Vocabulary aspects of advanced L2 French
Do lexical formulaic sequences and lexical richness develop at the same rate?

Fanny Forsberg Lundell and Christina Lindqvist
Stockholm University / Uppsala University

In her overview of research on the advanced L2 learner, Bartning (1997) aims at characterizing the advanced learner variety. This characterization is above all based on morphosyntactic traits. The aim of this contribution is to present additional characteristics of the advanced learner as defined by Bartning (1997), as well as to describe even more advanced levels based on recent research concerning spoken L2 French. More specifically, the main issue under investigation is whether two vocabulary measures, viz. lexical richness and lexical formulaic sequences, can be used to distinguish between different advanced levels and thus contribute to the characterization of the advanced learner of French. An additional issue investigated here is whether these two lexical aspects correlate with each other or whether they develop at different rates.

Keywords: advanced learner, vocabulary acquisition, lexical richness, formulaic sequences, L2 French

1. Introduction

The present study investigates two aspects of vocabulary acquisition in Swedish advanced learners’ oral production of L2 French: lexical richness and lexical formulaic sequences. Lexical richness is here conceived of as the proportion of low-frequency vocabulary in oral production. Lexical formulaic sequences (Lexical FSs) are word combinations (e.g. V + N or Adj. + N), with combinatorial restrictions, which reflect target language idiomaticity. Although there has been a growing interest in research on the acquisition of vocabulary in L2 French (cf. Daller 2007; Lindqvist 2010; Ovtcharov et al. 2006; Tidball & Treffers-Daller 2007, 2008; Treffers-Daller et al. 2008), few studies are concerned with the vocabulary of advanced learners. In addition, as pointed out by David et al. (2009: 147),
morphosyntax has been in focus in studies of the development of French learner language.

In her overview of research on the advanced learner, Bartning (1997) aimed at characterizing the advanced learner variety, but in this case the characterization was above all based on morphosyntactic traits. In a later study, Bartning and Schlyter (2004) presented a proposal for developmental stages in Swedish learners’ L2 French (based on the InterFra corpus, (http://www.fraitaklass.su.se/english/interfra), and the Lund corpus). Again, morphosyntax constituted the basis for the description of developmental paths in L2 French. As no lexical criteria have been taken into account in these earlier studies, we will focus on two lexical traits, namely lexical richness and Lexical FSs, in the present paper. Hopefully, these lexical traits will contribute to the characterization of the advanced learner of French. We will also discuss the use of lexical traits to characterize stages beyond the advanced stages by investigating the oral production of very advanced learners. In fact, Bartning et al. (2009) have already discussed the possibility of a stage beyond the most advanced stage (6), based on criteria such as formulaic language and elaboration of the information structure. The collection of new InterFra data of learners/users in a second language setting called for a refinement of the concept of the advanced learner. Indeed, the term ‘advanced learner’ is often used to refer to university students at different levels of advancement, but few researchers investigate learners who are long term residents in the L2 community and thus receive considerably more input than formal learners, enhancing their possibilities to attain nativelike levels of L2 use. In the present paper, we investigate the question of whether lexical richness and Lexical FSs reinforce the proposal of such a stage, preliminarily called stage 7. Moreover, we also include native speaker baseline data, in order to investigate whether it is possible to attain a nativelike level for these specific vocabulary aspects.

Finally, in addition to contributing to a description of aspects of lexical competence along the advanced acquisitional stages, the aim of this paper is to investigate whether there is a relationship between the development of lexical richness and that of Lexical FSs. Both Lexical FSs and lexical richness have proved to be useful measures for describing L2 development (Forsberg 2008; Lindqvist et al. 2011), but to our knowledge, no study to date has examined the relationship between these two measures for L2 French. As both aspects seem to be dependent on frequency of input, one could expect that they would develop in a similar way.
2. **Background: Formulaic language and lexical richness in advanced L2 French**

2.1 L2 Formulaic language and the advanced L2 learner of French

Numerous L2 researchers have taken an interest in these sequences that are not generated by the language grammar, seem arbitrary to the L2 learner, have to be memorized rather than generated, but make language use sound natural and fluent and actually account for language to quite an impressive extent (cf. Erman & Warren’s, 2000, figures 50–60% of the English language). As Forsberg (2010) suggests, formulaic sequences come in many different shapes, and they are probably acquired and processed differently and used for different functions in discourse. As already mentioned, the present study focuses on the subcategory Lexical FSs.

Few studies exist on L2 French and formulaic language, both as regards beginners and advanced learners. Concerning advanced learning — the topic of this article — researchers tend to consider formulaic language either from the perspective of fluency enhancers or from the perspective of idiomaticity and nativelike selection.

One of the first and best known studies is that of Raupach (1984), who studied German university students of L2 French before and after a study-abroad period in France, with special focus on fluency gains. Raupach (1984) found that learners made use of a considerable number of *formulae* (the term used by Raupach) in order to maintain fluency. Raupach divided the formulae into the subcategories of organizers, and fillers/modifiers. The first category are thus textual organizing devices such as *c’est* (‘it is’) and *il y a* (‘there is’). Fillers and modifiers function as ‘zones of safety’ and as modalizing markers (*je crois* ‘I believe’ and *je pense* ‘I think’). One of the major findings of Raupach’s (1984) study is that the learners, especially after their stay in France, overused both types, as compared to the native speaker control group.

The development of second language fluency is also studied in Towell *et al.* (1996). Similar to Raupach (1984), they set out to measure fluency gains in university students of L2 French after a study-abroad period in France. Among other things, they discovered that fluency gains are not due to speed but to longer and more complex utterances characteristic of spoken French, e.g. clefts, pseudo-clefts and the presentation devices *il y a* and *c’est*, the latter being the most frequent chunk in spoken French (L1 and L2) (Forsberg 2008). Furthermore, Towell *et al.* (1996) found a substantial number of ‘situational lexicalized phrases’, also contributing to the gains in fluency observed after their stay in France. All in all, both of these studies showed that formulaic sequences of certain types contribute to fluency.
Another perspective on formulaic language in French is taken by Edmonds (2010). She investigated whether anglophone learners of L2 French, having resided in France for short (4 months) or long (about 10 years) periods of time make the same judgments as regards naturalness of formulaic sequences (or *conventional expressions*, in her terminology). Surprisingly, she found no differences between learner groups and native speakers, suggesting that as regards judgment, these learners are well on their way towards developing nativelike selection.

Forsberg (2008) studied the use of formulaic sequences from a more general point of view at four different learner levels as compared to native speaker production, from beginner students to very advanced users of L2 French. The aim of the study was to find out whether use of formulaic sequences is indicative of second language development. Evidence strongly suggests that it is. There is an increase in the quantity of formulaic language use as the learner progresses, so that the largest proportion was found in the production of very advanced learners and native speakers. The most difficult category to acquire for L2 learners appears to be that of Lexical FSs (e.g. *poser une question* 'ask a question', *faire attention* 'pay attention') (Forsberg 2008; Lewis 2008). In Forsberg’s study, only the very advanced users who had lived in France for at least five years used this category to the same extent as the native speakers.

Bolly (2008) studied the collocations of two high-frequency verbs, *prendre* ('take') and *donner* ('give') (what Forsberg calls Lexical FSs), in the writings of Anglophone French university L2 learners and native speakers of French. Her main findings were that the learners tended to overuse verb + noun collocations with *donner* and underuse verb + noun collocations with *prendre*. Furthermore, she found that deviances affected either the verb or the appropriateness of the entire collocation with respect to the context of use. Bolly suggests that learners’ use of collocations might be influenced by the morphosyntax and the semantics of the verbs involved in the collocations.

Furthermore, several recent studies have shown that formulaic language measures succeed at discriminating between learners at different advanced levels, where grammar measures fail (Bartning et al. 2009; Forsberg & Bartning 2010). In the first study, three groups of advanced learners of L2 French with different degrees of exposure to French were studied with regard to morphosyntax, information structure and formulaic language. Formulaic language was the only measure that yielded significant differences between learners in a foreign language setting and learners in a second language setting. In the second study, learners placed by the DIALANG test at the six different CEFR-levels (A1-C2) performed two written tasks in French. Their productions were then analyzed in terms of morphosyntactic accuracy and formulaic language. While morphosyntactic measures discriminated between levels up to the B2 level, formulaic language succeeded at
discriminating between the most advanced levels, i.e. B2-C2. Taken together, these results indicate that language proficiency in terms of structural development, and that in terms of formulaic language development, do not follow the same rate.

In view of the research presented above, we can draw the conclusion that the production of lexical collocations seems to be a persistent problem, even at advanced or very advanced levels. For this reason it was decided in the present study to investigate the potential of this category to discriminate between different learner groups as well as to compare the development of the category with another lexical measure.

2.2 Lexical richness and the advanced L2 learner of French

According to Read (2000), lexical richness can be used as a superordinate term and covers several aspects of vocabulary acquisition: lexical density (the ratio between function words and content words), lexical diversity (lexical variation), lexical sophistication (the proportion of sophisticated words e.g. low-frequency words), and proportion of errors. There are several measures of lexical richness, depending among other things on what is being analyzed (see below).

Tidball and Treffers-Daller (2007) investigated lexical richness in L2 French using different measures (the D measure, the index of Guiraud, the Advanced Guiraud, and the Limiting Relative Diversity measure). They compared two groups of learners: first-year and third-year students at a British university, and one group of native French speakers. The results showed that there were significant differences between the first-year group and the native speakers, as well as between the third-year group and the native speakers, considering all measures of lexical richness. There were also significant differences between the learner groups for all measures except one. Thus, the most advanced learners in this study did not reach the level of the native speakers regarding lexical richness.

In a later study, Tidball and Treffers-Daller (2008) investigated different ways of measuring lexical sophistication in L2 French. They used the Advanced Guiraud and tested different modifications of this measure. They found that using teachers’ judgments was the best way to define a basic vocabulary. That is, instead of using the frequency of words, for example, as an indicator of word difficulty, experienced teachers were asked to judge whether words produced by the learners in a retelling should be considered basic or advanced. The Advanced Guiraud based on teachers’ judgments was the measure that best separated two learner groups of French. Again, there were significant differences between the learner groups as well as between the most advanced learners and a group of native speakers, which indicates that the most advanced group did not attain the native speaker level of lexical richness.
Treffers-Daller (2009) investigated lexical diversity in an oral retelling task in French in three groups: Dutch/French bilinguals (Dutch dominant), Flemish learners of French, and French/English bilinguals (French dominant). The results showed that there were significant differences between the three groups: the French dominant bilinguals had a more varied vocabulary than the Dutch dominant bilinguals. In a more detailed analysis of the use of nouns and verbs using the index of Guiraud, Treffers-Daller (2009) could conclude that there were significant differences between the groups and that these differences were of the same magnitude as the analysis of the whole production.

Ovtcharov et al. (2006) examined lexical richness in intermediate and advanced learners of French in Quebec. The participants were employed by the Canadian government and French was a requirement for their job position. The data consisted in oral proficiency interviews (test situation) with a native speaker of French. The learners were divided into two proficiency levels, based on the results of the interviews. Ovtcharov et al. used the Lexical Frequency Profile method (Laufer & Nation 1995) in order to measure lexical richness from a frequency-based perspective. Within this framework, the general assumption is that a relatively high proportion of low-frequency words is indicative of a rich vocabulary. Ovtcharov et al. found that the most advanced group had a significantly higher proportion of low-frequency words than the less advanced group. They also compared the learners’ lexical profiles with those of native speakers of French and found that there were no statistically significant differences between the most advanced learners and the native speakers in the use of low-frequency words in oral production. This led the authors to conclude that the most advanced learners actually seemed to reach the lexical richness of native speakers. This result thus runs counter to the ones reported on above, where learners were not found to reach native-like levels of lexical richness. This is perhaps not unexpected, however, as the learners in the Ovtcharov et al. study were most likely more advanced learners, having French as a tool in their professional activities.

Lindqvist et al. (2011) investigated lexical richness in the oral production of Swedish learners of French and Italian L2. They used a frequency-based method similar to the Lexical Frequency Profile (Laufer & Nation 1995), the Lexical Oral Production Profile (LOPP). For French, they compared the proportion of low-frequency words in two advanced groups of learners and one group of native speakers. The learner groups were divided on the basis of Bartning and Schlyter’s proposal of morphosyntactic stages. The results showed that the most advanced learners had a significantly higher proportion of low-frequency words than the less advanced group. In addition, there were no significant differences between the most advanced group and the native speakers, which suggested that the most advanced learners approached the native speakers with regard to this particular aspect of lexical competence.
In sum, earlier studies on lexical richness have found differences between learner groups at different proficiency levels using different types of measures. Only two studies have seen that advanced learners approach native speakers as regards lexical richness (Ovtcharov et al. 2006; Lindqvist et al. 2011).

3. Formulaic language, lexical richness and correlations with other linguistic measures

A few recent studies have investigated the relationship between formulaic language, lexical richness and other linguistic measures, but the results are far from conclusive to date. Mizrahi and Laufer (2010) studied the correlation between general vocabulary knowledge (as measured by the Vocabulary Levels Test) and collocational knowledge (a sub-group of formulaic language) in near-native speakers of L2 English (level estimated through self-rating). They found that Israeli near-native speakers of English attained nativelike levels of general vocabulary but failed to reach nativelike levels as regards collocations. Gyllstad (2007), on the other hand, found clear correlations between results on a vocabulary size test and a receptive collocation test, studying advanced learners of L2 English. However, these learners were not nearly as advanced as those in Mizrahi and Laufer’s study; they were Swedish L2 high school and university students. One explanation for the diverging results would be that the two aspects develop in parallel up to a certain level, but that collocational knowledge develops slower after some time. It is also important to note that Gyllstad’s study tested receptive knowledge whereas Mizrahi and Laufer’s tested productive knowledge. It has been suggested that collocations pose far more problems in production than in reception (cf. Warren, 2005). In the same vein as Gyllstad (2007), several other studies have also shown a clear correlation between overall language proficiency and mastery of formulaic sequences (Boers et al. 2006; Lewis 2008; Stengers et al. 2011).

The question concerning correlations between lexical richness and other measures has been investigated in David et al. (2009). They examined the relationship between lexical development and morphosyntactic development in British learners of L2 French. Lexical development was measured with the index of Guiraud, which measures lexical diversity. Grammatical development was measured using Mean Length of Utterance (MLU). David et al. (2009) found that lexical diversity correlated with MLU in that the more varied the vocabulary of the learner, the greater the MLU. However, in a similar study, Malvern et al. (2004) found no correlations between lexical diversity and MLU in L2 French. In David et al. (2009), lexical diversity was also compared to the use of grammatical gender. The results showed that there were no correlations between lexical diversity and gender. Thus,
it could be concluded that lexical diversity correlated with some of the morpho-syntactic measures, but not all. As previously mentioned, Tidball and Treffers-Daller (2007, 2008) investigated different measures of lexical richness. They found that all measures (see above) correlated significantly with a C-test, which measures overall proficiency. Lemmouh (2010) investigated the relationship between different lexical aspects in the written L2 English of Swedish learners. He found that there was a modest correlation between lexical richness, as measured by the Lexical Frequency Profile (Laufer & Nation 1995), and vocabulary depth (knowledge of collocations and derivations). There were no correlations between lexical richness and receptive or productive vocabulary size.

As very few studies have examined exactly what we want to investigate here, it is difficult to know what to expect on the basis of earlier studies. As regards formulaic language, it seems that it rarely develops at the same pace as other linguistic features. Concerning the relationship between lexical richness and other variables, there are no clear results. Some studies have shown correlations with certain measures, while others have not. The studies that have focused exclusively on different aspects of vocabulary do seem to have found a relationship, however. Such was the case in Lemmouh (2010) and Tidball and Treffers-Daller (2007, 2008). However, as previous studies have used tasks different from the one used in this study, we do not expect to obtain similar results. In the present study, in contrast to many of the earlier studies that used tests, it was deemed important to investigate spontaneous spoken data.

4. This study

4.1 Participants and task

Lexical richness and Lexical FSs will be analyzed in L2 French in three groups of Swedish-speaking informants. They will also be compared to a group of native speakers (see Table 2).

All informants were recorded within the InterFra project at Stockholm University. Group 1 consists of ten 4th and 5th term university students and doctorate students. They are advanced, semi-formal learners, i.e. they have learned French mostly in Sweden, but some of them have been to a French-speaking country for one–two years (mean Length of Residence, LoR: 1.4 years). They have been classified at stages 5 and 6 — the medium advanced and the superior advanced stages of the developmental continuum presented in Bartning and Schlyter (2004) — on the basis of morpho-syntactic traits. Groups 2 and 3 consist of Swedes who have moved to Paris and thus use French as a second language on
a daily basis. The ten ‘Juniors’ in Group 2 have lived in Paris for 5–15 years (mean LoR 6.7 years). Their ages on arriving in France were 18–19 years on average, and they were 25–32 years old when they were recorded. The ten ‘Seniors’ in Group 3 have lived in Paris for 15–40 years (mean LoR: 23 years). They were 40–55 years old at the time of the recordings, and came to France at approximately 21–24 years of age. The control group consists of ten native speakers of French. It is difficult to say whether there exist any differences in terms of linguistic proficiency between the two groups in Paris. A distinction can definitely be made between L2 learners in a foreign language setting (Group 1), L2 users in a second language context (Groups 2 and 3) and native speakers (Group 4, NS). It is quite obvious that both second language groups have been exposed to far more input than the learners in Group 1. If results from Groups 2 and 3 surpass those of Group 1, we find evidence in support for a stage 7, as mentioned in the introduction. This would also indicate that vocabulary measures are better at characterizing very advanced levels, than morphosyntactic criteria, at least as regards these data. Considering length of residence the literature has shown (e.g. Cummins 1981) that its effect tends to diminish after five years of residence in the target language community. However, since we are dealing with frequency-dependent phenomena, lexical richness and formulaic sequences, it is plausible to assume that there will be differences even between the two Paris groups, which is why it was decided to keep these two groups separated.¹ The fact that the learners in group 3 have spent more time in the target language community probably means that they have been exposed to more input generally. Thus, they should have had the possibility to add

¹ In Forsberg Lundell et al. (submitted), the participants in the two Paris groups are categorized differently, based on a listener test, which groups speakers into those who pass as native speakers of French and those that do not. The first group is labeled as ‘near-natives’, following Abrahamsson and Hyltenstam’s (2009) definition.
new words to their vocabulary to a larger extent than the learners in group 2. The longer time you spend in the TL environment the better the possibilities to encounter low-frequency words and extensive exposure also allows for collocational links to be strengthened.

All the informants performed the same task which was a 15–20 minutes long interview with a native speaker of French on subjects such as work/studies, spare time activities, family, etc. Studying the lexicon, it is important to be somewhat cautious when using such an ‘open’ task. Each informant’s life story has a non-negligible effect on the use of lexis, since the themes brought up might vary considerably. One may speak of ballet and another of advertising, while yet another may mainly talk about his/her family. It should also be noted that such tasks are probably easier than a more controlled task, since they give the speaker more freedom to talk about known topics. For example, when comparing the results from Forsberg (2008) and Forsberg and Fant (2010), it turns out that L2 speakers are more nativelike in their use of formulaic language in an interview than in a pragmatic role-play or a retelling task. The advantage is that we use the same type of interview with all the 40 speakers and we have the possibility to investigate spontaneous language use.

4.2 Methods

4.2.1 Identification and classification of lexical formulaic sequences

The issue of how to define and identify formulaic language is complex, probably due to the irregular and intuitive nature of formulaic language on the one hand, and to the large diversity of structures grouped under the label ‘formulaic’ on the other. The two most commonly used ways of identifying formulaic language in corpora are the statistical method and the phraseological method (cf. Granger & Pacquot 2008). Within the phraseological methodology, to be used here, the researcher identifies potentially formulaic/conventional sequences based on linguistic criteria related to syntactic, semantic and pragmatic restrictions.

The present study makes use of Erman and Warren’s (2000) original categorisation of prefabs (their term), which was slightly modified in Erman et al. (submitted), a model based on the phraseological tradition. In Erman et al. (submitted), formulaic sequences can be classified into Lexical and Qualifier FSs. In the present study, only Lexical FSs will be investigated. They incorporate at least one content word and are used for extralinguistic reference and denote actions (such as faire la fête ‘to party’), states (avoir peur ‘to be scared’), objects (année sabbatique ‘sabbatical year’) and so on. When it comes to the practical identification of Lexical FSs, Erman and Warren (2000) make use of the criterion restricted exchangeability. In order for a sequence to qualify as a prefab in their terminology, an exchange of one
of the words for a synonymous word must always result in a change of meaning or a loss of idiomaticity (Erman & Warren 2000: 32).

Forsberg (2008) proposed an operationalization of the restricted exchangeability criterion. The first step in identification is to manually identify, in the corpus, the Lexical FSs that meet the criterion based on researcher intuition. This is then complemented by searches on Google.fr. To test the extent to which restricted exchangeability applies to a sequence, an analogous sequence, which has been subject to one of the following modifications, is constructed.

- One of the words is exchanged for a synonymous word.
- One of the words is exchanged for an antonymous word (for example _ça marche mal_ 'it works badly' instead of _ça marche bien_ 'it works well').
- Change of article (from definite to indefinite or absence of article).
- Change of number (from plural to singular or vice versa).
- Change in word order (for example _égalité femmes/hommes_ 'equality women/men' instead of _égalité hommes/femmes_ ‘equality men/women').

For a sequence to be counted as formulaic, it has to appear at least twice as frequently on Google as any of the modified versions.²

### 4.2.2 Measuring lexical richness

Lexical richness is measured using a frequency-based method, the Lexical Oral Production Profile, developed in Lindqvist et al. (2011).³ The method is similar to the Lexical Frequency Profile (Laufer & Nation 1995), in that it assumes that frequency plays an important role in vocabulary acquisition, in the sense that the most frequent words of the language are acquired first, and that the less frequent words are acquired later in the acquisition process. However, as the learner data analyzed consisted of oral production, it was deemed necessary to create frequency bands that were also based on oral native speaker data. Thus, for French, the Corpaix corpus of spoken L1 French, which has been compiled at the Université de Provence (Campione et al. 2005), was used. The corpus contains 1 million words. For a detailed description of this corpus, see Pallaud and Henry (2004). A frequency list based on the Corpaix corpus, compiled by J. Véronis, is available online. The list consists of word forms. Lindqvist et al. lemmatized this list using

---

² Ideally, frequency bands for (lexical) formulaic sequences should have been used in order to match the lexical richness measure (cf. Section 4.2.2.). However, to our knowledge, such frequency bands are not available for French.

³ The method is currently being further elaborated, by integrating aspects other than frequency (Bardel et al. in press). However, as the first version of the method has proven to allow for separation between different proficiency levels, it will be used in the present study.
TreeTagger, which resulted in 2766 lemmas. On the basis of these, three frequency bands were created, each band containing approximately 1000 lemmas.\(^4\)

In order to calculate the lexical richness of a text, all the lemmas produced in an interview by each informant were compared to the frequency bands, which resulted in a lexical profile showing the proportion of lemmas (tokens) in the frequency bands. The lemmas that do not belong in any of the frequency bands are categorized as “off-list”, e.g. Band 1: 90%, Band 2: 2%, Band 3: 3%, Off-list: 5%. The general assumption is that a relatively high proportion of lemmas in Band 3 and Off-list indicates a rich vocabulary.\(^5\)

4.3 Research questions and hypotheses

The following research questions are posed:

1. Is it possible to discern lexical development between stages 5 and 6 (Group 1) and informants in a second language setting (Groups 2 and 3), in terms of lexical richness, on the one hand, and Lexical FSs, on the other? Based on these measures, is it reasonable to propose advanced stages beyond stage 6?

In view of the results of earlier studies (Bartning \textit{et al.} 2009; Forsberg 2010), where similar groups of informants were studied, it is hypothesized that we will find differences between Group 1 and Groups 2 and 3, in terms of Lexical FS use. It is also hypothesized that no differences will be found between the most advanced groups and the NSs, since this was the case for similar groups of learners and NSs in Forsberg (2010). However, more informants are included in the present study and some of the NSs are not the same. In Lindqvist \textit{et al.} (2011) it was found that the most advanced learners (stage 6) reached nativelike levels with respect to lexical richness. However, in that study, the NSs were students with no working

---

\(^4\) A one-million word corpus may seem small. However, it was not possible to have access to a larger corpus. It is difficult to find large corpora of spoken language, especially for French. Lindqvist \textit{et al.} (2011) estimated that \textit{Corpaix} was sufficient for the purposes of creating frequency bands, as it allowed them to create three frequency bands, each containing approximately 1 000 lemmas. This is to compare to the Lexical Frequency Profile (see www.lectutor.ca/vocabprofile), for which each frequency band contains 1 000 word families (for both English and French).

\(^5\) For both the formulas measure and the lexical richness measure, the counts are based on tokens. In future research, it would be interesting to also base the counts on types to see if the results will be different. As for the lexical richness measure for instance, some of the most frequent words are also often used several times in a conversation, which means that words in Band 1 such as \textit{je} or \textit{et} may increase the number of tokens in that frequency band, whereas infrequent words normally also have fewer occurrences.
experience. In the present study, the control group consists of NSs with some or many years of professional experience. This leads us to believe that they will have a more developed vocabulary than the NSs in the previous study. Our general assumption is that vocabulary is an aspect that learners constantly have the possibility to develop during the acquisitional process. As for the learner groups in the present study, there are important differences in exposure to the target language, and we expect that this will be reflected in their lexical richness. We hypothesize that Group 1 will have a less developed vocabulary, revealed by the relatively low proportion of low-frequency words in their oral production, than the learners who have moved to Paris and stayed there for several years. Given the differences in length of residence between Groups 2 and 3, we hypothesize that Group 3 will have the more developed vocabulary.

2. Is there a correlation in the development between the two vocabulary measures, i.e. lexical richness and Lexical FSs?

We hypothesize that Lexical FSs will be related to development of lexical richness since they are both dependent on frequency of exposure. In addition, earlier results show that both measures are promising tools for distinguishing between levels.

4.4 Results

4.4.1 Lexical formulaic sequences

Table 2 shows the results for Lexical FSs in the different groups. The results refer to total number of tokens of Lexical FSs per 100 words. All differences have been calculated through One Way ANOVA analysis with Tukey-Kramer post-test.

As regards Lexical FSs, significant differences can be found among several groups but not all. Firstly, a difference ($p < 0.05$) is found between Groups 1 and 2, which is in line with the results presented in Bartning et al. (2009). The least advanced learners (Group 1) also differ significantly from the NSs ($p < 0.0001$), which comes as no surprise. However, they do not differ significantly from Group 3, which is rather difficult to interpret. If Lexical FS use should be considered as

<table>
<thead>
<tr>
<th>Group</th>
<th>LFS/100 words (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.Stage 5+6</td>
<td>1.9 (SD 0.9)</td>
</tr>
<tr>
<td>2.Juniors</td>
<td>3.4 (SD 1.0)</td>
</tr>
<tr>
<td>3.Seniors</td>
<td>2.8 (SD 0.8)</td>
</tr>
<tr>
<td>4.Natives</td>
<td>4.1 (SD 1.0)</td>
</tr>
</tbody>
</table>
a measure of proficiency, Group 2 (Juniors) would then be more advanced than Group 3 (Seniors), since Group 2 have a significantly higher score than Group 1 but Group 3 does not. In addition, there is no difference between Group 2 and Group 4 (NS), whereas there is a significant difference ($p < 0.05$) between Group 3 and 4. However, the differences between Group 2 and 3 are not significant, but there is clear tendency for Group 2 scores to be higher. Since the number of participants is limited, it is possible that individual differences give rise to these somewhat puzzling results. As Table 1 shows, the standard deviations are fairly high in all groups. As will be discussed in Section 4.3, this leads us to view formulaic sequence use as not only a function of frequency of input, but also related to other factors (of socio-psychological orientation, as suggested by Dörnyei et al. 2004). All in all, the results allow us to suggest that formulaic language is a good measure of advanced L2 development. It succeeds in separating stages 5 and 6 from a more advanced stage 7, thus lending support for an elaboration of Bartning and Schlyter’s (2004) advanced stages. In addition, it shows that it is possible to approach nativelike performance on this measure (which regards quantity of Lexical FSs), which is the case for Group 2.

4.4.2 Lexical richness

Table 3 shows the lexical profiles of the groups, i.e. the mean proportions of lemmas (tokens) in the different frequency bands. What is of interest here is the proportion of low-frequency words used, i.e. the proportion of Band 3 and Off-list, which presumably indicates lexical richness (cf. Laufer & Nation 1995).

As Table 3 shows, the proportions of low-frequency words (Band 3+off-list) seem to reflect the proficiency level of the learners in the sense that Group 1 has the lowest proportion (3.86%) and Group 3 the highest proportion of the learner groups (5.25%). The NSs have the highest proportion (7.27%). In order to examine whether these differences were statistically significant we ran an ANOVA. The results showed that there were significant differences between Groups 1 and 3 ($p < 0.01$), and between Group 1 and the NSs ($p < 0.001$). However, no significant differences were found between Groups 1 and 2. This result does not coincide

<table>
<thead>
<tr>
<th>Group</th>
<th>Band 1</th>
<th>Band 2</th>
<th>Band 3</th>
<th>Off-list</th>
<th>Band 3 + off-list</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.Stage 5+6</td>
<td>94.11</td>
<td>2.02</td>
<td>0.69</td>
<td>3.24</td>
<td>3.86 (SD 0.60)</td>
</tr>
<tr>
<td>2.Juniors</td>
<td>93.65</td>
<td>1.84</td>
<td>0.80</td>
<td>3.71</td>
<td>4.52 (SD 1.07)</td>
</tr>
<tr>
<td>3.Seniors</td>
<td>92.74</td>
<td>2.01</td>
<td>0.75</td>
<td>4.51</td>
<td>5.25 (SD 1.55)</td>
</tr>
<tr>
<td>4.Natives</td>
<td>90.02</td>
<td>2.72</td>
<td>1.12</td>
<td>6.15</td>
<td>7.27 (SD 0.72)</td>
</tr>
</tbody>
</table>
with the results concerning Lexical FSs, where there were significant differences between precisely these two groups. The fact that Group 2 seems to have a vocabulary that is similar to that of Group 1 would indicate that Group 2 has not developed beyond stage 6 with respect to lexical richness.

Significant differences in lexical richness were also found between Group 2 and the NSs ($p < 0.05$). Interestingly, there are no significant differences between Group 3 and the NSs. This would indicate that the lexical richness of this very advanced group is similar to that of the NS group. This issue was also brought up in Lindqvist et al. (2011). They compared learners at stage 6 with NSs and found no significant differences in the proportion of low-frequency words. They interpreted this as suggesting that the learners of that study actually had a lexical richness similar to that of NS. However, in that study, the very advanced Paris-based Groups 2 and 3 were not investigated. As we have seen in the present study, these groups, especially Group 3, display more advanced vocabularies than Group 1. In light of the new data analyzed in the present study, we would argue that the fact that Lindqvist et al. did not find significant differences between stage 6 and NSs is probably because they used a different group of NSs: students at Stockholm University with very limited working experience. In the present study, the NSs all have several years of working experience, which plausibly gives them the opportunity to come into contact with a more varied vocabulary with different kinds of specialized terminology for instance. This is certainly reflected in their lexical richness. Some examples of low-frequency words used by the native speakers in the present study are *électromécanique* ‘electromechanical engineering’, *interstellaire* ‘interstellar’, *maquette* ‘scale model’, which definitely belong to a specialized vocabulary related to their profession. These results bring attention to the difficulties involved in choosing appropriate NS control groups as benchmarks for L2 performance.

The fact that Group 3 has a higher proportion of low-frequency vocabulary than the other learner groups is in line with our hypothesis that vocabulary is an aspect that develops continuously over time. The Seniors have lived in the target language environment for a long time and have consequently received a lot of input, plausibly of words that are not easily encountered in a formal setting. It seems that frequency of exposure actually plays a role in the acquisition of low-frequency words, since Group 3, with a considerably longer mean LoR, has a higher proportion of this category of words, than Group 2. Examples of such words are *entente* ‘agreement’, *clivage* ‘divide’, *bourrer* ‘cram full’, *attache* ‘tie’, *déclic* ‘trigger’, *estomper* ‘blur’, *intoxiquer* ‘brainwash’, *poireauter* ‘hang about’.
4.4.3 Correlation

Our second research question concerns the correlation between lexical richness and Lexical FSs: do they develop at the same rate in the individual learner? It was hypothesized that a measure of Lexical FSs would correlate with that of lexical richness, since both features are dependent on frequency. It has already been shown that the two measures display divergent results in group comparisons, but we also wanted to run a Spearman rank correlation between the two measures. The following result was obtained: \( (n = 30) \ r = 0.34, p = 0.06 \), suggesting that we are dealing with a weak correlation which is not quite significant. Our hypothesis is consequently not confirmed, but neither can it be completely rejected. Given the low number of participants, it cannot be excluded that we are actually dealing with a positive correlation. If we look at the separate results for the formulaic sequence measure and the lexical richness measure, we see that they both yield group differences, but not necessarily the same ones and whereas the formulaic sequences measure yields differences between Groups 3 and 4, the lexical richness measure does not. This would imply that lexical richness is actually something that develops continuously over time and that the long residence of Group 3 actually has an effect for this measure. But this does not, curiously enough, seem to be the case for formulaic language, since Group 3 does not produce more nativelike results than Group 2. It appears, then, that formulaic sequence use is not only dependent on frequency but also on other factors. It has been proposed that the use of formulaic sequences is dependent on factors such as motivation, aptitude and socio-cultural integration (Dörnyei et al. 2004; Kecskes 2002), which suggests that it is not a phenomenon that follows a linear progression but is dependent on individual differences. Given the limited number of participants and the spontaneous nature of the oral task at hand, these results should, however, be interpreted with caution.

5. Conclusions

In the present study we analyzed two aspects of vocabulary in advanced learners of French: lexical richness and Lexical FSs. We compared three learner groups — one group of foreign language learners and two groups of second language learners with different lengths of residence in the target language country — and one group of native speakers. As regards Lexical FSs, the results showed that there were differences between Groups 1 and 2, which suggests that the Group 2 learners have moved beyond stage 6 as regards formulaic sequence use. However, this is not the case for Group 3. Whereas the latter group performed in a nativelike way on the lexical richness measure, as regards Lexical FSs they do not. However, Group 2 does not differ significantly from the NS in terms of Lexical FSs. This means that
Group 3 is to be considered the most advanced group as regards lexical richness, whereas Group 2 is the most advanced with respect to formulaic sequences. The analysis, however, reveals that, for both measures, it is possible to attain nativelike levels (relatively high proportion of low-frequency lemmas vs. tokens of Lexical FSs).

For lexical richness, the proportion of low-frequency lemmas, the results indicate that it develops in the expected way: the less advanced learners had the lowest proportion, while the most advanced learners had the highest proportion. However, there were no statistically significant differences between Groups 1 and 2. This result seems to indicate that Group 2 has not moved beyond stage 6 in terms of lexical richness. In contrast, this seems to be the case for Group 3, which approaches the native speakers as there were no significant differences between these groups. We interpreted this as a possible effect of the amount of exposure to the target language and as a confirmation of our hypothesis that vocabulary is an aspect that constantly develops.

The results for both measures indicate that there seems to be evidence of an additional stage beyond stage 6, since both lexical richness and formulaic sequences continue to develop, albeit differently for the different groups. The new measures used in the present study can consequently describe linguistic development beyond stage 6, which morphosyntactic measures have failed to do (cf. Bartning et al. 2009). However, the present study does not clearly indicate parallel development between the two aspects of the lexicon investigated. A correlation analysis was conducted and no strong, significant correlation was obtained. The fact that different aspects of vocabulary do not always develop in parallel has also been noted in e.g. Lemmouh (2010) and Mizrahi and Laufer (2010) (cf. Section 3).

In future research, we will further investigate the relationship between different aspects of vocabulary in advanced learner French. One of the shortcomings of the present study was the low number of informants in each group. More data has recently been collected from advanced learners performing various vocabulary tasks. These data will hopefully further our understanding of the lexical competence of the advanced learner in L2 French and contribute to the characterization of the advanced learner variety.

Acknowledgements

We would like to thank two anonymous reviewers for their very useful comments on earlier versions of this paper.
References


Mizrahi, E. & Laufer, B. (2010). Lexical competence of highly advanced L2 users: is their collocation knowledge as good as their productive vocabulary size? Paper presented at Eurosla 20, Reggio Emilia, Italy.


**Résumé**

Dans son survol de recherches sur l’apprenant avancé, Bartning (1997) a pour but de caractériser la variété avancée. Cette caractérisation est surtout basée sur des traits morpho-syntaxiques. Notre objectif est d’enrichir la caractérisation des apprenants avancés par la prise en compte de traits supplémentaires, mais aussi de décrire des niveaux supérieurs, à partir des recherches récentes sur le français L2. Plus précisément, nous nous posons la question de savoir si deux dimensions lexicales, à savoir la richesse lexicale et les séquences préfabriquées lexicales, pourraient éventuellement distinguer des niveaux dans cette variété et, ainsi, contribuer à enrichir sa caractérisation. De plus, sera examinée ici la corrélation entre ces deux variables. S’agit-il de deux aspects qui se développent en parallèle ou non?