners had Dutch as their L1; one learner had Arabic as her mother tongue, and there was one Dutch/French bilingual learner. The data of the bilingual learner were removed from the dataset. As a result, the data of 50 learners were analyzed. In addition, data from 27 native speakers of French were collected in a secondary school in Wallonia, the French-speaking part of Belgium. Like the non-native learners, these participants were also in their last year of secondary education.

Since both the L2 and the L1 speakers produced two speaking tasks, our spoken corpus contained a total of 100 non-native speech samples and 54 native speech samples ($N = 154$).

Instruments

The data-collection instruments consisted of both a receptive and a productive vocabulary test and two semi-structured dialogic speaking tasks. The vocabulary tests were administered only to the L2 speakers. Learners all started with the productive test, which was followed by the receptive test. The vocabulary tests were administered on different days from the speaking tasks, but within the same week.

Receptive vocabulary knowledge test

In order to measure L2 learners' receptive vocabulary knowledge, a French frequency-based bilingual multiple-choice test was created and administered. Since the results on the monolingual receptive VocabLab test (Peters, Velghe, & Van Rompaey, 2019) used in a study by Noreillie, Kestemont, Heylen, Desmet, and Peters (2018) had shown that the scores of L2 learners of French (aged 14, 16, and 18) were very low on the 5K band, a vocabulary test assessing only the 4,000 most frequent words in French was created (Noreillie, 2019). The *Test de la taille du vocabulaire* (Batista & Horst, 2016) was not considered appropriate for our research purposes, given that it tests more lower-frequency words (the 2K, 3K, 5K, and 10K sections). Consequently, it would have been more difficult to differentiate between the participants.

Lonsdale and Le Bras's (2009) frequency list was used to develop the test. The test targets words from the 1,000 (1K), 2,000 (2K), 3,000 (3K) and 4,000 (4K) most frequent words in French. The 120 items on the test, consisting of 30 items per 1,000-word frequency band, are presented in isolation and are offered in a written multiple-choice format. Learners receive four translations in Dutch (i.e., the learners' L1), among which the correct answer and three distractors are found; there is also an "I don't know" option. The test had good internal consistency (Cronbach's $\alpha = 0.91$, $n = 50$).

An example of a multiple-choice question follows, with English glosses for reference:

lutte (f)
gevecht [battle]
verdriet [grief]
moord [murder]
uitdaging [challenge]
ik weet het niet [I don't know]

Productive vocabulary knowledge test

Learners' productive vocabulary knowledge was measured by means of a bilingual frequency-based vocabulary test developed according to the format of the English Productive Levels Test (Laufer & Nation, 1999). This test provides an estimate of learners' productive vocabulary knowledge at two frequency levels, namely the 1K and 2K bands, for which 30 items were sampled from the frequency list of Lonsdale and Le Bras (2009), as in the receptive vocabulary test. We focused only on the 2,000 most frequent words because a pilot study had shown that learners obtained only very low scores on the 3K and 4K frequency word bands. We also did this since learners tend to know fewer words productively than receptively (Laufer, 1998) and in order to avoid test fatigue because of the lack of knowledge of too many words. The test has a written fill-in-the-gaps format in which a French sentence is given. The frequency of the words in the sentences was the same as or of higher frequency than the word that had to be

filled out. The Dutch translation of the word is provided in parentheses, and the first or the first two letters are given as well as the number of letters the word contains, in order to minimize guessing. The test had good internal consistency (Cronbach's $\alpha = 0.90$, $n = 50$).

An example follows:

Le champion olympique, hospitalisé depuis trois mois et demi après un grave accident de ski, fait de petits p_____ (vooruitgang) encourageants.

[The Olympic champion, hospitalized for three and a half months after a terrible skiing accident, is making encouraging progress.]

Speaking tasks

The speaking tasks that were administered were semi-structured dialogic speaking tasks, in which a learner had to interact with the researcher. The topics of the two tasks, namely, “a doctor's visit” and “a job interview,” were chosen because they are described as such in the CEFR at the B1 level, which is the proficiency level that the L2 learners in the present study were supposed to have reached, and they are topics that typically appear in standardized B1 examinations, such as the Diplôme d'étude en langue française (DELF, Centre International d'Études Pédagogiques).

The L2 French learners in the present study were supposed to be able to interact about those two topics at the time they did the tasks. L1 speakers performed the same semi-structured dialogic speaking tasks, which would allow us to compare L1 and L2 speakers' output. Finally, two topics were chosen because genre and topic might have an effect on the output (e.g., Bygate, 1999; Pauwels, 2016).

Speaking task 1: A doctor's visit. Both the L1 and the L2 speakers received the instructions in French and were told that they did not feel well. They had to describe their symptoms to the doctor, that is, the researcher. Images were added in order to increase the comparability between the speakers' output and to guide the discussion, since a pilot study had shown that L2 learners found it useful and easier to talk about the topic when pictures that could inspire them were provided.

Task 1:

*Vous êtes allé(e) à la montagne et en rentrant vous vous rendez compte que vous êtes malade.
Vous vous rendez chez le médecin et vous lui décrivez vos symptômes.*

Vous pouvez utiliser les images ci-dessous pour vous aider.

[You went on a trip to the mountains and now you are not feeling well. You go to the doctor and describe your symptoms.

You can use the images below to help you.]

Speaking task 2: A job interview. The second semi-structured dialogic task was also one of the topics presented in the DELF at the B1 CEFR level. The difference between this task and the previous speaking task (a doctor's visit) was that there were no images added, since this task was more abstract. Both L1 and L2 speakers were told in French that they had to present themselves at an employment agency to find a student job and that they had two options: a student job at the bakery or one in the supermarket. The different options were given in order to provide participants with a choice, which, however, was limited to two topics that were supposed to be close to the participants' daily life. The instructions were also given on paper.

Task 2:

Vous vous présentez à une agence pour l'emploi pour obtenir un job d'étudiant.

Il y a deux propositions :

- *un job d'étudiant dans une boulangerie*
- *un job d'étudiant dans un supermarché*

L'employé vous demande de vous présenter (études, expérience, motivation) et vous essayez de le convaincre.

[You present yourself at an employment agency to find a student job. You can choose between a job at a bakery or in a supermarket.

You have to introduce yourself (studies, experience, motivation) and try to convince them that you are the person they need.]

Procedure

The data collection with the non-native speakers took place on different days and was conducted by two researchers. The data collection with the native speakers also took place on different days but was conducted by only one researcher. The participants all started with the doctor's visit. They first read the instructions and received two minutes of planning time before doing the task. Subsequently, they received the instructions for the second speaking task, that is, the job interview. As in the first task, they had two minutes to prepare themselves. Both speaking tasks took about fifteen minutes per participant for the L2 speakers and ten minutes for the L1 speakers. Even though the same questions were asked to both L2 and L1 speakers, L1 speakers answered more quickly and therefore needed less time in total. The procedure was the same for all the participants, regardless of their school or the researcher who administered the speaking task.

Rating instrument and human ratings

Participants' spoken output was measured in terms of holistic features of their performance and was evaluated by two raters from a panel of three. The three raters assessed the oral performances of both topics and of L1 and L2 speakers. The raters, who were native speakers of French and had a background in language studies and/or teaching, did not know in advance whether they would be rating an L1 or an L2 performance. As raters performed the ratings of the audio files on a voluntary basis, they completed them at their own pace, after having participated in a training session, in which they rated audio files from the pilot study in the presence of the researcher. Unlike Appel et al. (2019) and Crossley et al. (2015), we used the audio files for the rating procedure in order to stay as close to the authentic situation as possible.

The rating instrument that was used consisted of a holistic rating scale, in which the evaluations were prompted through a 5-point Likert scale. The holistic lexical rating scale was translated into French and adapted from the rating scale used by Crossley et al. (2015).

Scoring and analysis

VOCABULARY TESTS

The receptive test was scored dichotomously. A score of zero points was attributed when a learner indicated one of the three incorrect options or the "I don't know" option. A correct answer would lead to one point, so there was a maximum possible score of 120 points. Learners who obtained a score of 90% on a 1,000-word section – that is, 27 out of 30 on

a frequency band – were supposed to have mastered that level (Nation, n.d., as cited in Stæhr, 2009).

The productive test was also scored dichotomously: One point was given for every correct answer, zero points for every incorrect answer. Spelling mistakes were not taken into account. There was a Pearson correlation of 0.997 between the two raters who corrected the productive tests.

ELICITED SPOKEN DATA

Following previous research (e.g., Brown et al., 2005; Malvern & Richards, 2002), the spoken interactions were transcribed in CHAT (Codes for the Human Analysis of Transcripts) format (MacWhinney, 2000), and the Computerized Language Analysis (CLAN) package of programs (MacWhinney, 2000), which contains a large number of automatic analyses, was also used for our analyses. Lextutor (Cobb, n.d.) was used for determining the LFP of the participants' productions. Furthermore, the CLAN programs and the Gramulator 6.0 (McCarthy, Watanabe, & Lamkin, 2012) were used to calculate the measures of LD. D and HD-D were selected for our analyses, since Treffers-Daller (2013) showed that both measures are good indicators of language ability in French.

Given that the non-native data were collected by two researchers and that data were normally distributed, an independent *t*-test was run to verify whether the researcher who served as an interlocutor had an effect on the learners' use of vocabulary (number of tokens, types, and lemmas; LFP; measures of LD). The results showed that there was no interviewer effect³ and that the data collected by the two interviewers could be analyzed together. We also investigated, using an independent *t*-test, whether taking a three- or a four-hour French course had an influence on learners' vocabulary, which did not seem to be an issue.⁴ Therefore, no distinction was made between the two groups of learners. Finally, it should also be mentioned that we controlled for the words that speakers clearly did not know and repeated after the researcher produced them. These were not taken into account for the analyses.

Intraclass Correlation Coefficient (ICC) estimates were calculated using SPSS version 25, based on a random consistency model in order to check the consistency of raters' judgments. The results showed that there was a high degree of inter-rater reliability. The average ICC was 0.90, with a 95% confidence interval from 0.88 to 0.91, $F(1401) = 9.57$, $p < 0.001$.

First, in order to determine the factors that predict the holistic rating scores of L1 and L2 speakers' vocabulary use in oral output in two semi-structured dialogic speaking tasks, we ran a multivariate linear mixed effects model with (a) the holistic rating score as dependent variable, (b) topic as the fixed main factor, (c) number of types, number of lemmas 1K, number of tokens 2K, and HD-D as covariates, and (d) participants as random effect. This analysis was used because there were two tasks and two expert ratings per participant. We started with a model including all the aforementioned main factors before carrying out a backward stepwise analysis to arrive at the simplest model possible. Each time a non-significant factor was removed, the models were compared using *chi*-square tests and the AIC values in order to look for the best-fitted model.

In order to answer the second research question, independent *t*-tests were used for the normally distributed variables and Mann-Whitney U-tests for the variables that were not normally distributed. The number of tokens was kept constant between 90 and 450 tokens for calculating LD. Although a cut-off between 100 and 400 tokens is suggested by McCarthy and Jarvis (2007) for calculating D, some studies (e.g., Appel et al., 2019) have suggested slightly

more lenient cut-off scores in order to capture “a wider range of linguistic abilities represented in the corpus, as lower level speakers often produce shorter samples than their higher level counterparts” (Appel et al., 2019, p. 31). This was also the case in the present study.

Finally, Pearson correlations were computed to evaluate the relationship between the receptive and productive vocabulary knowledge tests and holistic ratings of the L2 speakers’ output, since data were normally distributed. This enabled us to answer our third research question.

Results

Predicting factors for experts’ ratings of speaking tasks

To better understand the importance of the holistic rating scores of participants’ spoken output in the two semi-structured dialogic tasks and the factors that could predict these ratings, a linear mixed effects model was conducted.

First, we checked for multicollinearity. As for all measures, the variables displayed Pearson correlations above 0.70, which is the threshold commonly used in research (e.g., Appel et al., 2019; Crossley et al., 2011b), they could not all be included in the regression model. Therefore, the variables that correlated best with the holistic rating score were retained as predictor variables for the analyses (Table 1). As a result, only four variables – namely, the number of types used, lemma 1K, tokens 2K, and HD-D – were used in the analyses. The results show that three parameters – HD-D, the number of types, and the number of 1K lemmas used – were positively and significantly related to raters’ holistic scores (Table 2). All the predictors together explained 64.57% of the variance of raters’ holistic scores. HD-D predicted most (i.e., 55.54%) of the variance in human ratings of spoken output based on the null model, followed by the number of types (30.91%), and finally the number of 1K lemmas (17.57%).

Number of words produced

Table 3 provides the mean number of tokens, types, and lemmas used by L2 and L1 speakers for the doctor’s topic and the interview topic, as well as the results of the Mann-Whitney U-test. The results show that L1 speakers used slightly more tokens, types, and lemmas for both topics than L2 speakers, even though they needed less time to perform the speaking tasks.

Since most of the variables were not normally distributed, Mann-Whitney U-tests were computed between L1 and L2 speakers to evaluate whether both groups statistically differed in the number of words they produced. The difference in the number of lemmas between L1 and L2 speakers for the doctor’s topic was normally distributed, and an independent *t*-test was used (Table 4). As can be seen in Tables 3 and 4, there were significant differences between both groups for both topics regarding the number of tokens, types, and lemmas used. Therefore, native speakers seem to use more tokens, types, and lemmas while speaking.

Lexical frequency profile of the words produced

In order to answer the second part of our second research question, we looked at the LFP of L1 and L2 speakers’ output, which is available in Tables 5 and 6. Independent *t*-tests were used for the variables that were normally distributed (Table 5) and Mann-Whitney U-tests for the variables that were not normally distributed (Table 6). It should be mentioned that these two task types are supposed to elicit mainly high-frequency words, as learners at the

Table 1: Significant correlations with holistic rating scores for both topics

Measures	Lexical variable	Pearson correlation
Word count	Tokens	0.25**
	Types	0.42**
	Lemmas	0.37**
Lexical frequency profile	Tokens 1K	−0.35**
	Tokens 2K	0.44**
	Tokens 2K+	0.21*
	Types 1K	−0.05
	Types 2K	0.34**
	Types 2K+	−0.01
	Lemmas 1K	−0.43**
	Lemmas 2K	0.39**
	Lemmas 2K+	0.23**
Lexical diversity	HD-D	0.70**
	vocd-D	0.68**

* significant at the 0.05 level;

** significant at the 0.01 level.

B1 level are supposed to interact with such words. The results show that this was the case for both L1 and L2 speakers. In addition, L1 speakers used slightly more low-frequency tokens, types, and lemmas than L2 speakers for both topics. Even though the differences between both groups were statistically significant, the differences in percentages are small and probably represent only a limited number of tokens, types, and lemmas. For the doctor's topic, the number of 2K types produced by L1 and L2 speakers was not significantly different. Both groups seemed to use a similar number of unique words. As far as the interview topic is concerned, the results show that there was no statistically significant difference between the number of low-frequency words (i.e., tokens, types, and lemmas beyond 2K) that L1 and L2 learners use.

Table 2: Parameter estimates linear mixed effects model for holistic rating score

Fixed effects	Estimate	SE	t-value	p-value	95% confidence interval	
					Lower bound	Upper bound
Intercept	7.097	1.084	6.542	0.000	4.96	9.23
Topic = Interview	−0.702	0.109	−6.429	0.000	−0.92	−0.49
Topic = Doctor	0	0				
Types	0.008	0.003	3.118	0.002	0.00	0.01
Lemmas 1K	−0.042	0.013	−3.232	0.001	−0.067	−0.016
HD-D	0.174	0.031	5.681	0.000	0.114	0.235
Residual	0.30	0.03		0.000	0.244	0.368
Intercept (participant)	0.50	0.11		0.000	0.332	0.760
−2LL	598.88					

Note: SE = standard error.

Lexical diversity of the words produced

Table 7 displays the descriptive results of the LD and the differences between L1 and L2 speakers. It shows that both L1 and L2 speakers produced output that was more lexically diverse for the interview topic than for the doctor’s topic, with higher values for D indicating a higher LD and with values closer to zero for HD-D indicating a more lexically diverse output. Moreover, the results indicate that the performance of L1 speakers was more lexically diverse than that of L2 speakers and that the variation between speakers was also higher for the L2 speakers, even though the number of tokens had been kept constant, that is, between 90 and 450 tokens. As the data were normally distributed,⁵ we computed independent *t*-tests (Table 7), which indicate that there were statistically significant differences for both topics between L1 and L2 speakers concerning LD, as measured by D and HD-D. L1 speakers’ output was more lexically diverse than L2 speakers’ output. The effect sizes, which are very large (Cohen, 1988) for both topics, are higher for HD-D than for D.

Descriptive results of vocabulary knowledge tests and experts’ ratings of speaking tasks

Table 8 shows the results of the receptive and productive vocabulary knowledge tests of the L2 learners per K level. It displays results that are higher on the receptive vocabulary test than on the productive vocabulary test. The correlation between the receptive and the productive vocabulary tests is $r = 0.67$ ($p < 0.001$, $N = 50$). With respect to the receptive vocabulary test,

Table 3: Descriptives and summary results of the Mann-Whitney U-test for the number of tokens for the doctor's topic and for the interview topic

Variable	L1 (<i>n</i> = 27)		L2 (<i>n</i> = 50)		Mann-Whitney U	Z-score	<i>p</i> -value	Effect size <i>d</i> ^a
	<i>M</i> (<i>SD</i>)	95% CI	<i>M</i> (<i>SD</i>)	95% CI				
Doctor								
Tokens	173.78 (67.23)	[147.18, 200.37]	143.30 (62.69)	[125.48, 161.12]	462.00	−2.27	0.023	0.54
Types	85.37 (25.80)	[75.17, 95.57]	61.80 (19.49)	[56.26, 67.34]	293.50	−4.07	0.000	1.05
Interview								
Tokens	362.81 (119.46)	[315.56, 410.07]	238.24 (80.93)	[215.24, 261.24]	242.50	−4.62	0.000	1.24
Types	151.93 (33.48)	[138.68, 165.17]	95.70 (24.41)	[88.76, 102.64]	93	−6.21	0.000	2.00
Lemmas	128.00 (27.01)	[117.32, 138.68]	86.12 (25.37)	[78.91, 93.33]	165	−5.45	0.000	1.58

^a Effect sizes in all tables were calculated using the tools retrieved from https://www.psychometria.de/effect_size.html.

Table 4: Descriptives and summary results of the independent samples *t*-test between L1 and L2 speakers for the doctor's topic

Variable	L1 (n = 27)		L2 (n = 50)		<i>t</i> -value	<i>p</i> -value	Effect size <i>d</i>
	<i>M</i> (SD)	95% CI	<i>M</i> (SD)	95% CI			
Lemmas	73.07 (21.13)	[64.71, 81.43]	57.62 (17.18)	[52.74, 62.50]	−3.47	0.001	0.83

learners seemed to be familiar with the 1,000 most frequent words in French ($M = 91.87\%$) and mostly familiar with the 2,000 most frequent words ($M = 87.20\%$), as obtaining 90% on a 1,000-word level is considered mastery of that K level (e.g., Read, 1988; Schmitt, Schmitt, & Clapham, 2001). Learners did not, however, obtain the 90% benchmark for mastering the 1K level in French ($M = 72.07\%$) on the productive test. Moreover, the standard deviation

Table 5: Descriptives and summary results of the independent samples *t*-test between L1 and L2 speakers for both topics

Variable	L1 (<i>n</i> = 27)		L2 (<i>n</i> = 50)		<i>t</i> -value	<i>p</i> -value	Effect size <i>d</i>
	<i>M</i> (<i>SD</i>)	95% CI	<i>M</i> (<i>SD</i>)	95% CI			
Doctor							
Tokens 1K	85.14 (3.13)	[83.91, 86.38]	90.08 (2.79)	[89.29, 90.87]	7.09	0.000	1.69
Tokens 2K	4.93 (1.11)	[4.49–5.36]	4.20 (1.62)	[3.74–4.66]	−2.08	0.041	0.50
Tokens beyond 2K	9.93 (3.21)	[8.66, 11.20]	5.41 (2.39)	[4.73, 6.09]	−7.00	0.000	1.67
Types 1K	74.37 (4.29)	[72.67, 76.06]	79.96 (4.80)	[78.60, 81.33]	5.65	0.000	1.35
Types 2K	9.11 (2.27)	[8.21, 10.01]	8.18 (2.66)	[7.42, 8.93]	−1.54	0.128	0.37
Types beyond 2K	15.69 (4.10)	[14.07, 17.31]	9.96 (4.12)	[8.79, 11.13]	−5.84	0.000	1.40
Lemmas 1K	76.61 (4.35)	[74.89, 78.33]	84.55 (3.84)	[83.46, 85.64]	8.26	0.000	1.97
Lemmas 2K	10.54 (2.81)	[9.42, 11.65]	8.64 (3.06)	[7.77, 9.51]	−2.66	0.010	0.64
Lemmas beyond 2K	12.85 (3.80)	[11.35, 14.36]	6.78 (3.08)	[5.90, 7.65]	−7.60	0.000	1.82
Interview							
Tokens 1K	87.61 (1.84)	[86.88, 88.34]	89.44 (2.40)	[88.76, 90.12]	3.45	0.001	0.82
Tokens 2K	4.24 (1.26)	[3.74, 4.74]	2.70 (0.73)	[2.49, 2.91]	−5.84	0.000	1.40
Tokens beyond 2K	8.03 (1.16)	[7.57, 8.49]	7.72 (2.09)	[7.12, 8.31]	−0.84	0.403	0.20

(Continued)

Variable	L1 (n = 27)		L2 (n = 50)		t-value	p-value	Effect size d
	M (SD)	95% CI	M (SD)	95% CI			
Types 1K	77.29 (2.47)	[76.32, 78.27]	77.43 (4.44)	[76.17, 78.70]	0.18	0.857	0.04
Types 2K	8.02 (1.86)	[7.29, 8.76]	6.20 (1.61)	[5.74, 6.65]	−4.49	0.000	1.07
Types beyond 2K	14.31 (2.35)	[13.38, 15.24]	15.04 (3.78)	[13.97, 16.12]	0.92	0.361	0.22
Lemmas 1K	77.75 (2.60)	[76.72, 78.78]	82.05 (3.42)	[81.07, −83.02]	5.69	0.000	1.36
Lemmas 2K	9.30 (2.22)	[8.42, 10.17]	6.82 (1.82)	[6.30, 7.34]	−5.25	0.000	1.25
Lemmas beyond 2K	12.96 (2.38)	[12.02, 13.90]	11.12 (2.88)	[10.30, 11.94]	−2.83	0.006	0.68

Table 6: Descriptives and summary results of the Mann-Whitney U-test for L1 and L2 speakers for tokens and lemmas for the doctor’s topic

Variable	L1 (<i>n</i> = 27)		L2 (<i>n</i> = 50)		Mann-Whitney U	Z-score	<i>p</i> -value	Effect size <i>d</i>
	<i>M</i> (<i>SD</i>)	95% CI	<i>M</i> (<i>SD</i>)	95% CI				
Doctor								
Tokens 2K	4.93 (1.11)	[4.49, 5.36]	4.02 (1.62)	[3.74, 4.66]	413.50	−2.79	0.005	0.67
Lemmas 1K	76.61 (4.35)	[74.89, 78.33]	84.55 (3.84)	[83.46, 85.64]	115.00	−5.98	0.000	1.86

of the productive vocabulary test is higher than of the receptive vocabulary test, which indicates that there might have been more variation in learners’ productive vocabulary than in their receptive vocabulary. Finally, the mean holistic rating scores for the L2 speakers ($N = 50$) are similar for both topics, that is, 2.70 for the doctor’s topic ($SD = 0.78$) and 2.72 for the interview topic ($SD = 0.79$) on a scale from 1 to 5.

Table 7: Descriptives and summary results of the independent samples t-test of the lexical diversity of L1 and L2 speakers for both topics

Variable	L1		L2		t-value	p-value	Effect size <i>d</i>
	<i>M</i> (<i>SD</i>)	95% CI	<i>M</i> (<i>SD</i>)	95% CI			
Doctor	<i>n</i> = 24		<i>n</i> = 43				
Vocd-D	46.20 (7.66)	[42.96, 49.43]	26.47 (9.02)	[23.70, 29.25]	9.04	0.000	2.30
HD-D	-1.75 (1.02)	[-2.18, -1.32]	-6.02 (2.43)	[-6.77, -5.27]	10.05	0.000	2.56
Interview	<i>n</i> = 21		<i>n</i> = 50				
Vocd-D	58.81 (8.00)	[55.17, 62.45]	33.30 (8.62)	[30.85, 35.75]	11.61	0.000	3.02
HD-D	-0.18 (0.71)	[-0.50, 0.15]	-3.93 (1.89)	[-4.47, -3.40]	12.15	0.000	3.16

Relationship between vocabulary knowledge tests and experts’ ratings of speaking tasks

In order to investigate the relationship between L2 learners’ vocabulary knowledge and the holistic rating scores given by raters, Pearson correlations were computed. We found only one positive, significant correlation between the receptive vocabulary test and the holistic ratings for the doctor’s topic, as can be seen in Table 9. No significant correlations were found with the productive vocabulary test.

Discussion

The current study aimed to determine which factors could predict holistic lexical rating scores of L1 and L2 speakers’ spoken output in two semi-structured dialogic speaking tasks. Next, it looked at how the output of L1 and L2 speakers differs and finally aimed to explore the relationship between low-intermediate L2 learners’ vocabulary knowledge and holistic lexical rating scores given by experts on learners’ output on two semi-structured dialogic speaking tasks.

Predictors of raters’ holistic scores of L1 and L2 speakers’ oral output

With respect to the first research question, we investigated the factors that could predict raters’ holistic ratings of L1 and L2 speakers’ vocabulary use in oral output. The measure that could best predict raters’ holistic scores was the LD of participants’ output, followed by the number of types and the number of high-frequency lemmas. Speech that contained more low-frequency words was scored higher than speech that contained more high-frequency

Table 8: Descriptives per K level for the receptive and productive vocabulary tests of the L2 learners (*N* = 50)

K level	M percentage (SD)	95% CI	M raw scores (SD)	95% CI
Receptive vocabulary test				
1K	91.87 (6.57)	[90, 93.73]	27.56 (1.98)	[27, 28.12]
2K	87.20 (8.12)	[84.89, 89.51]	26.16 (2.44)	[25.47, 26.85]
3K	72.13 (11.52)	[68.86, 75.41]	21.64 (3.46)	[20.66, 22.62]
4K	69.73 (12.36)	[65.22, 74.25]	20.92 (4.76)	[19.57, 22.27]
Total	80.23 (9.30)	[77.59, 82.88]	96.28 (11.16)	[93.11, 99.45]
Productive vocabulary test				
0–1,000	72.07 (12.36)	[68.55, 75.58]	21.62 (3.71)	[20.57, 22.67]
1,001–2,000	61.73 (16.01)	[57.18, 66.28]	18.52 (4.80)	[17.15, 19.89]
Total	66.90 (13.48)	[63.07, 70.73]	40.14 (8.09)	[37.84, 42.44]

Note: The mean raw scores are out of 30 per frequency band, out of 120 for the total receptive test and out of 60 for the total productive test.

Table 9: Pearson correlation vocabulary tests and mean holistic rating score (*N* = 50)

Type of test	M rating doctor	<i>p</i> -value	M rating interview	<i>p</i> -value
Receptive vocabulary test	0.32* $r^2 = 0.10$	0.025	0.19 $r^2 = 0.04$	0.185
Productive vocabulary test	0.27 $r^2 = 0.07$	0.054	0.05 $r^2 = 0.003$	0.742

* significant at the 0.05 level.

words. This is in line with previous research in speaking and writing (Kyle & Crossley, 2015), even though the methodologies are not entirely comparable. Nevertheless, despite differences in tasks and language, these studies show that speaking samples of high proficiency display more low-frequency words.

Moreover, the results showed that the number of tokens did not predict ratings, whereas the number of types did. These results confirm previous research (Iwashita et al., 2008; Lorenzo-Dus & Meara, 2005) that demonstrated that the number of unique words (i.e., types) a speaker uses is a strong predictor for raters' scores of speaking proficiency. Such trends indicate that raters will better evaluate speakers who produce a higher number of unique words. More proficient speakers can access and use a higher number of (unique) words than less proficient speakers, which is indicative of a larger vocabulary.

Finally, the results of LD are also in line with previous research (Crossley et al., 2011a; Yu, 2010). Whereas HD-D explained up to 55.54% of the variance based on the null model in spoken output in the present study, the results were lower in previous research, which might have been due to the LD measure that was used. Despite differences in number, the evidence shows that speaking samples that are more lexically diverse receive better ratings, which might indicate that a more diverse vocabulary is associated with a larger vocabulary knowledge.

Differences in lexical output between L1 and L2 speakers

The second focus of the study was to explore how L1 and L2 speakers' output in two semi-structured dialogic tasks differs regarding the number of tokens, types, and lemmas; the LFP; and the LD. The results show that L1 speakers used a higher number of tokens, types, and lemmas than L2 speakers for both topics. These results are in line with the findings of Treffers-Daller (2013), who found that French native speakers produced more tokens and types than L2 learners of French. It should be noted that there were considerable differences between the two topics in terms of number of tokens and types used. This holds for the L1 as well as the L2 speakers and thus illustrates that topic can have an impact on the number of words produced. We assume that the doctor's task was a narrower task, whereas the job interview allowed participants to elaborate on their profile and motivation for the job.

In our study, the data were elicited through two B1 tasks, which should elicit a high percentage of high-frequency words. This is confirmed, as both groups of speakers drew the majority of the tokens, types, and lemmas from the 1,000 most frequent words in French. These findings are also in line with Nation's (2013, p. 22) claim that a small group of high-frequency words covers a very large proportion of the words in texts. Even though L1 speakers used slightly more low-frequency words than L2 speakers in both tasks, both groups employed mainly high-frequency words. The results confirm previous research looking into the frequency of the words that low and highly proficient speakers use (Laufer & Nation, 1995; Lindqvist, 2010; Ovtcharov et al., 2006), which concluded that the more low-frequency or rare words speakers use, the more proficient they are. L1 speakers are often assumed to use more low-frequency or rare words. Furthermore, the results of our study indicate that speakers using more low-frequency words were also better rated, which confirms previous research (Crossley & McNamara, 2013; Kyle & Crossley, 2015).

Finally, as far as the LD of both L1 and L2 speakers is concerned, the results show that L1 speakers produced output that was more lexically diverse than L2 speakers' output, as measured by D and HD-D, for both topics. This finding is in line with previous research

(Treffers-Daller, 2013) in which the LD of L1 speakers was shown to be higher than that of L2 speakers. In sum, more lexically proficient French speakers not only produce more words but also use less frequent words and more different words.

Relationship between vocabulary knowledge and experts' ratings of speaking tasks

Finally, we investigated the relationship between L2 learners' vocabulary knowledge and holistic lexical ratings. The results are not conclusive, as there was a significant correlation only between the receptive vocabulary test and the holistic rating of the doctor's topic, but not that of the interview topic. This means that learners who had a higher score on the receptive vocabulary test received a better rating score for their output in the doctor's task. These findings might be explained by the use of fewer 1K words and more 2K and beyond 2K words for the doctor's topic. Another possible explanation might be the fact that the content of a doctor's visit is more predictable and thus more comparable than that of a job interview. Different correlations were found in Uchiyama and Clenton (2018), but this could be due to differences in the task type (monologic versus interactive speaking tasks in the present study). It was expected that a productive vocabulary test would be a better predictor of vocabulary use than a receptive test, as speakers need to access their L2 or even L1 productive lexicon (Milton, 2009, p. 121). The L2 learners' limited productive vocabulary knowledge and the fact that – as a result – only two frequency bands were tested suggest that there might not have been enough variance to reveal any correlation between productive vocabulary knowledge and learners' vocabulary use in the two speaking tasks. It is at this stage too early to draw any firm conclusions regarding learners' productive vocabulary knowledge. The receptive test contained items from four frequency bands and had 120 items compared with 60 in the productive test. As a result, it was probably easier to differentiate between learners' vocabulary in the receptive than in the productive test. This might explain the differences between the receptive and productive test.

Limitations and future research

There are a number of limitations to this study that need to be acknowledged and addressed in future research. First, as the aim of this study was to focus on low-intermediate L2 learners, future research should also include other proficiency levels. Second, although two task topics were given to the speakers to control for a topic effect, more research into other task types, such as cognitively more demanding task types, is warranted, since such task types could reveal differences in learners' vocabulary knowledge and lead, for instance, to the use of fewer words or the use of more high-frequency words. Moreover, the tasks were not counterbalanced, since all learners first did the doctor's task. This might have led to the production of a lower number of words due to test anxiety. Third, only D and HD-D were used to investigate the LD of speakers' output. Unlike the case for English, there are not many automated measures or lexical tools that allow a lexical analysis of spoken language for French. Therefore, studies focusing on the development of such tools are needed (see Lindqvist, Gudmundson, and Bardel, 2013, for an exception). Fourth, the productive vocabulary test in this study tested only up to the 2,000 most frequent words in French. Future research could include a test measuring knowledge of more frequency bands or a test depending on the learners' proficiency level. Finally, in order to better predict experts' holistic ratings, we should not only take into account complexity as measured by lexical parameters but also focus on accuracy and fluency.

Conclusion

This study sought to explore which factors could predict holistic lexical ratings of L1 and L2 speakers' oral output in two semi-structured dialogic speaking tasks. In addition, this study investigated the differences between L1 and L2 speakers' output in the same speaking tasks and explored the relationship between L2 learners' receptive and productive French vocabulary knowledge and experts' holistic ratings of learners' lexical output. With respect to the predictors of holistic human ratings of vocabulary use in two semi-structured dialogic speaking tasks, the results show that the more diverse the spoken output is, the higher the ratings are. Both the number of types and the number of high-frequency lemmas have an influence on how experts perceive speakers' output. Second, a comparison between the lexical output of L1 and L2 speakers shows that L1 speakers not only use more tokens, types, and lemmas but also use more low-frequency words and produce output that is more lexically diverse. Finally, the results for the relationship between L2 learners' receptive and productive vocabulary knowledge and experts' holistic lexical ratings suggest that a receptive vocabulary test might predict experts' holistic lexical ratings better than a productive test. This study adds to our understanding of the lexical evaluation of French spoken output in semi-structured dialogic tasks at an intermediate level of language proficiency.

Ann-Sophie Noreillie is a post-doctoral researcher at the KU Leuven. Her main research interests include French vocabulary acquisition and learning, language testing, and vocabulary use in listening and speaking.

Piet Desmet is a full professor at the KU Leuven. His research interests are second language acquisition and technology, with applications in computer-assisted language learning, learning analytics, and language technology and corpus linguistics.

Elke Peters is an associate professor at the KU Leuven. Her main research interests involve deliberate and incidental FL vocabulary learning inside and outside the classroom and how different types of input can contribute to vocabulary learning (single words and multiword units).

Notes

1. For a detailed discussion of lexical diversity, we refer to Jarvis (2013).
2. The distinction between French as a foreign language and French as a second language could be interpreted in different ways. According to Cuq and Gruca (2005, p. 96), a second language differs from a foreign language as far as its social and/or legal status is concerned and by the degree of appropriation of the language in a community, that is, bilingual or plurilingual. As a consequence, French in Flanders, which is officially a monolingual region, would be considered a foreign language even though French is one of the official languages in Belgium. In Brussels, however, it would definitely be a second language, given the authorization and use of both French and Dutch in administration and daily life. Nevertheless, Cuq and Gruca agree that it is difficult to qualify an official language of a country or a first language of an important group of a population of that country as foreign, as would be the case for Belgium, Canada, and Switzerland, for instance. Therefore, in our opinion, both choices could be argued for.
3. An independent *t*-test showed that interacting with one or the other researcher did not have an effect on learners' vocabulary. The differences for both groups on the different measures were not significant, $t(74) = 1.83$, $p > 0.05$. Therefore, we analyzed data of those learners as part of the same group.

4. An independent *t*-test showed that having a three-hour course or a four-hour course did not affect learners' vocabulary. The differences for both groups and both tests were not significant, $t(48) = 1.24, p > 0.05$ for the receptive test and $t(48) = 0.96, p > 0.05$ for the productive test. Therefore, we analyzed data of those learners as part of the same group.
5. As the data of the non-native speakers for the doctor's topic were lightly skewed for both measures of LD, we also performed a Mann-Whitney U-test in order to have a more robust analysis. The results were very similar and showed that both groups were statistically significant for vocd-D ($U = 53.000, z = -6.054, p < 0.000, r = 0.74$) and for HD-D ($U = 40.000, z = -6.224, p < 0.000, r = 0.76$).

References

- Appel, R., Trofimovich, P., Saito, K., Isaacs, T., & Webb, S. (2019). Lexical aspects of comprehensibility and nativeness from the perspective of native-speaking English raters. *ITL International Journal for Applied Linguistics*, 170(1), 24–52. <https://doi.org/10.1075/itl.17026.app>
- Batista, R., & Horst, M. (2016). A new receptive vocabulary size test for French. *The Canadian Modern Language Review*, 72(2), 211–233. <https://doi.org/10.3138/cmlr.2820>
- Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series*, 2005(1), 1–157. <https://doi.org/10.1002/j.2333-8504.2005.tb01982.x>
- Bygate, M. (1999). Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks. *Language Teaching Research*, 3(3), 185–214. <https://doi.org/10.1177/136216889900300302>
- Carroll, J. (1964). *Language and thought*. Englewood Cliffs, NJ: Prentice Hall.
- Cobb, T. *Compleat Web VP v.2* [Computer program]. Retrieved from <https://www.lex tutor.ca/vp/comp/>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Laurence Erlbaum Associates.
- Crossley, S.A., & McNamara, D.S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115–135. <https://doi.org/10.1111/j.1467-9817.2010.01449.x>
- Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, 17(2), 171–192.
- Crossley, S.A., Salsbury, T., & McNamara, D.S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, 36(5), 570–590. <https://doi.org/10.1093/applin/amt056>
- Crossley, S.A., Salsbury, T., McNamara, D.S., & Jarvis, S. (2011a). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, 45(1), 182–193. <https://doi.org/10.5054/tq.2010.244019>
- Crossley, S.A., Salsbury, T., McNamara, D.S., & Jarvis, S. (2011b). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, 28(4), 561–580. <https://doi.org/10.1177/0265532210378031>
- Cuq, J.-P., & Gruca, I. (2005). *Cours de didactique du français langue étrangère et seconde*. Grenoble, France: Presses universitaires de Grenoble.
- Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24(2), 197–222. <https://doi.org/10.1093/applin/24.2.197>
- de Jong, N.H., Steinel, M.P., Florijn, A.F., Schoonen, R., & Hulstijn, J.H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, 34(1), 5–34. <https://doi.org/10.1017/S0272263111000489>
- Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Dordrecht, Netherlands: D. Reidel.
- Iwashita, N. (2010). Features of oral proficiency in task performance by EFL and JFL learners. In M.T. Prior, Y. Watanabe, & S.-K. Lee (Eds.), *Selected proceedings of the 2008 Second Language Research Forum: Exploring SLA, perspectives, positions, and practices* (pp. 32–47). Somerville, MA: Cascadia Proceedings Project.

- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24–49. <https://doi.org/10.1093/applin/amm017>
- Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63(S1), 87–106. <https://doi.org/10.1111/j.1467-9922.2012.00739.x>
- Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research*, 4(5), 900–913. <https://doi.org/10.4304/jltr.4.5.900-913>
- Kyle, K. (2020). Measuring lexical richness. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 454–476). London, England: Routledge.
- Kyle, K., & Crossley, S.A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757–786. <https://doi.org/10.1002/tesq.194>
- Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, 19(2), 255–271. <https://doi.org/10.1093/applin/19.2.255>
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322. <https://doi.org/10.1093/applin/16.3.307>
- Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1), 33–51. <https://doi.org/10.1177/026553229901600103>
- Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15–30.
- Lindqvist, C. (2010). La richesse lexicale dans la production orale de l'apprenant avancé de français. *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, 66(3), 393–420. <https://doi.org/10.3138/cmlr.66.3.393>
- Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to the measuring of vocabulary in L2 oral production. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 109–126). Eurosla Monograph Series 2. Amsterdam, Netherlands: Eurosla.
- Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French*. New York, NY: Routledge.
- Lorenzo-Dus, N., & Meara, P. (2005). Examiner support strategies and test-taker vocabulary. *International Review of Applied Linguistics in Language Teaching*, 43(3), 239–258. <https://doi.org/10.1515/iral.2005.43.3.239>
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, 96(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x
- MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk (3rd ed.). Part 2: The CLAN program. Mahwah, NJ: Lawrence Erlbaum Associates. <https://doi.org/10.21415/T5G10R>
- Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19(1), 85–104. <https://doi.org/10.1191/0265532202lt221oa>
- Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. London, England: Palgrave Macmillan. <https://doi.org/10.1057/9780230511804>
- McCarthy, P. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* (Unpublished doctoral dissertation). University of Memphis, Memphis, TN.
- McCarthy, P., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459–488. <https://doi.org/10.1177/0265532207080767>
- McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392. <https://doi.org/10.3758/BRM.42.2.381>
- McCarthy, P., Watanabe, S., & Lamkin, T.A. (2012). The Gramulator: A tool to identify differential linguistic features of correlative text types. In P. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 312–333). Hershey, PA: IGI Global.

- Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, England: Multilingual Matters.
- Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59–81. <https://doi.org/10.1353/cmlr.2006.0049>
- Nation, I.S.P. (2013). *Learning vocabulary in another language*. Cambridge, England: Cambridge University Press. <https://doi.org/10.1093/elt/56.1.91>
- Nation, I.S.P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 15–29). London, England: Routledge. <https://doi.org/10.4324/9780429291586>
- Noreillie, A.-S. (2019). *It's all about words: Three empirical studies into the role of lexical knowledge and use in French listening and speaking tasks* (Unpublished doctoral dissertation). KU Leuven, Leuven, Belgium.
- Noreillie, A.-S., Kestemont, B., Heylen, K., Desmet, P., & Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and French as foreign languages: An approximate replication study of Stæhr (2009). *ITL - International Journal of Applied Linguistics*, 169(1), 212–231. <https://doi.org/10.1075/itl.00013.nor>
- Ovtcharov, V., Cobb, T., & Halter, R. (2006). La richesse lexicale des productions orales : mesure fiable du niveau de compétence langagière. *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, 63, 107–125. <https://doi.org/10.3138/cmlr.63.1.107>
- Pauwels, P. (2016, July). Lexical development in/through writing. Paper presented at the Vocabulary Special Interest Group of the British Association for Applied Linguistics, Nottingham, England.
- Peters, E., Velghe, T., & Van Rompaey, T. (2019). The VocabLab tests: The development of an English and French vocabulary test. *ITL - International Journal of Applied Linguistics*, 170(1), 53–78. <https://doi.org/10.1075/itl.17029.pet>
- Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, 19, 12–25.
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18(1), 55–88.
- Stæhr, L.S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139–152. <https://doi.org/10.1080/09571730802389975>
- Stæhr, L.S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31(4), 577–607. <https://doi.org/10.1017/S0272263109990039>
- Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French. An exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–103). Philadelphia, PA: John Benjamins. <https://doi.org/10.1075/sibil.47>
- Uchihara, T., & Clenton, J. (2018). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*. Advance online publication. <https://doi.org/10.1177/1362168818799371>
- Uchihara, T., & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *Language Learning Journal*, 47(1), 64–75. <https://doi.org/10.1080/09571736.2016.1191527>
- van der Jeught, S. (2017). Territoriality and freedom of language: The case of Belgium. *Current Issues in Language Planning*, 18(2), 181–198. <https://doi.org/10.1080/14664208.2016.1243883>
- Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17(1), 65–83.
- Wu, T. (1993). An accurate computation of the hypergeometric distribution function. *ACM Transactions on Mathematical Software*, 19(1), 33–43. <https://doi.org/10.1145/151271.151274>
- Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 31(2), 236–259. <https://doi.org/10.1093/applin/amp024>