Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability

CHAPTER 3

Measuring lexical diversity among L2 learners of French

An exploration of the validity of D, MTLD and HD-D as measures of language ability*

Jeanine Treffers-Daller
University of Reading

In this study two new measures of lexical diversity are tested for the first time on French. The usefulness of these measures, MTLD (McCarthy and Jarvis (2010 and this volume) and HD-D (McCarthy and Jarvis 2007), in predicting different aspects of language proficiency is assessed and compared with D (Malvern and Richards 1997; Malvern, Richards, Chipere and Durán 2004) and Maas (1972) in analyses of stories told by two groups of learners (n = 41) of two different proficiency levels and one group of native speakers of French (n = 23). The importance of careful lemmatization in studies of lexical diversity which involve highly inflected languages is also demonstrated. The paper shows that the measures of lexical diversity under study are valid proxies for language ability in that they explain up to 62 percent of the variance in French C-test scores, and up to 33 percent of the variance in a measure of complexity. The paper also provides evidence that dependence on segment size continues to be a problem for the measures of lexical diversity discussed in this paper. The paper concludes that limiting the range of text lengths or even keeping text length constant is the safest option in analysing lexical diversity.

Keywords: Lexical diversity, SLA, vocabulary, validity, French

* I am very grateful to Phil McCarthy for his advice in working with the Gramulator and to Scott Jarvis and Tom Salsbury for their detailed comments on earlier versions of this chapter. All remaining errors are mine.

Introduction

Vocabulary is an essential component of language. As Bates and Goodman (1997) have shown, it is knowledge of words which drives the acquisition of grammar. Lexical knowledge is also one of the main prerequisites for academic achievement of monolingual and bilingual children (see Daller, 1999; Dickinson & Tabors, 2001) and it has been shown to be an important factor in, for example, reading ability in L1 and L2 (Anderson & Freebody, 1981; Hu & Nation, 2000). As can be seen in Read's overview (2007), a plethora of tools and measurements have been developed to measure lexical knowledge. While for a small minority of widely spoken languages there are standardized vocabulary tests, such as the Peabody Picture Vocabulary Task (Dunn & Dunn, 1959/2006), tests do not exist for the vast majority of languages. Given the fact that market factors drive the development of tests, it is unlikely that standardized tests will be developed in the near future for languages which lie outside the top twenty. For this reason, researchers interested in less widely spoken languages will greatly benefit from using vocabulary measures which can be applied to any language, are freely available, and can be calculated on samples of naturalistic speech which do not require the researchers to buy expensive tests or equipment. In earlier studies (Treffers-Daller, 2009; 2011) I have shown how indices of lexical diversity (LD) can provide reliable and valid information to researchers who want to measure (lexical aspects of) language ability in bilinguals or L2 learners but do not have access to standardized tests or consider that using such tests is inappropriate for some reason.

The most widely known index of LD is the type-token ratio or TTR, often attributed to Templin (1957), but probably first introduced by Johnson (1939; 1944)[1]. In his 1944 publication, he acknowledged the problem that it is dependent on text length and offered some alternatives (such as Mean Segmental TTR) to overcome the problem. Others have proposed different mathematical transformations of TTR, e.g. the Index of Guiraud (Guiraud, 1954), also known as Root TTR, and log corrections such as the Index proposed by Maas (1972). All of these try to capture the lexical diversity of texts in the form of a ratio of types (V) over tokens (N) but compensate to some extent for the text length issue. The way the Maas Index (from now on: MAAS) is calculated is given in (1). A full discussion of different ratios is provided by Baayen and Tweedie (1998).

1. According to Johnson (1944) several authors, including Carroll (1938) and Jersild and Ritzman (1938) were investigating the proportion of types and tokens in text independently of each other and came up with similar ideas at the time, but not all were aware of the text length dependency of TTR. I am very grateful to Scott Jarvis for pointing me in the direction of these early sources.

$$(1) \quad MAAS = \frac{\log N - \log V(N)}{Log2(N)}$$

As this article focuses on French it is worth pointing out that there is an important French research tradition in studying lexical richness (see for example Cossette, 1994 and Dugast, 1980). One of the first authors to investigate lexical richness in French learner language is Dewaele (1993) who also presents an overview of the earlier studies done on French. In recent years great advances have been made in measuring lexical diversity in samples of oral or written speech, and more sophisticated indices have been developed, which will be discussed below, even though some researchers (e.g. Le Normand, Parisse, & Cohen, 2008) continue to measure lexical diversity by counting the number of different words or the total number of words in a text.

Many researchers use the D-measure (also known as VOCD, the command used in CLAN to calculate the D-value), first proposed by Malvern and Richards (1997) and further developed and tested in detail in Malvern, Richards, Chipere and Durán (2004). D has not only been empirically tested in first and second language acquisition, but also among bilinguals and in clinical contexts (Owen & Leonard, 2002; Silverman & Ratner, 2002). It has also been applied to a number of different languages, including Arabic (Ntelithelos, Idrissi, Tibi, Al Qahtani, Tamimi, Al Eisaei ms.); Cantonese (Klee, Stokes, Wong, Fletcher, & Gavin, 2004), Dutch (Treffers-Daller, 2011), English (Malvern & Richards, 1997; Jarvis, 2002; Yu, 2010), French (Malvern & Richards, 2002; David, 2008; Lindqvist, 2010; Macaro & Masterman, 2006; Tidball & Treffers-Daller, 2008), Italian (Spinelli, 2011), Brazilian Portuguese (Lonngren-Sampaio, in prep.) and Spanish (Malvern et al, 2004; Perez-Bazan, 2004).

More recently, two alternatives were proposed: the first of these is a Measure of Textual Lexical Diversity (MTLD), which was first proposed by McCarthy (2005) and later tested by Crossley, Salsbury and McNamara (2009) and McCarthy and Jarvis (2010). This measure is calculated as the mean length of sequential word strings in a text that maintain a given TTR value (which the authors have chosen to be 0.720 (see McCarthy & Jarvis, 2010, for details). MTLD calculates the TTRs in a sentence until the TTR drops to 0.72, at which point the first factor is complete and TTRs are counted from scratch again: as in the following example: *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) |||FACTORS = FACTORS = 1||| *for* (1.00) *the* (1.00) *people* (1.00) . . . and so on (McCarthy & Jarvis, 2010: 384)[2]. Subsequently MTLD is obtained by dividing the total number of words by the total number of factors. Thus, if the text is 360 words long and there are

2. The example is kept brief for reasons of space. Factors do not normally consist of so few words.

4 factors, the MTLD value is 90. The programme does not discard remaining data but calculates a partial factor for remainders of the data (see McCarthy & Jarvis, 2010 for details). The final version of MTLD is obtained by running the programme forward and backward through the data and calculating an average of the outcome of both. According to McCarthy (2005) and Crossley et al. (2009), MTLD does not vary as a function of text length for text segments whose length is in the 100–2,000-word range.

The second of the new measures is HD-D (McCarthy & Jarvis, 2007), which is similar to D but based on the hypergeometric distribution function (Wu, 1993). HD-D calculates, for each lexical type in a text, the probability of encountering any of its tokens in a random sample of 42 words drawn from the text (Jarvis & McCarthy, 2010: 383).

As MTLD and HD-D have only recently been developed, McCarthy and Jarvis (2010: 381) call for further validation of these measures, and a systematic comparison with other, more established measures, such as D and older measures of LD such as Maas (1972). So far MTLD and HD-D have only been tested in English, apart from an unpublished study on Spanish by McCarthy (personal communication) and an unpublished study on L2 Finnish data by Jarvis (personal communication), and it seems therefore relevant to find out whether these measures work for languages other than English. To the best of my knowledge, the current study is the first one in which these new indices are applied to French.

One of the key issues in validating a new measure is providing evidence for its concurrent validity; that is, the extent to which it correlates with "a criterion which we believe is also an indicator of the ability being tested" (Bachman, 1990: 248). Unfortunately, an independent standard for diversity does not exist apart from the measures of LD under study. While the measures could possibly be validated by correlating the LD scores with judgements of the diversity of the texts under study (as is done in Crossley, Salsbury, & Macnamara, this volume), a disadvantage of this approach is that the construct is not easy to understand. It would be difficult to find out whether judges have indeed assessed the diversity of the texts or whether they have instead assessed a different property (e.g. overall quality of the texts). As lexical diversity measures have often been used as a general purpose measure of spoken and written language development (Malvern et al., 2004, p. 8), I have chosen to use a C-test as the anchor test to assess to what extent the LD measures can predict general language proficiency. Strictly speaking, this is not a validation of the measures as indices of lexical diversity, but an assessment of their predictive validity: their usefulness in predicting a different – albeit related – construct. In this context we will also discuss the incremental validity of the measures. Incremental validity is an evaluation of the extent to which new

measures contribute to explaining variance above and beyond other measures (see McCarthy & Jarvis, 2010).

In a study of over 800 students of German as a Foreign Language, Eckes and Grotjahn (2006) use Rasch analysis to provide evidence for the unidimensionality of their German C-test and for the claim that it measured the same underlying ability as their criterion measure, the TestDAF (Test for German as a Foreign Language), namely general language proficiency. In a previous study (Tidball & Treffers-Daller, 2008), I have shown that the French C-test we developed is a valid tool to measure general language proficiency in French. If the new measures are found to correlate strongly with this external criterion, this will constitute important evidence for their usefulness as a proxy for general language ability.

The internal validity of MTLD, HD-D and D will be studied through an analysis of their dependency on text length. In addition, following McCarthy and Jarvis (2010) I will look at the convergent validity of the measures, that is, the principle that measures of theoretically similar constructs should be highly intercorrelated (Trochim, 2006). In this particular case, we interpret this to mean that new measures of LD should correlate strongly with existing widely accepted measures of lexical richness, in particular the D-measure (Malvern & Richards, 1997; Malvern et al., 2004) and also – but perhaps to a lesser extent – with derivatives of TTR, such as MAAS (1972).

Complementary to convergent validity, the notion of discriminant or divergent validity refers to the principle that measures of theoretically different constructs should not correlate highly with each other (Campbell & Fiske, 1959). In the current study, an index of complexity is used as the criterion against which divergent validity of the lexical diversity measures are assessed. The mean number of words per T-unit (Hunt, 1965) was used as an index of the complexity of students' language (see under methods for a definition and examples of T-units in French), because it is the single most employed measure of complexity (Norris & Ortega, 2009). Although according to Wolfe-Quintero, Inagaki and Kim (1998) the mean length of T-units measures fluency rather than complexity, Ortega (2009) provides evidence that the mean length of T-units is best seen as a measure of complexity. Whilst complexity and fluency may well be related to lexical diversity, the constructs are certainly not identical, and can therefore not be assumed to covary in the same way as two different measures of LD. The most important point for the current study is therefore that the mean length of T-units should provide a good criterion for assessing divergent validity, whether one believes it is a proxy for complexity or fluency. For a full discussion of different measures of complexity and fluency, the reader is referred to Ortega (2009).

The study also aims to illustrate the importance of lemmatizing the data prior to calculating any measure of lexical richness. For obvious reasons, the operational

definition of types and tokens needs to be valid if any of the measurements that are based on this crucial distinction are to have construct validity. Thus, researchers need to decide whether all inflected forms of a word count as one type or as different types. In highly inflected languages such as French, the best solution is to use the base word as the unit of measurement rather than the inflected form, to ensure that different inflected forms of verbs, such as *arriver, arrive, arrives, arrivons* "work" etc., nouns such as *bureau, bureaux* "office(s)" and adjectives such as *petit, petite, petits, petites* "small" and articles such as *le, la, l' les* "the" are counted as different tokens of one type rather than as different types. If data from highly inflected languages are not lemmatized, values of lexical richness can be strongly inflated, as can be seen for example in Ntelitheos et al (ms.) who find D-values of over 200 among Emirati Arabic-speaking children between the ages of three and five, whilst for English-speaking children between the ages of seven and fourteen Malvern et al (2004: 169) found average D- values ranging from 40 to 73, with no one scoring higher than 106. Clearly, comparing D-values across languages is problematic and differences in D-values can be ascribed to typological differences between languages. However, such large discrepancies can be avoided if data are lemmatized appropriately (see also Treffers-Daller, 2011 and Treffers-Daller & Korybski, in prep. for a discussion of cross-linguistic comparisons of lexical richness measures).

2. Method

The data used for the analysis are taken from the Learner Language Project (see Tidball & Treffers-Daller, 2007; 2008). The participants consisted of a group of undergraduates who studied French at a British university. Only students who were native speakers of English took part in the study. There were 21 first year students (level 1) and 20 final year students (level 3), as well as a group of 23 native speakers of French who were studying English as a foreign language at the same university. Two comic strips from Plauen's ([1952] 1996) father-and-son stories were used to elicit narratives from each student individually in the interpreting laboratory. The stories chosen were *Unbeabsichtigte Helden* "involuntary heroes", where the father and the son witness a bank robbery and *Erfolglose Anbiederung* "unsuccessful ingratiation", where the two protagonists play fetch the stick with a dog. These two were chosen because they were found to be most appealing to students in a pilot study. Students were asked to tell the stories in their own words. They were encouraged to prepare their stories prior to recording and to start

recording when they were ready. They told the stories in any order they wanted[3]. In addition, each student filled in a French C-test, which was used to measure students' general language ability in French.

All data were transcribed in CHAT format (MacWhinney, 2000) and carefully lemmatized on the main tier with the help of the change string command. For this purpose a changes.cut file was created which listed all the changes that needed to be made in the data. The lemmatization involved replacing all inflected forms of nouns, verbs and adjectives with the corresponding lemma, so that forms such as *cherche, cherchons, chercha* etc. were changed to *chercher* "to look for". Articles, demonstratives, pronouns and question words were all replaced with the masculine singular form. Thus, for example *ils* "they" was replaced with *il* "he" and *tous* "all-pl" with *tout* "all", and *la* "the-fem" with *le* "the-masc", and *quel(le) (s)* "which" with *quel*. Form variants of conjunctions, such as *que* and *qu'* "that," were all standardized to one form, in this case *que*. The lemmatization process is illustrated in (2).

(2) Example of the lemmatization applied to the data from a level 3 learner of French

```
*542:  et il ne devait [: devoir] pas entrer dans la [: le]
       banque .
*542:  et ensuite ils [: il] font [: faire] la [: le] fête
       avec tous [: tout] les [: le] banquiers [: ban-
       quier] et le laron est [: être] arrêté [: arrêter]
       par la [: le] police .
*542:  je crois [: croire] qu' [: que] ils [: il] font
       [: faire] des [: de] photos [: photo] il y aura
       [: avoir] peut+être un [*] dans le journal .
```

Hyphenated words such as *peut-être* "maybe" and compound prepositions such as *parce que* "because" and nominal compounds such as *coup de poing* "punch" were linked with "+" symbols as is common practice in CHAT transcription. In the case of *parce que* this was necessary because *parce* is not a word in French and failure to link it to *que* would result in *parce* and *que* being counted as separate types (one of which does not exist). For compounds, this was done because knowledge of compounds constitutes advanced knowledge well beyond the knowledge represented by the use of the individual parts. Transcribing the above word sequences as

3. I checked whether the LD scores were affected by the order in which the stories were told by comparing two ways of calculating the LD scores on a sample of nine texts: I first calculated the scores on the text in which the stories were told in the original order and then I reversed the order of the stories and recalculated the scores. Three texts from each level were randomly sampled for this purpose. The mean scores for HD-D were identical, those for D differed by 0.02 and those for MTLD differed by 0.05. As the differences were so small it is unlikely that the order in which the stories were told had a significant effect on the results.

peut+être, *parce+que* and *coup+de+poing* ensured they were counted as separate types, different from the use of, for example, *être* and *que*, which can be used on their own as a verb or a conjunction respectively in other constructions. Finally, when students switched to English (e.g. the use of *stick* in *Erfolglose Anbiederung* "unsuccessful ingratiation") these switches were excluded from the analysis.

The students produced on average 96 types (lemmatized) and 325 tokens per story, as can be seen in Table 1. The number of tokens and types ranged from 127 tokens and 39 types (minimum) to 1350 and tokens and 290 types (maximum). As one might expect, level 1 students produced the lowest number of types and tokens on average and the native speakers the highest number. An ANOVA revealed that these differences were significant for the types (F (2, 61) = 17.95, $p < .001$) and the tokens (F (2,61) = 6.03, $p < .01$). A Tukey post hoc analysis showed that the level 1 learners were significantly different from the level 3 learners and from the native speakers with respect to their use of types and tokens ($p < .05$), but the level 3 learners and the native speakers were not significantly different from each other. The high standard deviation for the native speakers is due to the fact that one speaker in this group produced an exceptionally high number of words (1350) in telling the stories. If this speaker is excluded, the mean for the native speakers is reduced to 343.8, which is less than the mean for the level 3 students, and the number of types to 114.8, which is higher than the corresponding result for the level 3 students. This simple comparison of the frequency of types and tokens seems to suggest that the level 3 students are more verbose in telling the stories than the native speakers, but the stories of the former tend to be less diverse than those of the latter. Similar verbosity of L2 learners has often been found, and is generally referred to as the "waffle phenomenon", a term coined by Edmondson and House (1991). As differences between the level one group and the native speakers remain significant if this student is excluded, and the differences between level three students and native speakers continue to be non-significant, it was decided not to discard this student. Instead, I decided to investigate whether the performance of LD measures was negatively affected by the wide range of text lengths in the database. Calculations are therefore performed twice: first on the entire data set, and

Table 1. Types and tokens produced by all groups

	Types (lemmatized) M (SD)	**Tokens M (SD)**
Level 1	63.86 (19.82)	224.76 (107.12)
Level 3	100.1 (22.13)	359.2 (91.17)
native speakers	122.43 (46.50)	387.52 (238.66)
Total	96.23 (40.40)	325.27 (176.60)

then on a more limited range of text lengths (from 200 to 666 words), which McCarthy and Jarvis (2010) recommend as a safe range for using MAAS (1972).

Unfortunately, among the different measures of LD under investigation here only D is available under CLAN. Therefore MTLD, HD-D and MAAS were calculated with a different tool, namely McCarthy's Gramulator, which at the time this study was carried out, was freely available from McCarthy's webpages: <https://umdrive.memphis.edu/pmmccrth/public/software_index.htm>. To facilitate analyses with the Gramulator, all CHAT coding had to be taken out of the transcripts and all transcripts needed to be converted to text format. This was done with the help of the flo command under CLAN. Each file was converted to text in two different versions: a non-lemmatized version and a lemmatized version, which made a comparison between calculations based on either version possible. It is important to know that the values obtained for HD-D are all negative if HD-D is calculated with McCarthy's Gramulator, with those closest to zero indicating high diversity, and values far below zero representing low diversity in a text[4].

While in many studies correlations between text length and measures of LD are reported as providing evidence for text length dependency of such measures, this approach is flawed because more proficient speakers or writers can produce longer texts than less capable ones, and the former can also be expected to possess more diverse vocabularies than the latter (see also McCarthy and Jarvis, this volume). Therefore, one would expect *all* measures of LD to correlate positively with text length. To measure text length dependency, we need calculations of LD from *samples of different lengths from the same text.* This way, we can find out whether an index of LD under study decreases if the same speaker uses more words.

In order to establish whether or not they were text-length-dependent, calculations of D, MTLD and HD-D were made on data samples of 30 students who produced at least 300 words. The cut-off point of 300 words was chosen to make sure almost half of the students who took part in the study could be included (N = 30).

Prior to the calculation of the measures of LD, the texts were treated as follows. To begin with, using McCarthy's Gramulator, a segment of 300 words was taken from the middle of each lemmatized transcript. These 300 word segments were subsequently divided into multiple, equally-sized segments with the help of the Gramulator: each text was divided into three segments of 100 words, and also into two segments of 150 words. For these five segments, I then calculated the LD values. After that, the mean LD value for each length was calculated: for the 100-word length, the text's LD value was the mean of the LD values for all three individual 100-word segments. Then, for the 150-word length, the text's LD value was the

4. The final HD-D calculation is normed to a set of narrative and expository texts for ease of interpretation (McCarthy, p.c.).

mean of the LD values for both 150-word segments. In addition, I calculated the LD value for the one 300-word segment of each text. This approach made it possible to compare LD values by length whilst comparing exactly the same content across all length conditions. If the measures of LD are truly independent of text length, the measures across the three different lengths should not decrease or increase with text length.

For the analysis of complexity, I counted the number of T-units (Hunt, 1965) in the data. According to Bardovi-Harlig (1992: 390) a sentence has two (or more) T-units when independent clauses (with subjects and finite verbs) are conjoined as in (3), but a single T-unit when one or more clauses are embedded in an independent clause as in (4). The end of the first and the second T-unit is indicated in (3) with square brackets. Example (4) consists of one T-unit only, because all clauses in the utterance depend on the main clause.

(3) Il y a un homme une fille et un chien au+bord d' un lac [1] et la fille a jeté une canne dans le lac 1 [2] (level 3 student 540)
"There is a man, a girl and a dog on the shore of lake [1] and the girl throws a stick into the lake [2]."

(4) l' enfant et le mec avec qui elle parlait avant sont dans la banque où il y a l' homme qui a frappé le jeune avec deux pistolets (level 3 student 540)
"The child and the guy with whom she was talking before are in the bank where there is the man who has hit the young man with two pistols."

The index of complexity was calculated by dividing the number of words (tokens) by the number of T-units for each informant.

3. Results

This section first addresses the effect of lemmatization on the scores obtained on the four measures (3.1), and then different aspects of the validity of each measure will be discussed in turn. First of all, we look at their predictive validity; that is, to what extent each measure correlates with the chosen anchor point, the French C-test (3.2). The issue of the measures' internal validity – whether or not they are dependent on text length – is taken up (3.3), after which their convergent validity, divergent validity and incremental validity are assessed (3.4).

3.1 The effect of lemmatization

As explained in Section 2, French is highly inflected. In calculating LD scores, this can be taken into account by lemmatizing the data: this process reduces the

number of types in the data, because inflected forms are no longer counted as different types. Therefore, the results of calculations based on lemmatized data are normally lower than the results based on non-lemmatized data. Put differently, scores based on non-lemmatized data are likely to be strongly inflated. The exception to this rule is the log transformation of TTR proposed by Maas (1972), because in this ratio the value of the numerator *increases* if the number of types is reduced, whilst the value of the denominator remains the same.

To demonstrate the effect of lemmatization on scores, I have calculated each measure on the original, non-lemmatized version of the transcripts and then on the lemmatized version, and computed the differences between each with a paired samples t-test.

As can be seen in Table 2, the results for all four measures are significantly different when the data are lemmatized. For D, the differences between both versions amount to a reduction of 35% in the values obtained, whilst for MTLD the values are reduced by 23%, and for HD-D by 68%. The values for MAAS increase by 15%, but this increase needs to be interpreted as a decrease in diversity on this measure, because high MAAS scores indicate low diversity. It is interesting to compare the MTLD scores to those obtained by Crossley, Salsbury and McNamara (2009) who report MTLD values ranging between 28 and 35 over a one-year study of learners who were enrolled in an intensive EFL learning programme. The average scores I calculated on the lemmatized data are roughly comparable to those of Crossley et al., whilst the non-lemmatized data look inflated by comparison with their data. This analysis provides some support for the view that lemmatizing data prior to calculating LR measures is indeed a useful step, because it makes comparing the results of LR measures between languages easier (see also Treffers-Daller, 2011 where the same point is made).

Clear evidence that lemmatizing data prior to the analysis also increases the explanatory power of LR measures was obtained by studying to what extent the lemmatized and the non-lemmatized data are able to discriminate between the

Table 2. Measures calculated on non-lemmatized and lemmatized data (N = 64)

	Non-lemmatized M (SD)	**Lemmatized (M, SD) M (SD)**	**t**
MAAS	141.54 (15.53)	162.87 (16.85)	25.90**
D	41.95 (13.29)	26.98 (8.3)	19.82**
MTLD	40.27 (9.68)	30.64 (6.91)	14.19**
HD-D	–3.62 (2.19)	–6.07 (2.43)	20.06**

**differences significant at $p < .001$.

Table 3. Effect sizes (η^2) of measures calculated for the three groups on non-lemmatized and lemmatized data (n = 64)

	Non-lemmatized data	Lemmatized data
HD-D	.585	.682
D	.586	.659
MAAS	.362	.429
MTLD	.352	.354

three groups of speakers involved in the current study. Table 3 reveals that for all measures the calculations based on lemmatized data are more powerful in predicting group membership of the informants.

3.2 Predictive validity

In the current study, a C-test was used to measure the students' general language ability. This C-test proved to be highly reliable (Cronbach's alpha = .96) and, as can be seen in Table 4, discriminated extremely well between the three groups. The results from an ANOVA (F (2,61) = 105.37, p < .001) and the Tukey post hoc tests revealed that all groups were significantly different from each other (p < .001). In addition, the Eta-Squared value was extremely high (η^2 = .776), higher than that of any of the LD measures reported on in the previous section.

It is interesting to note in this context that the C-test results correlate significantly, though not very strongly, with the number of tokens produced by the students in the story-telling task (Pearson r = .396, p < .001). As we have seen in Section 2, students at the lower levels produce shorter stories than those at the higher levels. One would therefore expect a link between text length and the C-test results if both of these are indices of language ability. The C-test results also correlate fairly strongly with the number of types (Pearson r = .582, p < .001), which may indicate that the C-test taps to a certain extent into lexical aspects of language ability, as Little and Singleton (1992) and Daller and Xue (2009) suggest. If the correlations between the measures of LD under study in the current paper and the C-test can be shown to be stronger than the correlations between a very basic measure of

Table 4. French C-test results for all three groups (n = 64)

	C-test (M)	SD
Level 1	51.38	12.46
Level 3	75.6	8.9
Native speakers	91.65	5.18

LD ("number of different words") and the C-test, this will constitute important evidence that the sophisticated measures are indeed better measures of language ability than such simple measures.

In Table 5a the results are reported for all informants in the data, and these show that HD-D and D correlate most strongly with the C-test, although the correlation between HD-D and the C-test was slightly stronger than the correlation between D and this external criterion. MAAS or MTLD, by contrast, correlated less strongly with this external criterion, although correlations beyond .5 still count as strong according to Cohen (1988). The excellent results obtained for HD-D and D show that these two do indeed constitute better measures of language ability than a measure such as "number of different words". I also ran a series of simple regressions with the C-test as the dependent variable and the measures of LD as predictors[5]. Table 5a reveals that HD-D and D explain the largest proportion of the variance in C-test scores, as indicated by the R^2s.

In Table 5b the same results are given for students whose stories were longer than 200 words but shorter than 666 words. This interval was chosen because McCarthy and Jarvis (2010: 384) propose reducing the variation in sample size when calculating MAAS. Among the intervals they recommend, the range between 200 and 666 words corresponds best to the current data set.

The results for MAAS improve slightly when the correlations are calculated on a smaller range of sample sizes. The drawback of this approach, however, is that fewer students are then included in the study (50 instead of 64) and this has a negative impact on the strength of the correlations and the explained variances for

Table 5a. Correlations between measures of LD with the C-test and adjusted R^2 (N = 64)

	MAAS[1]	**D**	**MTLD**	**HD-D**
Pearson r correlations with C-test (adjusted R^2)	–.556** (.298)	.763** (.575)	.571** (.326)	.791** (.620)

[1]The correlation with MAAS is negative because low MAAS values indicate high diversity.

Table 5b. Correlations between measures of LD and the C-test, and adjusted R^2 for sample sizes between 200 and 666 (N = 50).

	MAAS	D	MTLD	HD-D
Pearson r (adjusted R^2)	–.637** (.393)	.712** (.494)	.505** (.239)	.762* (.571)

5. It was not possible to run a multiple regression with different LD measures as the predictors because of multicollinearity: several of the correlation coefficients in the correlation matrix in Table 8a are higher than .8 (Field, 2005: 175).

the other measures. Although the results for MAAS and MTLD are moderate, the results presented in Tables 5a and 5b provide powerful evidence for the ability of at least two of these measures (D and HD-D) to predict scores on a measure of general language ability.

3.3 Internal validity: Dependence on text length

As pointed out in Section 2, the issue of text length dependency of LD measures is sometimes studied by correlating the tokens in a text with the scores on LD measures. In the current study, LD measures were also found to correlate significantly with text length: D and HD-D correlated equally strongly with the number of tokens ($r = .61$, $p < .001$), MTLD slightly less strongly ($r = .47$, $p < .001$), whilst MAAS did not correlate significantly with the number of tokens. To investigate the measures' dependence on text length, a more sophisticated approach is needed, whereby the measures of LD are calculated on the *same text but on different segments of this text*. For each story I therefore calculated the mean LD value across three segments of 100 words and across two segments of 150 words, as well as the LD value for the 300-word segment. If the measures are independent of text length one should get the same result for each of these calculations. However, Table 6 shows that the results are not exactly the same when the measures are calculated on segments of different sizes: the results for MTLD decrease, whilst for HD-D the scores increase with sample size (keeping in mind that the HD-D values are negative). For D the picture is less clear: the results for the 100-word segment are slightly higher than one would expect on the basis of the results for the other segment. If we disregard the 100-word segment for D, the scores appear to increase with segment size from 28.81 (150 words) to 31.39 (300 words).

Table 7 provides an overview of the paired samples t-tests that were used to test whether the differences between the scores were significant. Given the large number of tests applied (9), the Holm-Bonferroni correction (Holm, 1979) needed to be applied to avoid interpreting differences as significant when they could be

Table 6. Mean and standard deviations for LD scores measured on different sample sizes (N = 30)

	D	HD-D	MTLD
100 words (mean of three segments)	30.19 (8.29)	–5.74 (1.76)	35.55 (7.88)
150 words (mean of two segments)	28.81 (8.18)	–5.70 (1.73)	34.60 (8.13)
300 words	31.39 (8.02)	–5.08 (1.65)	33.95 (7.76)

Table 7. Differences between LD scores measured on different sample sizes, paired samples t tests (N = 30)

	D		HD-D		MTLD	
	t	p	t	p	t	p
100–150 words	2.22	.034	0.59	.560	2.24	.033
100–300 words	1.82	.079	6.52	**.000**	3.90	**.001**
150–300 words	11.36	**.000**	10.58	**.000**	2.31	.028

due to chance[6]. Thus, for example, the differences between the mean D values of the 100 and the 150 word segments would be significant without this correction. In Table 7 the results that were still significant after applying this correction are given in bold.

To check whether any trends in the relationship between scores and text length become clearer when larger samples are studied, I have recalculated the measures for those students who produced 420 words or more. For obvious reasons, the number of students who produced 420 words is relatively small (n = 10). The procedure followed was exactly the same as described in the methods section: first I cut out 420 words from the middle of the transcripts. Then each sample of 420 words was divided into three segments of 140 words and two segments of 210 words. The measures of LD were calculated on all five segments, after which the mean of the LD measures for the three 140-word segments and the mean of the measures for the two 210-word segments was calculated.

Table 8 shows that the results for D do indeed become clearer now. The values of this measure increase linearly with sample size. This is also the case for HD-D. The results for MTLD are less clear in this table, because the value for the 210-word sample is higher than that for the 140-word sample, which would not be expected if the measure is dependent on text length. It is interesting to note that the D and the MTLD values in Table 8 are higher than those in Table 6, which is probably due to the fact that the sample of ten informants on which Table 8 is based consists of the top performers in the group, namely those who produced the longest texts.

6. Holm-Bonferroni's correction mechanism reduces the chance of a type I error (rejecting the null hypothesis whilst the differences are not significant) using the following formula: $\alpha = \alpha / (k - i + 1)$.Nine t-tests were carried out for the comparisons in Table 7. After putting the p-values in Table 7 in rank order, the new α's were calculated as follows: the lowest p-value (.000) needed to be lower than $\alpha = 0.05/9 = 0.0028$, and the second lowest p-value (.000) needed to be lower than $\alpha = 0.05/8 = 0.0022$, etc..

Table 8. Mean and standard deviations for LD scores measured on different segment sizes (n = 10)

	D M (SD)	HD-D M (SD)	MTLD M (SD)
140 words (mean of three segments)	32.20 (6.63)	–4.93 (1.47)	37.12 (6.11)
210 words (mean of two segments)	32.91 (6.77)	–4.77 (1.46)	37.41 (6.69)
420 words	35.29 (7.00)	–4.25 (1.42)	36.36 (6.81)

Table 9. Differences between LD scores measured on different sample sizes, paired samples t tests (n = 10)

	D		HD-D		MTLD	
	t	p	t	p	t	p
140–210 words	3.82	**.004**	3.32	**.009**	0.53	.607
140–420 words	8.34	**.000**	5.8	**.000**	1.14	.284
210–420 words	6.62	**.000**	4.95	**.001**	2.66	.026

Whilst a sample of ten informants is obviously very small, the differences for D and HD-D are significant, even after applying the rule of Holm-Bonferroni (Holm 1979). In Table 9 the results that remain significant after applying this correction are given in bold.

An explanation for the unexpected result for MTLD could possibly be sought in the computation of this measure. The calculation of factors for this measure generally leaves a remainder of data for which a complete factor could not be calculated. For the remainder a factor size is calculated on the basis of how far the TTR has progressed towards the default factor size of .720. McCarthy and Jarvis (2010: 384) point out that shorter texts are more difficult to evaluate with confidence because of the relatively strong impact of the factor, which is calculated for the remainders of the text. The inclusion of a factor for the remainder always reduces the final MTLD score: for a text of 330 words which consists of 8 complete factors, the MTLD value would be 41.25 if no remainders are taken into account. If, however, a factor of .5 is included to account for the remainder, the final MTLD value is reduced to 38.82 (i.e. a reduction by 5.9 percent). For shorter texts, the impact is proportionally higher. For a text of 100 which consists of three factors, the MTLD value would be 33.33 without any factors for remainders. If a factor of .5 is included to account for the remainder, the final value is reduced to 28.57 (i.e. a reduction of 14.3 percent). The effect of the inclusion of a remainder on the final MTLD score is therefore larger if texts are shorter.

As this example illustrates, in the current study the MTLD value for the sample of 140 words in Table 8 may well be too low, because remainders have been included[7]. If this is indeed the case, the apparent rise from the score for the 140 word sample to the score for the 210 word sample may well be an artefact of the relatively stronger impact of the remainder on the 140 word sample. It is entirely possible that the MTLD would start to fall from the 140 word sample onwards already (and not only from the 210 word sample) if remainders could be calculated differently or excluded from the calculations.

In summary, the results from these analyses show that there is evidence that D, MTLD and HD-D vary with sample size. Whilst the differences in values are often subtle, for D and HD-D they are significant in all comparisons of values involving sample sizes of 140, 210 and 420 words. The fact that not all comparisons of the LD values for samples of 100, 150 and 300 words led to significant differences could be due in part to difficulties involved in calculating these measures on very small samples (namely 100 words). The evidence for text length dependency is strongest for HD-D and D, whilst for MTLD the differences between values calculated for different sample sizes are significant in one case only.

3.4 Convergent, divergent and incremental validity

As all measures of LD can be assumed to tap into the same construct, one would expect these to correlate significantly and strongly with each other (convergent validity) but not necessarily with indices that tap into a different trait (divergent validity). The aim of this section is to establish whether this is the case for our data.

Table 10a summarizes the correlations between all measures of LD under study here, as calculated on all stories, irrespective of text length, and Table 10b provides the same correlations but for sample sizes between 200 and 666.

Table 10a. Correlations between measures of LD (n = 64)

	D	HD-D	MTLD	MAAS	TTR
D	–	.93**	.77*	–.61**	.24
HD-D		–	.77**	–.62**	.22
MTLD			–	–.47**	.16
MAAS				–	–.85**
TTR					–

7. Of course this means that the value for the 100 word sample in Table 6 may also be too low. This possibility is entirely consistent with the interpretation given in this section. In fact, if this was the case, the differences between the 100 word and the 150 word sample would possibly become significant.

Table 10a shows that D and HD-D correlate so strongly with each other that they can be considered to be virtually the same. Correlations of D and HD-D with MTLD are slightly less strong, however, and those with MAAS range between medium and strong. The absence of correlations between the newer measures of LD on the one hand and TTR on the other hand provides evidence for McCarthy and Jarvis' (2010) claim that a measure which is so strongly dependent on text length should not correlate well with more sophisticated measures of LD. The exceptionally strong association between MAAS and TTR found in the present study is therefore worrying and raises doubts about the validity of MAAS.

Limiting the effect of text length by selecting only those texts whose length ranges from 200 to 666 words does, however, improve the results for MAAS quite considerably: other LD measures now correlate strongly with MAAS, and there are medium to strong correlations with TTR and the other measures, as Table 10b reveals. The correlations between MAAS and TTR even increase to above .9 under these conditions, which suggests these measures are virtually the same once variation in sample size has been limited. HD-D and D continue to correlate extremely strongly too, but this correlation exists irrespective of sample size, as is evident from Table 10a.

After studying the correlations among the different measures of LD, we look at the correlations between these LD measures and a variable which measures a different component of language ability, namely complexity. As a detailed analysis of complexity is beyond the scope of the current study, only one measure is used here, namely the mean number of words per T-unit (Hunt, 1965). The three groups differ in predictable ways from each other with respect to this index of complexity, and the overall differences are significant ($F(61,2) = 13.66$, $p < .001$). The differences between the level one and the level three students are not large enough to become significant in a post hoc test, but all other intergroup differences are significant. Table 11 gives an overview of the mean scores and standard deviations of the complexity index.

Table 10b. Correlations between LD measures calculated on sample sizes between 200 and 666 (N = 49)

	D	HD-D	MTLD	MAAS	TTR
D	–	.921**	.705**	–.763**	.575**
HD-D		–	.711**	–.771**	.551**
MTLD			–	–.503**	.369**
MAAS				–	–.915**
TTR					–

Table 11. Mean scores and standard deviations of the complexity index

	Mean (SD)
Level one	10.73 (1.78)
Level three	11.44 (1.56)
Native speakers	14.03 (2.94)

Table 12. Correlations of the index of complexity with LD measures, and adjusted R2s

LD measures	**D**	**HD-D**	**MTLD**	**MAAS**	**TTR**
Complexity	.581** (.326)	.517** (.267)	.503** (.241)	–.268* (.057)	.05 (ns) (.057 (ns))

Table 12 shows that correlations between the complexity measure used in this study and the different measures of LD are significant, but less strong than the correlations of the LD measures among each other, which confirms the expectations. Even lower correlations would perhaps have been obtained if complexity had been operationalised differently, e.g. as the number of subordinate clauses per T-unit, but this was beyond the scope of the current project. Table 12 also demonstrates that D, HD-D and MTLD explain a considerable amount of variance (between 24 and 33 percent) in the complexity scores, which is still considerable but much less than the explained variance in the C-test scores (see Table 5a). These R^2s were obtained by running a simple linear regression with the LD measures as predictors and the index of complexity as the dependent variable.

Given the fact that complexity is an aspect of language ability, and the C-test is assumed to measure general language ability, one might expect these two variables to correlate too, which was indeed the case ($r = .410$, $p < .001$), although the correlation was slightly less strong than one might have expected.

Finally, it is desirable for a new measure to be informative above and beyond indices that have already been proposed and are considered to measure the same construct (incremental validity). Earlier in this chapter, we have already seen that HD-D explains a little more of the variance in C-test results than the other measures (see Table 5a), which means this measure is indeed informative over and above the others. The results for MTLD were less convincing in this calculation. Using an ANOVA, I have also investigated to what extent the LD measures can predict whether the informants are level one, level three students or native speakers. The Eta-squared values reported in Table 13 reveal that HD-D is the best predictor of group membership, followed by D, whilst MAAS and MTLD are less successful and TTR is least successful. In other words, HD-D performs slightly better than the other measures, but the effect size for MTLD as a predictor of

Table 13. Group membership as predicted by LD measures (Eta Squared)

	Eta squared (all samples, N = 64)	Eta Squared (samples from 200–666 words only) (N = 49)
HD-D	.682	.570
D	.659	.563
MAAS	.429	.593
MTLD	.354	.244
TTR	.253	.483

group membership is substantially lower. Again the picture changes when only text samples with a length in the range between 200 and 666 words are included: MAAS now becomes the strongest measure, followed by HD-D and D whilst, MTLD performs less well than TTR. These results suggest that text length continues to affect the power of most measures to a certain extent, although TTR is clearly most dramatically affected.

4. Discussion and conclusion

In this paper, the focus was on assessing the usefulness of two new measures of LD in predicting different aspects of language ability. The measures under study were the Measure of Textual Lexical Diversity (MTLD) developed by McCarthy (2005) and HD-D (McCarthy & Jarvis, 2007; 2010) and they were compared with more established measures such as D (Malvern et al., 2004) and a traditional measure of LD, namely MAAS (Maas, 1972). In this process, various aspects of the validity of the measures were also addressed, in particular their dependence on text length (internal validity), to what extent they constitute an improvement of existing measures of LD (incremental validity) and how the correlations of LD with each other compare with correlations of the same measures with a measure of a different construct, namely complexity (divergent validity). The present paper is the first to test these new measures on French, a highly inflected language, which poses a particular challenge for measures of LD, which are based on analyses of types and tokens. The data used in the study were transcriptions of oral narratives based on two picture elicitation tasks, carried out by two groups of L2 learners, of different proficiency levels and one group of native speakers of French.

The focus of the current study was, first of all, on establishing the effect of lemmatization on the measures' ability to predict whether participants were level-one or level-three students or native speakers (as measured by Eta Squared). The results show that the performance of the measures can be improved considerably if

the data are carefully lemmatized. A first key finding of this study is therefore that lemmatization is an essential step that needs to be taken prior to calculating LD scores, particularly in highly inflected languages, to avoid obtaining scores that are strongly inflated.

Next, different aspects of the validity of the measures were investigated. First of all, the predictive validity of the measures was assessed against a measure of general language ability, a French C-test, which in prior studies of the same groups had been shown to be a highly reliable and valid instrument for measuring their proficiency in French. The outcome of this analysis showed that HD-D correlated most strongly with this external criterion, even more strongly than D, whilst the results for MTLD were good, and comparable to those for MAAS, but less impressive than those for D or HD-D.

The issue of the internal validity of the measurements was addressed by calculating LD scores on segments of different text lengths (100, 150 and 300 words) that were drawn from the narratives of 30 students who had produced at least 300 words. A comparison of the mean LD scores calculated for the 100, 150 and 300 word segments revealed that the values for D and HD-D increased from the smallest until the largest sample (but for D only if the results for the 150 word sample were disregarded), whilst for MTLD all values fell linearly with sample size (but see Chapter 2 of this volume for evidence that MTLD does not vary with sample size). After applying the Bonferroni-Holm correction not all of the differences were found to be statistically significant, although for each measure at least one difference was significant. As the results for D were not entirely clear, another analysis of the same data was carried out, based on slightly larger samples that were taken from ten students who had produced at least 420 words. The calculations based on segments of 140, 210 and 420 words confirmed that both D and HD-D values increase significantly with sample size. The results for MTLD were not very clear, which may be due in part to the fact that calculating MTLD values on very small samples is problematic, as McCarthy and Jarvis (2010) point out.

The correlations between the different measures of LD turned out to be strong, in particular for D and HD-D, which confirms the results of McCarthy and Jarvis (2007) and of McCarthy and Jarvis (2010), who also found correlations of over .9 between these two. The correlations of D and HD-D with MTLD were less strong but still substantial. There was therefore strong evidence for the convergent validity of the new measures. The most surprising result was perhaps the strong correlation between MAAS and TTR, which became even stronger ($r = .915$) when sample size variation was limited to samples between 200 and 666 words. The strength of this correlation indicates that these two measures are in fact interchangeable in studies in which text length does not vary too much.

Equally clear evidence for the divergent validity of all measures was obtained through correlations between a complexity measure (the number of tokens per T-unit) and the measures of LD. These correlations are much lower than the correlations of the LD measures among each other, which was exactly what one would hope to find when looking for divergent validity. Finally, a comparison of the ability of each measure to discriminate between the three groups revealed that HD-D and D were the most powerful among the LD measures, whilst MTLD and MAAS were less powerful. When the variation in sample size was reduced, quite unexpectedly MAAS turned out to be the strongest predictor of group membership.

In conclusion, the current study clearly provides unambiguous evidence that HD-D and D are good indicators of language ability in French: they are good predictors of C-test results and scores on these measures correlate strongly with the number of words produced by each speaker, which is also an indication of a speaker's language ability. In fact, it appears to be the case that the LD measures that are most consistently positively correlated with text length are also the best predictors of proficiency, namely D and HD-D.

Whilst D and HD-D are the clear winners in the current study, measures such as MAAS can still be used if text length is controlled for or the range of text lengths is kept within reasonable limits, such as those suggested by McCarthy and Jarvis (2007). This may well be a good solution to limit the impact of text length in the calculation of other measures too. The results of the analyses with MTLD point to the importance of addressing the issue of the impact of the remainders in calculating MTLD scores (see also McCarthy & Jarvis, 2010). This measure may well become more powerful if this aspect of its calculation can be improved.

The fact that measures of LD correlate so well with a test of general language ability is good news for those looking for tools to assess language skills in languages for which no standardized language tests exist, as well as for those seeking to assess bilingual ability, as I have argued elsewhere (Treffers-Daller, 2011). The value of these measures could of course be enhanced substantially if norms were developed for different populations and different languages, as is common for standardized vocabulary tests such as the Peabody Picture Vocabulary Task (Dunn & Dunn, 1959/2006). As the research into LD has been so successful and there is now a substantial body of evidence testifying to the validity of some of the measures, establishing such norms is probably the most important task for future studies in this field.

We also know little about the changes in the lexical diversity of learners' output over time. This issue could not be investigated in the current project because it reports on a cross-sectional study of different groups of learners. As Daller, Turlik and Weir (this volume) show, a longitudinal study of lexical diversity can provide important new insights into learners' development over time, which cross-sectional studies cannot reveal.

References

Anderson, R.C. & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie (Ed.),*Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: OUP.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*, 390–395.

Bates, E., & Goodman, J.C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia, and real-time processing. *Language and Cognitive Processes, 12*, 507–584.

Bradac, J.J. (1977). The role of prior message context in evaluative judgments of high- and low-diversity messages. *Language and Speech,* 20, 295–307.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Carroll, J.B. (1938). Diversity of vocabulary and the harmonic law of word frequency distribution. *Psychological Record, II*, 379–386.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cossette, A. (1994). *La richesse lexicale et sa mesure.* Paris: Champion-Slatkine.

Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59*, 307–334.

Daller, H. (1999). *Migration und Mehrsprachigkeit. Der Sprachstand türkischer Rückkehrer aus Deutschland. Spracherwerb und Sprachverlust (Migration and Multilingualism. The Language Proficiency of Turkish Returnees from Germany)*. Frankfurt: Peter Lang.

David, A. (2008). A developmental perspective on productive lexical knowledge in L2 oral interlanguage. *French Language Studies, 18*, 315–331.

Daller, M.H., & Huijian Xue (2009). Vocabulary knowledge and academic study success: A study of Chinese students in UK Higher Education. In B. Richards, M.H. Daller, D.D. Malvern, P. Meara, J.L. Milton & J.Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition. The interface between theory and applications* (pp. 179–193). Houndmills, Basingstoke: Palgrave Macmillan

Dewaele, J-M. (1993). Extraversion et richesse lexicale dans deux styles d'interlangue française. I.T.L., *Review of Applied Linguistics*, *100*, 87–105.

Dickinson, O.K. & Tabors, P.O. (Eds.). (2001). *Beginning literacy with language: Young children learning at home and school*. Baltimore, MD: Paul H. Brookes.

Dugast, D. (1980). *La statistique lexicale.* Genève: Slatkine.

Dunn, L.M., & Dunn, D.M. (1959/2006). *Peabody picture vocabulary scale*. San Antonio, TX: Pearson Assessments.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*, 290–325.

Edmondson, W., & House, J. (1991). Do learners talk too much? In R. Phillipson, E. Kellerman, R. Phillipson, L. Selinker, M. Sharwood Smith, & M. Swain (Eds.), *Foreign/second language pedagogy research* (pp. 273–86). Clivedon, UK: Multilingual Matters.

Gavin, W.J. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research, 47*, 1396–1410.

Hu, M., & Nation, I.S.P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Hunt, K.W. (1965). Grammatical structures written at three grade levels. NCTE Research Report No 3. Champaign, IL: National Council of Teachers of English.

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

Jersild, A.T. & Ritzman, R. (1938). Aspects of language development: The growth of loquacity and vocabulary. *Child Development*, *9*, 243–259.

Johnson, W. (1939). *Language and speech hygiene* (General Semantics Monographs, No. 1). Lakeville, CT: Institute of General Semantics.

Johnson, W. (1944). Studies in language behavior. A program of research. *Psychological Monographs, 56*(2)1–15. doi: 10.1037/h0093508.

Klee, T., Stokes, S. R., Wong, A. M.-Y., Fletcher, P., & Gavin, W. J. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research, 47*, 1396–1410.

Le Normand, M.-T., Parisse, C. & Cohen, H. (2008). Lexical diversity and productivity in French preschooolers: developmental, gender and sociocultural factors. *Clinical Linguistics & Phonetics, 22*(1), 47–58.

Lindqvist, C. (2010). La richesse lexicale dans la production orale de l'apprenant avancé de français. T*he Canadian Modern Language Review/La revue canadienne des langues vivantes, 66*(3), 393–420.

Little, D., & Singleton, D. (1992). The C-test as an elicitation instrument in second language research. In. R.Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (pp. 173–92). Bochum: Brockmeyer.

Lonngren-Sampaio, C. (In prep.). Portuguese-English code-switching. University of Hertfordshire.

Maas, H.D. (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, *8*, 73–79.

Macaro, E., & Masterman, E. (2006) Does intensive explicit grammar instruction make all the difference? *Language Teaching Research, 10*(3), 297–327.

Malvern, D.D., & Richards, B.J. (1997) Anew measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Clevedon, UK: Multilingual Matters.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical richness and language development: Quantification and assessment*. Houndmills, Basingstoke: Palgrave Macmillan.

McCarthy, P.M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). *Dissertation Abstracts International*, *66*(12). (UMI No. 3199485).

McCarthy, P.M., & Jarvis, S. (2007). A theoretical and empirical evaluation of vocd. *Language Testing, 24*, 459–488.

Norris, J.M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578.

Ntelitheos, D., Idrissi, A. Tibi, S., Al Qahtani, S., Tamimi, O., & Al Eisaei, F. (ms.) The development of morphosyntactic complexity in Emirati Arabic. United Arab Emirates University.

Owen, A.J. & Leonard, L.B. (2002). Lexical diversity in the spontaneous speech of children with specific language impairment: Application of D. *Journal of Speech and Hearing Research, 45*, 927–937.

Perez-Bazan, M.J. (2005) ¿Qué será, será?: A methodological tool for predicting early bilingualism in a family setting. In J. Cohen, K.T. McAlister, K. Rolstad, & J. MacSwan (Eds.), Proceedings of the 4th International Symposium on Bilingualism (pp. 1821–1841). Somerville, MA: Cascadilla Press.

Plauen, E.O. (1952/1996). *Vater und Sohn*, Band 2. Ravensburg: Ravensburger Taschenbuch.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, *7*(2), 105–125.

Richards, B.J. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language, 14*, 201–209.

Silverman, S., & Bernstein Ratner, N. (2002). Measuring lexical diversity in children who stutter: Application of VOCD. *Journal of Fluency Disorders, 27*(4), 289–303.

Spinelli, M. (2011). La valutazione della competenze lessicale nei bambini: Una confronto tra indici (The assessment of lexical competence in children: A comparison of three indexes). Thesis Laurea vecchio ordinamento. Scienze della formazione primaria, University of Modena and Reggio Emilia.

Templin, M. (1957). *Certain language skills in children*. Minneapolis, MN: University of Minnesota Press.

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

Treffers-Daller, J. (2009). Language dominance and lexical diversity: How bilinguals and L2 learners differ in their knowledge and use of French lexical and functional items. In B. Richards, H. Daller, D.D. Malvern, P. Meara, J. Milton & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition. The interface between theory and applications* (pp. 74–90). Houndmills, Basingstoke: Palgrave Macmillan.

Treffers-Daller, J. (2011). Operationalizing and measuring language dominance. *International Journal of Bilingualism, 15*(2), 147–163.

Trochim, W.M.K. (2006). The multitrait multimethod matrix. <http://www.socialresearchmethods.net/kb/mtmmmat.php> (7 November 2011).

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. 1998. *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center

Wu, T. (1993). An accurate computation of the hypergeometric distribution function. *ACM Transactions on Mathematical Software, 19*, 33–43.

Yu, G. (2010) Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236–259.