Analysing lexical richness in French learner language: what frequency lists and teacher judgements can tell us about basic and advanced words

Article

Accepted Version

Analyzing lexical richness in French learner language: what frequency lists and teacher judgements can tell us about basic and advanced words[1]

Françoise Tidball and Jeanine Treffers-Daller (UWE Bristol)

[1] We would like to thank Kate Beeching, Jean-Yves Cousquer, Annie Lewis, Gareth Lewis and John Tidball for their help in collecting and/or transcribing the data, the tutors who provided the judgements on the vocabulary items, Brian Richards for his guidance on the reliability analyses of the data and his detailed comments and two anonymous reviewers for their valuable suggestions.

Abstract

In this paper we study different aspects of lexical richness in narratives of British learners of French. In particular we focus on different ways of measuring lexical sophistication. We compare the power of three different operationalisations of the Advanced Guiraud (AG) (Daller, van Hout and Treffers-Daller, 2003): one based on teacher judgement, one on ‘le français fondamental 1er degré’ and one on frequency of lexical items. The results show that teacher judgement is a highly reliable tool for assessing lexical sophistication. The AG based on teacher judgements is better able to discriminate between the groups than the other operationalisations. It also works better than Vocabprofil (the French version of Laufer and Nation's (1995) Lexical Frequency Profile).

1. Introduction

In this paper we aim to come to a better understanding of a particular aspect of lexical richness, namely lexical sophistication, in learner language of British learners of French. As is well-known, lexical richness is a multidimensional feature of written or spoken language. Read (2000: 200) distinguishes four dimensions of lexical richness, and one of these is lexical sophistication, which he defines as 'the use of technical terms and jargon as well as the kind of uncommon words that allow writers to express their meanings in a precise and sophisticated manner.' The key question is of course how to operationalise what counts as a sophisticated word or expression. Many measures of lexical richness are based on the assumption that the key factor behind the difficulty of a lexical item is its frequency. Laufer and Nation's (1995) Lexical Frequency Profile, for example, is based on the assumption that frequent words are easier than infrequent words. This view is shared by Vermeer (2000) and Meara and Bell (2001). Similarly, Malvern, Richards, Chipere and Durán (2004: 3) define lexical sophistication as the appropriate use of low frequency vocabulary items.

The question is however whether frequency is the only dimension that counts. The psycholinguistic literature shows that the cognate status of items is also an important factor in processing. Cognate items, i.e.,

translation pairs in which the words are similar in sound and spelling are processed faster than non-cognates (Van Hell and De Groot, 1998: 193) confirming that, as one would predict, cognates are easier than non-cognates.[2] For British learners of French therefore, many infrequent items are easy, because the French and English translation equivalents are cognates, e.g. French *détester* 'to detest', which is infrequent but probably highly transparent to learners.

Support for the fact that cognates play an important role in L2 acquisition comes from Laufer and Paribakht (1998) who demonstrate that French-speaking ESL students obtain higher scores on a test of English controlled active vocabulary than learners with no French because of the large number of French-English cognates. In a similar vein, Horst and Collins (2006) in their study of the longitudinal development of French learners of English in Canada show that learners initially prefer certain low frequency items such as *respond* over the high frequency alternative *answer*, because the former is a cognate of the French translation equivalent *répondre*.

[2] We have adopted the psycholinguistic definition of cognates given in Van Hell and De Groot (1998: 193) rather than a historical linguistic definition of cognates because many speakers will not know whether the words in French and English are derived from the same source.

Experienced teachers, in putting together textbooks or other material for learners will therefore not rely *solely* on the frequency of lexical items, but they will use additional criteria, such as the cognate status of items, in judging which words learners need. It is interesting to find out to what extent teacher judgements can help us to get a better understanding of the lexical richness of learner language.

We assume that measures of lexical sophistication which involve teacher judgement are better than those that are solely based on frequency. More specifically, the key hypothesis of the current article is that measures of lexical richness which are based on a basic vocabulary list which is derived from teacher judgements should be better able to discriminate between groups than measures that are based on a basic vocabulary list which consists of the most frequent words or on a traditional basic vocabulary list such as *Français Fondamental Premier Degré* (FF1). The latter is based on a variety of criteria and contains several items that are now out of date (see Tidball and Treffers-Daller, 2007).

We test the key hypothesis of this article by investigating how different operationalisations of the concept of basic vocabulary affect the power of a measure of lexical sophistication which was recently proposed by Daller, Van Hout and Treffers-Daller (2003), the Advanced Guiraud (AG) (see below for a description). The current paper is a

follow-up to Tidball and Treffers-Daller (2007) in which we found that the AG was less able to discriminate between groups than measures of lexical diversity that do not make use of external criteria such as a basic vocabulary list. The choice of a different basic vocabulary can possibly improve the performance of the AG.

We will compare different operationalisations of the concept of basic vocabulary by calculating AG in a variety of ways. These results will subsequently be compared with those obtained with the help of Vocabprofil, the French version of Nation's Range programme which gives a Lexical Frequency Profile of texts (Laufer and Nation, 1995). We will also establish how these measures compare with measures of lexical diversity, which do not make reference to any external criteria or lists, such as the Index of Guiraud (Guiraud, 1954) and *D* (Malvern et al, 2004).

Before going into the details of the current study, we will first present different ways of measuring lexical sophistication (section 2), a brief appraisal of recent work on word frequencies in French and how this relates to the operationalisation of the concept of a basic vocabulary (section 3). In section 4 we present the methodology of the current study, and section 5 gives an overview of the results. Section 6 offers a discussion of the results and a conclusion.

2. Measuring lexical sophistication

We agree with Meara and Bell (2001) that it is important to assess the quality of vocabulary used by L2 learners by making reference to external criteria, such as basic vocabulary lists or frequency lists of lexical items, if one wants to gain a better understanding of lexical richness. As Meara and Bell's (2001: 6) now rather famous examples (1) to (3) show, measures of diversity which are based on distribution of types and tokens in a text will produce the same result for each of these examples.

(1) The man saw the woman
(2) The bishop observed the actress
(3) The magistrate sentenced the burglar

These three sentences are however quite different in the quality of the vocabulary used, as the words in (1) are less difficult (and more frequent) than those in (2) and (3). As Malvern et al (2004: 124) notice, the dimensions of diversity and rarity (sophistication) are of course linked, because 'over a longer stretch of language, diversity can only increase by the inclusion of additional different words, and the more

these increase, the more any additional word types will tend to be rare.' The question is therefore whether the development of lexical resources in L1 or L2 learning is due to an increase in the number of low frequency words, or whether the children or students make better use of a wider range of higher frequency words.

Hayes and Ahrens (1988; in Malvern et al, 2004) and Laufer (1998), found that the percentage of low frequency vocabulary did not increase in the spoken or (free active) written data of their informants. Recently, Horst and Collins (2006) have shown that 11- and 12-year-old francophone learners of English in Québec do not use a higher number of low frequency words after 400 hours of tuition, but a larger variety of high frequency words (up to k1 layer), and they draw less upon cognates (see above). This illustrates most clearly that other factors, such as the cognate status of items, in addition to frequency, play an important role in lexical development, and that frequency bands to which vocabulary items belong do not always provide a good indication of students' progress.

In the present study we use the Advanced Guiraud (AG), as proposed by Daller et al (2003), to measure the differences in lexical sophistication in the speech of British learners of French and a French native speaker control group. The AG is derived from the Index of Guiraud (Guiraud, 1954), which is the ratio of types (V) over the square

root of tokens (N) as expressed in the following formula ($V/\sqrt{N}$). The AG is ratio of advanced types over the square root of the tokens ($V_{adv}/\sqrt{N}$). For its calculation, one needs to distinguish basic and advanced vocabulary and in this paper we do this in three different ways; with the help of a) the traditional *Français fondamental premier degré* (FF1); b) a list of basic words based on teacher judgements; c) a list of basic words derived from the Corpaix frequency list (Véronis, 2000).

As teacher judgements of the difficulty of words were a reliable tool in measuring lexical richness among Turkish-German bilinguals (Daller et al, 2003), the second operationalisation is based on teacher judgements. A list of the frequency of words in a corpus of spoken French, the Corpaix oral frequency list (Véronis, 2000), forms the basis of the third operationalisation. We wanted to find out whether frequency lists are able to successfully capture words which intuition tells us belong to basic vocabulary. We believe with Gougenheim, Rivenc, Michéa and Sauvageot (1964: 138) that 'Ils [les mots concrets] semblent se dérober à la statistique' ('concrete words seem to escape statistics'). A comparison of basic vocabulary lists based on frequency with those based on teacher judgement may well be able to shed new light on the validity of this claim.

We compare different operationalisations of the AG with a well-known measure of lexical diversity, *D* (Malvern et al, 2004), which represents the single parameter of a mathematical function that models the falling TTR curve (see also Jarvis, 2002 and McCarthy and Jarvis, 2007 for an appraisal of this measure). The different measures can give us an indication to what extent the students from the three groups differ from each other in the quantity and/or in the quality of the vocabulary they use.

Finally, we compare these results with those obtained with the help of the frequency bands in Vocabprofil. The output of the programme is a Lexical Frequency Profile (Laufer and Nation, 1995), which gives the frequency of words according to the following four frequency layers: the list of the most frequent 1000 word families (K1), the second 1000 (K2), the Academic Word List (AWL) and words that do not appear on the other lists (NOL). Laufer (1995) shows that a condensed version of the LFP, which distinguishes between the basic 2000 words and the 'beyond 2000' words can also be used to measure lexical richness across different levels of proficiency.

The frequency data on which Vocabprofil is based are derived from a written corpus (see below), but Ovtcharov, Cobb and Halter (2006) claim that Vocabprofil can be used to analyse vocabulary in oral data. It would be useful to know to what extent the profiles for oral and written

data as produced by Vocabprofil differ but the authors do not offer such a comparison. They do however show that the profiles of advanced Canadian learners of French are not significantly different from those of Beeching's corpus of oral data from French native speakers, which is also freely available on the internet. Vocabprofil differs from the LFP in that the third frequency layer (K3) contains words which occur at a frequency of 2001-3000 in the corpus, French having no equivalent to the Academic Word list (Cobb and Horst, 2001).

For the purposes of the current article it is important to note that Beeching's corpus contains a very high proportion of NOL words, namely 10.87%, and the same is true for the learners in Ovtcharov et al's study: their scores for the NOL category range from 4.02% for learners at the lowest level to 8.71% for learners in the top group. As the percentages are so high, it is likely that the NOL category does not only contain exceptionally rare words, but also many words that may be frequent in spoken language but not in written language, and which therefore do not occur in the written corpus on which the frequency profiles are based. Whether or not this is the case in our data will be investigated below.

3. Basic vocabularies and word frequencies in French

Until recently, the only existing basic vocabulary list was *Le Français fondamental premier degré* (Gougenheim, 1959) which has been widely used as a reference in many studies on vocabulary. *Le Français fondamental premier degré* (FF1) is largely based on an oral corpus. This list is not solely based on frequency: to the most frequent words were added 'available words', i.e. common words (e.g. *fourchette* 'fork', *chocolat* 'chocolate', *autobus* 'bus') which did not appear in the corpus because they are topic-specific and therefore have a lower frequency, but were frequently mentioned in additional surveys on specific themes, or were deemed essential for teaching French as foreign language.

Given the importance of lexical frequency in language processing, a selection of the most frequent words is an obvious alternative to FF1. We chose to use the Corpaix frequency list for oral French (Véronis, 2000) as this frequency list of 4,592 tokens is based on oral data, and it is freely available on the internet (in unlemmatized form). The corpus of one million words from which the list was derived is based on 36 hours of recordings of interviews held in real-life situations, collected over 20 years at the Université de Provence (now part of the DELIC team).

Because the list is drawn from a relatively limited corpus, some contexts are clearly represented more than others. A few examples can

illustrate that there are some unexpected results in this list. In the corpus *orthographe* 'spelling' occurs 235 times, and it has a higher frequency than *finir* 'to finish', regarder 'to look at', and *bactérie* 'bacteria', which occur 144 times. It is also surprising that *fromage* 'cheese' and *pipette* 'pipette' occur with the same frequency (16) and that *fromage* 'cheese', which appears in FF1, does not feature in the first 1000 words of Corpaix (see appendix for more examples).

In addition we used the frequency profiles that can be obtained with Vocabprofil. The frequency information on which the programme is based stems from a corpus of 50 million words (Verlinde and Selva, 2001) from two newspapers *Le Monde* (France) and *Le Soir (*French-speaking part of Belgium).

It is doubtful whether information about frequency of lexical items in written texts can be used for an analysis of oral data, because of the discrepancies between spoken and written French, but we felt it was interesting to see whether this tool is able to uncover the differences in lexical richness between our three groups, what percentage of the students' tokens belongs in the category NOL (not-on-lists) and whether the lexical frequency profiles are better able to discriminate between the groups in this study.

4. Methods

The participants were two groups of British undergraduates studying French as part of a Languages Degree at the University of the West of England (UWE), Bristol - 21 level 1 (first year), 20 level 3 (final year) - and a control group of 23 native French speakers, also students at UWE. All students undertook the same task under the same conditions: they were asked individually to record their description of two picture stories presented as cartoon strips of 6 pictures each (Plauen, [1952] 1996]). The corpus contains 23,332 tokens (January 2008).

The general language proficiency of each participant was measured by means of a French C-test which provided a useful external criterion against which the different measures could be validated. This test was highly reliable (Cronbach's alpha = .96, 6 items).

The data were transcribed and coded in CHAT, lemmatized and analysed using CLAN (MacWhinney, 2000). More details on the informants, the C-test, the transcription and the lemmatization are given in Tidball and Treffers-Daller (2007).

For this project we operationalised the concept *basic vocabulary* in three different ways. First of all we used a list based on frequency, availability and judgement (FF1); second, a list based on oral frequency

(Corpaix); and third, an intuition- based list (judgements of teachers). We will present each briefly here.

We used FF1 as our first operationalisation, even though this list is rather old and contains several items which relate to rural life in France, such as *charrue* 'plough' and *moisson* 'harvest', which are probably no longer part of the basic vocabularies of speakers living in cities.

Our second list is based on the non-lemmatized oral frequency Corpaix list (Véronis, 2000). This list contains many elements that are typical for spoken data, such as the interjections *euh* (20,897), *ben* (2,936) or *pff* (298). It also unfortunately splits up words that contain an apostrophe, such as *aujourd'hui* 'today' into two words, giving a frequency of 261 for each part.

Homographs constitute another issue: *voler* 'to steal/to rob' (in our data) also means 'to fly' (a bird/ aeroplane), *vol* 'theft/robbery' or 'flight'. The frequency lists on which Vocabprofil is based (see below) do not differentiate between the two meanings and the frequency rank of the word is therefore not entirely meaningful. FF1, on the other hand, gives the two different meanings of *voler* under two different entries. VocabProfil lists *vol* in K1 (first thousand words), *voler* in K2 (1100-2000) and *voleur* 'thief/robber' is NOL (beyond the first 3000 words). The latter is listed in FF1.

As we wanted to compare the results based on the Corpaix list to those based on FF1 (which only contains lemmas), we needed to lemmatize this list. The lemmatization gave us a frequency list of 2767 types. The methodology followed for the lemmatization can be found in Tidball and Treffers-Daller (2007).

For our third basic vocabulary list we used the judgement of three experienced tutors of French, two of whom were French native speakers, and one was a bilingual who had grown up with English and French. They were given a list of all 932 types produced by our learners and asked to rank them on a scale on 1 to 7, according to how basic or advanced they judged them to be, with 1 being the most basic and 7 the most advanced. A reliability analysis showed the raters' judgements correlated almost perfectly with each other (Cronbach's Alpha = .943 (N=3). Two weeks later we asked the tutors to give us a second judgement of a random sample of 10 percent of these judgements, which enabled us to carry out a test-retest reliability analysis. The scores given to each item by individual judges in the first and the second rounds correlated strongly and significantly with each other for the first two judges (r = .88 and .84) and significantly but less strongly for the third judge (r =.56). We also calculated Cohen's kappa to establish to what extent raters agree on what constitutes a basic and a non-basic word in the two rounds. Agreement turned out to be substantial for raters one

and two (k = .624 and .601; p <.001) but only fair for the third rater (k =.252; p < .001). The latter was therefore excluded from further calculations.

We defined our basic vocabulary as follows: First we totalled the scores given by the two remaining raters. Then we selected all words which obtained total scores in the lower quartile (i.e. scores of 4 or less out of a possible 14) for our basic vocabulary list. This gave us a list of 246 basic words.

In order to make a comparison between the different operationalisations possible, we used the Corpaix frequency list to create three different basic vocabulary lists: the first one (Corpaix 246) contained the same number of words as the judges' file (246 words), the second one (Corpaix 1378) contained the same number of words as FF1 (1,378), and the third (Corpaix 2000) corresponded to Laufer's Beyond 2000 measure (i.e. it contained the 2,000 most frequent words in the list).

5. Results

In this section we first present the results of the C-test, to show how the language proficiency of the three groups differs on a measure that is independent of the story telling task. We then discuss to what extent the

different operationalisations of the concept *basic vocabulary* overlap (section 5.2) and in section 5.3 we will present the results of the analysis of lexical sophistication using different measures based on those basic vocabularies.

5.1 The C-test

The C-test results demonstrate that there are significant differences between the French proficiency of the two learner groups and the native speakers (ANOVA, F (df 2,61) = 105.371. $p < .001$). The Tukey post hoc test shows that all groups are significantly different from each other. This information is important as one would expect that measures of vocabulary richness should be able to demonstrate the existence of such a clear difference between the learner groups and between learners and native speakers. The power of the C-test to discriminate between groups turned out to be very high as can be seen in the Eta^2 of .776 (see section 5.3 for more details on Eta squared).

5.2 The overlaps between the basic vocabularies

Before going into a discussion of the different measurements of lexical sophistication, it is interesting to see to what extent the different basic vocabularies overlap. No two lists, even those drawn from very large corpora of similar origin, will overlap completely. Comparisons of the

first 1000 words of three existing frequency lists derived from different large French literary corpora, of which the *Trésor de la langue Française* (TLF) (INALF, 1971) is one, showed that they had 80% of words in common, whereas *le Français fondamental* had 65% in common with TLF (Picoche, 1993).

We have used CLAN to compare the content of FF1 with two other operationalisations: the Corpaix oral frequency list and the basic vocabulary list which is based on the judgements of the teachers. As FF1 contains 1378 words, we compared the first 1378 words of the Corpaix oral frequency list with FF1, and found that 725 words (52.6 percent) of FF1 are also found in the first 1378 words of the Corpaix frequency list. The judges' file shares 236 words (95.9 percent of the 246 words it contains) with FF1.

Subsequently, we entered all different operationalisations into Vocabprofil, to find out what percentage of the words in FF1, the Corpaix list and the judges' file belongs in the different frequency bands distinguished by Vocabprofil. The results of these analyses can be found in Table 1. It shows that FF1 and the first 1378 words of Corpaix have roughly similar profiles, with approximately 60 percent K1 words, whilst the first 2000 words of Corpaix contains almost 50 percent K1 words. The judges' file and the first 246 words in Corpaix contain a far larger proportion of K1 words: respectively 89 and 94 percent.

TABLE 1 HERE

The percentage of words that does not appear in any list is very high for FF1 (21.9 percent) and Corpaix (14.44 percent) and this is probably due to the fact that Corpaix contains many elements that are frequent in spoken language but which are not found in the written corpora.

Examples of words from our corpus which are NOL - apart from the many interjections mentioned in section 4 - are nouns such as *chapeau* 'hat' and *voleur* 'thief', an adjective such as *gentil* 'nice' and verbs such as *nager* 'to swim' and *repartir* 'to set off again'. The problem, from our perspective, is that Vocabprofil puts these very common words (all of which are in FF1) in the same category as *bousculer* 'knock down' and *canne* 'walking stick', which are highly specific and very infrequent, and which are not found in FF1 or Corpaix.[3] This illustrates the difficulty of using a written frequency list for the analysis of oral data. It is very unlikely that these words would all be classified in the same category if Vocabprofil was based on an oral frequency list.

[3] Corpaix does list CANNES, but this is presumably the city, and not the plural form of *canne* 'walking stick'.

5.3 Measures of vocabulary richness

Table 2 gives an overview of the results obtained for our different measures of vocabulary richness, including list-free measures. All measures show that there are significant differences between the groups in the vocabulary used. The smallest basic vocabularies – defined by the judges or the Corpaix 246 list - were successful at differentiating between all groups. The AG (judges) and the AG (Corpaix 246) yielded significant differences between the two learner groups (level 1 and level 3) but the AG (FF1)[4] and AG (Corpaix 1378) or AG (Corpaix 2000) did not.

TABLE 2 HERE

[4] In Tidball and Treffers-Daller (2007), the results for the AG (FF1) were marginally significant, but in the current study, with four more informants, this was no longer the case. As two of these additional learners had C-test scores well below the Mean of 75, this may be an indication that these learners are rather weak. The person with the lowest C-test score also had an exceptionally low score for advanced words, namely 3, which is more typical of the level 1 group.

In order to find out which measure is best able to discriminate between groups, we calculated Eta^2. Eta squared is the percentage of the variance in the dependent variable that can be accounted for by the independent variable (i.e group membership in this case). As Table 2 shows, the AG (judges) obtains a higher *Eta*^{2} than the Index of Guiraud. The *Eta*^{2} for the AG (judges) and D are virtually identical. The AG (judges) compares very positively with the other operationalisations of the AG, including its closest ally, the AG (Corpaix 246). This clearly shows that using teacher judgement is a better way to obtain a basic vocabulary list than using frequency data.

We also submitted the data to Vocabprofil, to find out to what extent the frequency layers as distinguished in Vocabprofil could help to distinguish our three groups (see Table 3).

TABLE 3 HERE

The results given in Table 3 show that, as one might expect, the native speakers make less use of words belonging to the highest frequency layer than the level 3 learners, and the latter use fewer words from the highest frequency layer than level 1 learners. The percentages are

comparable to those of Beeching's corpus, although a smaller proportion of the words used by Beeching's informants belonged to the K1 category (83.99 percent), and more of their words are NOL (12.07 percent). This is probably due to the fact that our informants did not produce free speech but they all narrated the same two stories, which limits the choices for the informants. Of the words used by Beeching's informants 3.94 percent belonged to the K2 layer and 1.2 percent to the K3 layer, and these percentages are very similar to those for our informants.

All three groups differ significantly from each other in their use of K1 words and also in their use of NOL words (see ANOVA/Tukey post hoc in Table 3). It is interesting that the groups do not differ from each other with respect to the K2 and K3 layers of Vocabprofil. While level 3 students use more words from the K2 layer than level 1 students, these differences are too small to become significant. As for the K3 layer, the students seem to use even fewer words of this frequency layer than the level 1 students.

Table 3 shows the effect size of the measurement of the vocabulary used at each frequency layer distinguished by Vocabprofil. It clearly demonstrates that the choice of the words which are *not* in the frequency list is the most powerful indicator of the differences between the groups, but the Eta^2 of the scores obtained on the basis of Vocabprofil are

clearly lower than those obtained with the help of the basic vocabulary lists (see Table 2).

The Eta2 obtained for the C-test outshines all the results found for the lexical richness measures, as the C-test obtained an Eta^2 of .776. A simple C-test may therefore well be a more effective way to distinguish language proficiency levels among learner groups.

The use of cognates by learners also conveys important information that needs to be taken into account in analyses of vocabulary richness. In our data, for example, speakers from all groups describe the thief in the second story most often as a *voleur*, but the learners' second most popular word for this character is the cognate *criminel* 'criminal', which is not used at all by the native speakers, even though it can be used as a noun in standard French. This word belongs to the K2 layer in Vocabprofil, and it is not listed in the Corpaix oral frequency list at all. Students who use this word would get higher scores on Vocabprofil or on the AG (corpaix) as these measures are exclusively based on frequency. The use of cognates is however not necessarily an indication that the speaker possesses a rich vocabulary. Rather, it shows that the speaker knows how to strategically exploit similarities between languages in telling a story. How we can account for the strategic use of cognates in the context of studies on vocabulary richness therefore deserves to be investigated further.

6. Discussion and conclusion

The results presented in section 5 clearly show that the power of a list-based measure such as AG depends on the way in which researchers operationalise the concept of basic vocabulary (see also Daller and Xue 2007, who make a similar point). We found that an operationalisation based on teacher judgements was more powerful than different operationalisations based on frequency, in that the AG (judges) was better able to discriminate between the groups. It is possible though that the performance of the AG based on frequency data can be improved if alternative frequency lists are being used. The Corpaix frequency list may not have been ideal for the current purposes, as it was drawn from a relatively small corpus. Alternative frequency lists based on the CRFP or *Lexique* could be used in future studies on this topic.

The results of the Vocabprofil analyses show that the students make better use of a range of relatively easy words. The learners used more NOL words at level 3, such as *gentil* 'nice' and *chapeau* 'hat', which are common in spoken language but happen not to occur in the written corpora on which Vocabprofil is based. Our results confirm those of Horst and Collins (2006) whose learners did not use a higher number of low frequency words after 400 hours of tuition either but a larger variety

of high frequency words which belong to the k1 layer. As many researchers have found that the percentage of K2 words (and beyond) is very low in learner language, it may be important to further differentiate between different frequency layers among the k1 group. In this paper we have shown that using a *small* basic vocabulary (n=246 words) in calculations of the AG works better than using a large basic vocabulary (n= 1378 or n=2000). In Tidball and Treffers-Daller (in prep.) we illustrate this further by focusing on motion verbs. While the level 1 learners prefer to use basic deictic verbs (*aller* 'go' and *venir* 'come') over path verbs such as *entrer* 'to enter', all of which belong in the K1 layer of the vocabulary frequency lists, level 3 learners increasingly use *entrer* to express the same motion event, which shows they have progressed in comparison with the level 1 learners. The percentage of low frequency words is therefore for these learners possibly a less suitable indicator of the differences in the lexical richness of their speech.

The analysis of our data with Vocabprofil provided interesting information about the frequency layers of the vocabulary used by our learners, but the large number of words in the NOL category is worrying in that this category contains both very rare words such as *bousculer* 'to knock over' and highly frequent items which are characteristic of spoken rather than written language. It would therefore be very useful

for the research community if a new version of Vocabprofil could be created which is based on frequency data for oral language.

Finally, a preliminary analysis of the use of *criminel* by the learners indicates that cognates play an important role in L2 vocabulary acquisition, which confirms the results of Laufer and Paribakht (1998) and Horst and Collins (2006). The strategic use learners make of cognates is an area that deserves further attention in future studies of vocabulary richness.

References

Cobb, T. and Horst, M. (2001). Is there an academic word list in French? In: P. Bogaards and B. Laufer (eds.), *Vocabulary in a second language*. Benjamins, pp. 15-38.

Daller, H. and Huijuan Xue (2007). Lexical richness and the oral proficiency of Chinese EFL students. In: H. Daller, J. Milton and J. Treffers-Daller (eds), *Modelling and assessing vocabulary knowledge*. Cambridge: CUP.

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

Goodfellow, R., Jones, G. and Lamy, M.-N. (2002). Assessing learners' writing using Lexical Frequency Profile. *ReCALL* 14: 129-142.

Gougenheim, G. (1959). *Le Français fondamental 1er degré*. Ministère de l'Education Nationale. Paris : Institut National de Recherche et de Documentation Pédagogiques.

Gougenheim, G., Rivenc, P., Michéa, R. et Sauvageot, A. (1964). *L'élaboration du français fondamental.* Paris : Didier

Guiraud, P. (1954). *Les caractéristiques du vocabulaire.* Paris: Presses Universitaires de France

Hayes, D.P. and Ahrens, M.G. (1988). Vocabulary simplification for children: a special case of 'motherese'. *Journal of Child Language* 15: 395-410.

Horst, M. and Collins, L. (2006). From *Faible* to Strong: How Does Their Vocabulary Grow? *The Canadian Modern Language Review*, 63(1): 83-106.

INALF (1971). *Dictionnaire des fréquences du Trésor de la Langue Française*. Paris: Didier

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

Laufer, B. (1995). Beyond 2000. A measure of productive lexicon in a second language. In: L. Eubank, L. Selinker and M. Sharwood Smith (eds). *The Current State of Interlanguage. Studies in Honor of William E. Rutherford*. Amsterdam/ Philadelphia: John Benjamins, 265 - 272.

Laufer, B. (1998). The development of active and passive vocabulary in a second language: same or different? *Applied Linguistics*, 16: 307-22.

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

Laufer, B. and Paribakht, T.S. (1998). The relationship between active and passive vocabularies: effects of language learning context. *Language Learning* 48 (3), 365-391.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum.

Malvern, D.D., Richards, B.J., Chipere, N. and Durán, P. (2004). *Lexical Diversity and Language Development: Quantification and Assessment.* Houndmills etc.: Palgrave Macmillan.

McCarthy and Jarvis (2007). VOCD a theoretical and empirical investigation. *Language Testing*, Vol. 24, No. 4, 459-488

Meara, P. and Bell, H. (2001). P_Lex: a simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect* 16: 5-19.

Ovtcharov, V., Cobb, T. and 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*. Volume 63 (1) : 107-125.

Picoche J. (1993). *Didactique du vocabulaire.* Paris:Nathan

Plauen, E.O. [1952] 1996. *Vater und Sohn*, Band 2. Ravensburger Taschenbuch.

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Tidball, F. and Treffers-Daller, J. (2007). Exploring measures of vocabulary richness in semi-spontaneous French speech: A quest for the Holy Grail? In: H. Daller, J. Milton and J. Treffers-Daller (eds) (2007). *Modelling and assessing vocabulary knowledge*. CUP, 133-149.

Tidball, F. and Treffers-Daller J. (in prep.). Variability in the expression of motion events by British learners of French: the role of transfer and simplification.

Van Hell, J. G. and de Groot, A.M.B. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition* 1 (3), 193-211.

Verlinde, S. and Selva, T. (2001). Corpus-based versus intuition-based lexicography: Defining a word list for a French learner's dictionary. In: P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds). *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster: Lancaster University, University Centre for Computer Corpus Research on Language, pp. 594- 598.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing* 17, 65-83.

Véronis, J. (2000). *Fréquence des mots en français parlé*. http://www.up.univ-mrs.fr/~veronis [accessed 1 October 2007]

Websites:

The Range programme (Paul Nation):

http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Vocabprofil:

http://www.lextutor.ca/vp/fr/

Beeching's corpus of spoken French:

http://www.uwe.ac.uk/hlss/llas/iclru/corpus.pdf

DELIC team (Description Linguistique Informatisée sur Corpus):

http://sites.univ-provence.fr/delic/

Véronis' frequency lists:

http://sites.univ-provence.fr/veronis/donnees/index.html

word count: 6273

Table 1. Percentage of words in each Vocabprofil frequency band in each corpus.

Vocabprofil Bands	FF1 (1378 words)	Corpaix first 1378 words	Corpaix first 2000 words	Judges (246 words)	Corpaix first 246 words
K1 Words (1 to 1000)	60.12	59.29	49.10	89.0	93.57
K2 Words (1001 to 2000)	14.15	21.19	23.0	6.46	3.61
K3 Words (2001 to 3000)	3.84	5.08	6.6	0.76	0.40
NOL (not in list of first 3000 words)	21.90	14.44	21.3	3.80	2.41

Table 2. Mean scores on different measures of vocabulary richness and results of a one-way ANOVA / Tukey post hoc.

	Level 1 (N=21)	Level 3 (N=20	Native speakers (N=23)	F-value, df (2,61)	Tukey post hoc			
					1-3	1-NS	3-NS	Eta^2
Guiraud	4.29	5.28	6.27	57.0	**	**	**	.651
D (lemmat.)	18.78	26.53	34.87	58.9	**	**	**	.659
AG (FF1)	0.30	0.54	1.21	32.4	-	**	**	.515
AG (judges)	0.08	0.10	1.50	58.7	*	**	**	.658
AG (Corpaix, 246)	1.68	2.56	3.35	47.8	*	**	*	.610
AG (Corpaix, 1378)	0.48	0.69	1.18	36.8	-	**	*	.547
AG (Corpaix, 2000)	0.30	0.45	0.87	31.2	-	**	*	.505

* = p <.05; ** p <.001

Table 3. Percentage of the tokens belonging to different frequency layers (Vocabprofil) for all three groups (One-way ANOVA/Tukey post hoc).

	1-1000 (K1)	1001-2000 (K2)	2001-3000 (K3)	NOL (not on lists)
Level 1 (N=19)	92.77	4.47	.69	2.07
Level 3 (N = 20)	90.87	4.61	.15	4.37
Native speakers (N= 25)	88.83	4.99	.916	5.63
F (2, 61)	19.03	.741	1.80	31.58
P (Tukey)	<.001 (all groups sign. different)	n.s.	n.s.	<.001 (all groups sign. different)
Eta2	.384	.024	.056	.509

Table 4. Diferent keywords from the stories in Vocabprofil, FF1, Corpaix and the teachers' judgements

Keywords	Vocabprofil	FF1	Rank order in Corpaix	Teacher judgements (out of 14)
banque	K1	no (FF2)	926	4
bâton	NOL	yes	2362	5
canne	NOL	no	no	7
chien	K2	yes	649	3
coup	K1	yes	224	6
criminel	K2	no	no	10
lac	K1	yes	2434	6
vol	K1	no	1438	8
voler	K2	yes	1480	8
voleur	NOL	yes	no	7

Table 5. Examples of keywords from the stories and their allocation to different frequency layers in Vocabprofil

K1	K2	K3	NOL
banque	chien	amuser	bâton
vol	voler	(se) promener	voleur
aller	content		chapeau
retrouver	enlever		nager
lancer	rapporter		bouche
autant	emmener		braqueur
pouvoir	ramener		bousculer
se rendre compte			désarroi