Corela
Cognition, représentation, langage

HS-21 | 2017
Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage

Algorithmic and subjective measures of lexical diversity in bilingual written corpora: a discussion

Audrey Bonvin et Amelia Lambelet

Édition électronique
URL : https://journals.openedition.org/corela/4843
DOI : 10.4000/corela.4843
ISSN : 1638-573X

Éditeur
Université de Poitiers

Référence électronique
Audrey Bonvin et Amelia Lambelet, « Algorithmic and subjective measures of lexical diversity in bilingual written corpora: a discussion », *Corela* [En ligne], HS-21 | 2017, mis en ligne le 10 mars 2017, consulté le 07 février 2024. URL : http://journals.openedition.org/corela/4843 ; DOI : https://doi.org/10.4000/corela.4843

Algorithmic and subjective measures of lexical diversity in bilingual written corpora: a discussion

Audrey Bonvin et Amelia Lambelet

Introduction[1]

1 Vocabulary composition and size have been widely investigated in research on monolingual and bilingual language acquisition (i.e. during the first years of life) and have been proven to be dependent on factors such as the parents' socio-economic status, the quality and quantity of input, reading habits in the family, and other variables (see De Houwer, Bornstein, & Putnick, 2014; Vermeer, 2001). The influence of these individual factors increases over time and is even more pronounced in bilingual children (see discussion in De Houwer et al., 2014). The majority of previous studies aiming to assess vocabulary development in young children have focused on the production of individual lexical items or comprehension tests using lists: the child is, for instance, asked to choose the correct picture when listening to a word or to produce the word that corresponds to a picture. Parental reports have also been used to assess linguistic competence in younger children.

2 In the study presented in this paper, we are interested in semi-free productive vocabulary development in older children with an immigration background. The critical issue here is not the respective advantages or disadvantages of monolingualism vs. bilingualism, as has been the case for decades in research on first language(s) acquisition ("is bilingual lexical development slower or faster than monolingual lexical development?"). Instead, we hope to better understand a dimension of productive lexical knowledge that influences raters' subjective evaluation of linguistic competence and, more particularly, that impacts

teachers' assessments (Meara, & Bell, 2001): the lexical diversity (in the following LD) in written productions.

3 Lexical knowledge is a key component of linguistic competence and is therefore an often explicitly trained skill. At school, for instance, pupils are habitually encouraged to use synonyms and to avoid repeating vocabulary when writing. This ability furthermore has an impact on school success in general, as it increases the learners' ability to understand oral and written texts, an important part of every school subject (cf. Dickinson, Flushman, & Freiberg, 2009, p. 23; Henrichs & Schoonen, 2009). More important in our view, however, is that lexical knowledge is a skill that figures in linguistic proficiency evaluations either explicitly (vocabulary tests) or implicitly (evaluation of global linguistic competence). This implicit role of lexical knowledge is particularly striking when it comes to immigrant children, whose lexicon is by definition different from that of monolingual children, and whose lexical development in the school language may have suffered from a lack of input compared to monolingual children (Henrichs & Schoonen, 2009, p. 1).

4 The concept of LD[2] has been of interest to linguists since the 1930's – an interest that has led to the definition of various indices to measure it (see section 2.1 Lexical diversity). LD has therefore been applied in various fields of L1 and L2 acquisition research when, for instance, establishing language dominance in bilingual individuals (Treffers-Daller & Korybski, 2015), investigating the effect of bilingual education on vocabulary knowledge (Zydatiß, 2007), or describing L1 language development (Duran, Malvern, Richards, & Chipere, 2004). Despite the results of these studies, several questions remain open on the very definition of LD and the best ways to describe and measure it quantitatively and/or qualitatively. The aim of this paper is to test the applicability of several measures of LD in short written productions and to participate in the discussion begun by Scott Jarvis in his 2013 paper, in which he argues for a LD measure that considers other factors than solely (a lack of) word repetition. This approach is particularly appealing as it questions the way (quantitative) linguists conceptualise the very notion of diversity (see section 2.2).

5 Our entire discussion is based on a corpus of written productions and on the results of global proficiency tests (C-tests) taken by Portuguese immigrant children in Switzerland; the corpus forms part of a larger research project at the Institute of Multilingualism (University of Fribourg, and University of Teachers' Education, Fribourg, Switzerland). In this project, longitudinal data were collected in the children's heritage language (Portuguese) and the language used in the region to which their parents immigrated (French or German). In this paper we focus on the French and German written productions from the last (e.g. third) data collection in a subset of the sample (n=105 out of 518).

Lexical diversity

6 The most trivial definition of LD can be formulated as such: LD is a quantitative measure of the number of "different" words (types) of a text. According to this definition, the key idea is the notion of "non-repetition" ("different words"), without taking into account other features of the words (e.g. their frequency). This definition obviously fails to consider the importance of the text length; it also explains why indices to gain a relative measure of the concept have been developed since the beginning of the 1930s. Such measures are discussed in the next section (2.1). This limited definition of LD also ignores

the features of the words counted: it can be argued that the relative frequency of the words used by a child, their degree of conceptual complexity, their rarity/frequency in the targeted age group's lexicon should also be taken into account when measuring the LD of a text. This point is discussed in 2.2.

Indices to measure lexical diversity

7 A common problem of current LD measures is that counting the number of different words of texts excludes assessment and comparison of texts of different length. This observation led to the development of several relative measures and indices (Duran et al., 2004). Consequently, resistance of an algorithm to the influence of the length of the analysed texts has become a main factor of acceptance.

8 One of the first measures developed to counter the size-effect problem was the type-token ratio (TTR), i.e. the number of types divided by the number of tokens, proposed by Johnson. For instance, in sentence (1), the type-token ratio would be of 1, whereas in sentence (2), the type-token ration would be of 0.7.

9 (1) John is walking with his dad to the toy shop (10 types/10 tokens, TTR =1)

10 (2) John went to school, then John went to the shop (7 types / 10 tokens, TTR =0.7).

11 TTR is very intuitive but unfortunately sensitive to text length, thus rendering comparisons between TTR of samples of different lengths impossible. In fact, the more tokens a text contains, the more repetitions of already existing types, especially grammatical words (e.g. "the", "and") occur and the less new types appear, causing LD to effectively decrease as text length increases. For this reason, the use of TTR is not a satisfactory solution for short texts produced by primary school children: their LD may decrease as they grow older and write longer texts.

12 To counter this problem, many researchers have developed new indices with diverse algebraic transformations of TTR (e.g. Johnson's MSTTR, 1944; Guiraud, 1954; Herdan, 1964; Maas, 1972). Nevertheless, these TTR variations do not eradicate the problem of text length influence. Later, other new measures based on the rank frequency (i.e. how many words occur how many times in a text), such as Yule's Characteristic Constant, or techniques based on the probability of encountering new types in an increasingly long language sample (e.g. Sichel type-token Characteristic) were widely tested on texts containing several thousands of tokens. Text length influence nevertheless remains a significant factor in language acquisition research dealing with much shorter texts (see Duran et al., 2004, p. 222).

13 Recent measures have emerged with the development of computational linguistics, which proposes other modifications of TTR. One of them is the measure D, designed to calculate the speed at which TTR decreases in a language sample. One part of its calculation is to run a sampling series: it evaluates TTR for 100 random samples of 35 tokens, for 100 random samples of 36 tokens and so on, until samples of 50 tokens have been compiled. The result is an approximation of the value with all possible random samples (McCarthy & Jarvis, 2010, p. 383).[3] Duran et al., (2004) tested D with short texts in the first and foreign language, and concluded that it is a good indicator of language development. Nevertheless, D is sensitive to text length (McCarthy & Jarvis, 2007; Owen & Leonard, 2002 quoted in Fergadiotis, Wright, & Green, 2015) especially with short samples (less than 150 tokens) (Koizumi, 2012, p. 67).

14 HD-D proposed by McCarthy & Jarvis (2007, 2010) is an alternative to D. Concretely, HD-D determines the probability for each type in a text to meet any of its occurrences in a random sub-sample of 42 words. The LD index is then calculated by the sum of the probabilities for every existing type in the text (see McCarthy & Jarvis, 2010, p. 383).

15 The Measure of Textual Lexical Diversity (MTLD) is another modern method developed by McCarthy (2005). The MTLD measures TTR after every word of a sample until it reaches a given value (0.72). Then the TTR measurement starts again with the next token, and so on, until the last token of the sample is considered. Then, the length of the text is divided by the total number of TTR of 0.72 counted (Fergadiotis et al., 2015; McCarthy & Jarvis, 2010). Subsequently, a second MTLD measurement is made in the opposite direction, i.e. from the last to the first word. The average of the forward and backward MTLD scores provides the final MTLD index (see Fergadiotis et al., 2015; Koizumi, 2012).

16 The MTLD measure presents several advantages. According to some studies, it is more robust with regard to text length variations than D or HD-D (e.g. Fergadiotis, Wright, & West, 2013; Treffers-Daller, 2013) and it demonstrates no text length bias for text samples containing between 100 and 2,000 tokens (Crossley, Salsbury, & McNamara, 2009; McCarthy, 2005 quoted in Treffers-Daller, 2013, p. 82). Nonetheless, as Koizumi (2012, p. 67) points out, even if MTLD is more resistant to sample size effect than other measures in some configurations of short texts, it is still sensitive with samples range of 50 to 100 tokens, 100 to 200 tokens, and 50 to 200 tokens. In light of these results, Koizumi (2012) concludes that MTLD should be used with texts having at least 100 tokens and, if possible, with a maximum of a 50-token difference between texts.

17 MTLD and HD-D have not frequently been used on French corpora aside from Treffers-Daller's (2013) comparison of Maas, MTLD, D and HD-D measures on transcriptions of oral narratives (picture elicitation tasks) from two groups of L2 learners and one group of native French speakers. The results revealed that D and HD-D correlate with C-tests, leading to the conclusion that these LD measures can represent an appropriate tool for assessing general language ability. Furthermore, HD-D correlates positively with text length, which is considered a positive indicator of a speaker's linguistic competence. According to Treffers-Daller (2013), the correlation of HD-D and D with C-tests is higher than the correlation with MTLD and Maas, and there is a strong positive correlation between the number of tokens each learner produced and both HD-D and D; by contrast the same results analysed using MTLD are less clear (negative correlation, not always significant). Thus, the indices D and HD-D seem better suited to measuring language proficiency in French. The positive correlation with text length is a feature that is interesting to take into account when measuring children's vocabulary, since a longer text should be positively evaluated at the primary school level, where the tendency is to write short texts. The correlation with C-test results (i.e. global language proficiency) is particularly interesting for our data because, if LD measures and predicts general language ability, this would potentially have useful applications in teaching and assessment methods.

18 To summarise, two modern algorithms represent the most promising methods of measuring LD: HD-D and MTLD. In comparison to other more traditional algorithms, these two measures have the advantage of being less negatively affected by variation in text length; moreover HD-D has secured promising results in the analyses of French samples. For these reasons, we have decided to use these two measures in our own study on lexical development in immigrant children. It should, however, be stated that both

methods also have shortcomings that must be taken into account; these weaknesses are particularly evident when very short texts are concerned – and children often produce short texts. Indeed, MTLD's great resistance to text length evidently decreases with texts shorter than 100 tokens, and HD-D is still in process of validation and thus less corroborated than MTLD (Treffers-Daller & Korybski, 2015).

19 Developing adequate measures of LD is the first step for research on L2 lexical development. Once these measures are established, they can be used, for instance, to document the increase of active vocabulary used by children in longitudinal studies, or to test the influence of input on lexical learning.

Another perspective on lexical diversity

20 As became evident in the previous section, scholars have been continually refining LD measurements and improving their algorithms in the interest of incorporating factors such as text length. Yet, as Jarvis argues (2013), "language researchers have neglected the question of what it is that [these indices] are actually measuring". (Jarvis, 2013, p. 94). In his astute discussion of the very notion of LD, Jarvis (2013) compares the way linguists and biologists conceive diversity, and argues for a comprehensive understanding of this concept that goes beyond the simple equation *diversity ≠ repetition*. He points out that texts should be assessed as a whole and that a more qualitative vision of their lexical diversity should be taken into account. He advocates for a consideration of seven properties of diversity in LD measure: (1) *Size* (number of tokens); (2) *Richness* (number of types); (3) *Effective number of types*; (4) *Evenness* (defined as "the degree to which tokens are distributed equally across types"); (5) *Disparity* (i.e. "the proportion of words in a text that are semantically related"); (6) *Importance* ("the relative frequency with which the words in a text occur in the language as a whole", e.g. larger representative corpora); and (7) *Diversion* ("the average interval between tokens of the same type").

21 This vision of LD diversity is particularly appealing to an applied perspective, although achieving an empirical measurement poses certain difficulties. Nevertheless, teachers assessing written (and oral) productions of their students are influenced by such factors, even if they have not previously been explicitly defined. We therefore believe it is extremely interesting and useful to apply LD measures having greater subjectivity to algorithmic measures of LD. In this paper, we use a method developed by Jarvis to gain insight into subjective or "naïve" assessments of the LD diversity of the texts constituting our corpus and to explore them in light of the other algebraic measures.

The study

22 The data used to create the corpus were collected in a project aiming to describe the development of literacy skills in Portuguese children aged eight to ten and living in French and German-speaking Switzerland. More precisely, the project was designed to test the hypothesis that literacy skills can be transferred from one language into another without further training (for more details, see Berthele & Lambelet, in prep.; Lambelet, Desgrippes, Decandio, & Pestana, 2014[4]). To achieve this goal, longitudinal data were collected (three data collection points) in the participants' heritage language (Portuguese) and in the school language (either German or French) in the framework of various school tasks, including reception and production exercises. Parents also filled in a

questionnaire on the family's linguistic habits (relative input in heritage and school language), the parents' linguistic competence in the school language, and their socio-economic status (SES). In this paper, we only analyse part of the data collected: the participants' written productions at Time 3 and their results on a global proficiency test (C-tests).

Aims and research questions

23 The goal of the present study is to gain better insight into measuring LD in short texts from two perspectives. Our first aim is to explore the algebraic methods HD-D and MTLD, and to describe their distribution in our data and to chart their correspondence to global linguistic proficiency. This serves to answer the following research questions:

24 – Are these two methods suited to measure LD in short written productions in French and German ?

25 – Can LD, as measured by HD-D and MTLD, be used to rate linguistic proficiency ?

26 Our second aim is to explore LD from a more subjective perspective. In this case, subjective assessments, by untrained raters, of the same short texts are used to respond to the following questions:

27 – How reliable are subjective ratings of LD of short written productions ?

28 – Do subjective ratings correlate with the algebraic measures ?

29 – Does the combined use of subjective and algebraic measures allow a better understanding of what LD is ?

30 The first set of questions is particularly relevant from an empirical linguistic point of view. As discussed in the first sections, the algebraic measures of LD are still in a process of validation, especially in languages other than English and for texts produced by children and/or L2 learners. The aim of our study is to test the applicability of these measures in the kind of short texts that are typically produced by children when developing their literacy skills. The relevance of the second set of questions is equally more theoretical (the very notion of LD) and more practical from a pedagogical standpoint. The goal is to create a link between teaching practices (implicit and/or explicit evaluation of lexical knowledge/linguistic competence) and linguists' theorisation of the concept.

Corpus

31 To compile the data necessary for this discussion, 105 written productions from the extended corpus were first lemmatised and their LD measured using HD-D and MTLD. Then, the LD in the texts texts was evaluated subjectively by eighteen untrained raters (fourteen women, mean age=31). All but two of the raters hold a Master degree, thirteen of them in a field of linguistics; a small number (n=4) are teachers in public schools and others have some teaching experience outside of their current occupation. French or German is their L1 (or dominant language). None of them, however, had previously worked on lexical diversity. The evaluations were made on orthographically corrected (but not lemmatised) versions of the texts.

32 Two tasks were designed for the written productions (see Berthele & Lambelet, in prep. for more details). In the first task, pupils were asked to write a letter to an aunt in which they develop an argument for their choice of transportation (airplane or car) for the upcoming holidays. In the second task, they were required to write a descriptive text about their last school trip (narrative text). Because these two tasks are similar to common school exercises, it can be ruled out that the learners were unsettled by the novelty of the tasks; it was therefore expected that the results would correspond to their general scholastic performance.

33 C-tests were used to measure the participants' general linguistic proficiency in the heritage language and the school language. Based on Eckes & Grotjahn (2006; Grotjahn, 1992, 2002), the C-tests were constructed for each language (n= 4 per language) in the form of age-appropriate short texts requiring no specific vocabulary or knowledge of content. For each of the selected texts, half of every second word was deleted and the number of missing letters indicated by number of underlined gaps (except in the first and last sentences). Participants were then asked to fill in the gaps using the letters they believed would correctly complete the words. Contrary to other studies, we clarified the instructions using a short, joint demonstration; this was introduced due to our participants' young age (between eight and ten).

34 Selection of the sub-corpus

35 For the purpose of this study, a sub-corpus based on text length was compiled from the entire corpus. To counterbalance the small size of the texts, we added narrative and argumentative texts for each subject. The corpus of French texts initially contained 106 pairs of texts (argumentative and narrative), the corpus of German texts 93:

	French (N= 106)	German (N= 93)
Min. number of tokens	25	5
1st quantile (25%)	71	64
Median	97	83
Mean	110.5	94.7
3rd quantile (75%)	146	111
Max. number of tokens	346	335

TABLE 1: DESCRIPTIVE ANALYSIS OF THE CORPUS ACCORDING TO NUMBER OF TOKENS

36 After running a descriptive analysis of the data, we selected texts between the two quantiles (0.25 and 0.75), because at least 50% of the texts with a small length difference are located between these two quantiles. 53 French texts having a length between 71 and 146 tokens were selected. For the participants in the German-speaking part of Switzerland, there were 50 learners[5] between the two quantiles whose texts contained between 64 and 111 tokens. Because the difference between 111 and 64 is smaller than 50, we also included two participants with 112 tokens and one with 113, thereby constituting a sample of 52 participants.

Lemmatisation and LD calculation

37 Using lemma unity is particularly convenient because children often commit numerous spelling and grammatical errors, a factor that potentially influences the lexical analysis. In our study, we used both automatic and manual lemmatisation. In a first step, two researchers corrected the spelling and grammar in all the texts. Proper nouns and Arabic numerals were also eliminated. In a second step, automatic lemmatisation was realised in R with the script Tree Tagger (package R_Korpus) that attributes to every token its corresponding lemma (for instance, infinitives for verbs). In a third step, errors in the lemmatised texts were corrected manually. The few words remaining in a foreign language (mostly English or Portuguese) or German dialect were retained.

38 After this lemmatisation phase, the LD in all 105 texts was calculated. HD-D and MTLD were computed using the Gramulator 6.0 software for corpus linguistics and textual analysis.[6]

LD subjective evaluation

39 The 105 texts of the sub-corpus were also evaluated by eighteen (nine in each language) untrained raters who were instructed to read each text quickly and rate its level of lexical diversity on a scale of 1 (lowest) to 10 (highest). The only explanation given on the concept of lexical diversity was that it describes the "variety of words" in the text – and not writing quality or language proficiency. Following Jarvis's methodology, we also provided a sample text representing a 5 on the lexical diversity scale; in doing so, we selected texts as close as possible to the median H-DD and MTLS scores to serve as examples. The order of the texts was randomised to avoid habituation and co-occurrence effects.

Results

40 The results section is organised as follows: in the first sub-section, we present the results of the two algebraic LD measures (MTLD and HD-D) and explore the correspondence between these results and the participants' general linguistic proficiency (C-tests) and text length. In a second step of the analysis, we focus on the subjective evaluations of LD, describing these results and comparing them to the two algebraic measures, the C-tests, and the texts' length.

HD-D and MTLD as a measure of LD and linguistic proficiency

41 Spearman's rank correlations (ρ) were calculated between HD-D and MTLD measures to test their convergence and validity. Indeed, as both algorithms aim to measure the same construct, they should be related and therefore correlate. This expectation was confirmed, and we furthermore ascertained a strong correlation between both measures throughout the sample (ρ= .87, p <.001, see Fig. 1). The same pattern appears in both the German (ρ=. 83, p <.001) and French (ρ= .87, p <.001) sub-samples. In line with these results, a strong (and similar in French and German) correlation between HD-D and MTLD measures was observed. It is therefore possible to conclude that both algorithms measure a similar concept. Nevertheless, as demonstrated in Figure 1, MTLD appears to be more

sensitive than HD-D regarding high scores (from -2): texts having similar HD-D scores show greater variations in MTLD measures (between 60 and 90).

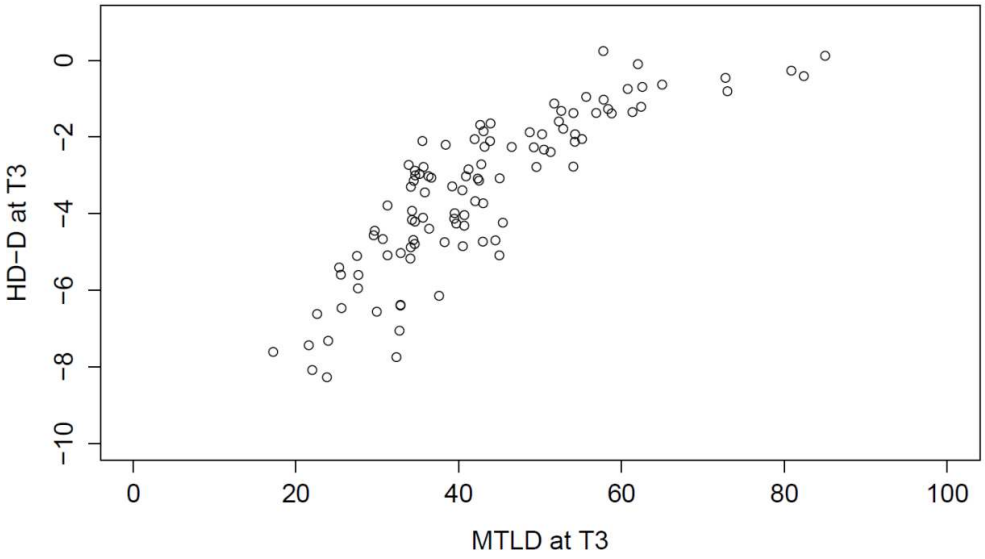

Figure 1: Correlation between both LD results for the whole sample. Each point represents the HD-D and MTLD scores for a text.

42 A second validity criterion for LD measures that is particularly relevant in short texts like ours is their tolerance to text length variations (see discussion in section 2.1.). Spearman's correlations run on our data show that neither HD-D, nor MTLD are sensitive to text length in either language (French text length shows neither correlation with HD-D [ρ= .18, *p* >.05] nor MTLD [ρ= .11, *p* >.05]; German text length shows neither correlation with HD-D [r= .13, *p* >.05] nor with MTLD [ρ= .02, *p* >.05]). We can therefore conclude that both measures are suited to measure LD in short French and German written productions.

43 Although MTLD and HD-D appear to be valid measures of LD in our data, the question of their correspondence to general linguistic proficiency remains open. To explore this topic, we calculated Spearman's correlations in each language between the LD measures and the measure of general linguistic proficiency. In German, the results show moderate correlations between C-tests and the LD as measured by MTLD and HD-D (C-tests and MTLD: ρ= .35, *p* <.05, C-tests and HD-D: ρ= .41, *p* <.05). In French, however, C-tests do not significantly correlate – neither with MTLD (ρ= .02, *p* >.05) nor with HD-D (ρ= .13, *p* >.05). These results must be qualified, however, because the descriptive analysis of the C-test results at Time 3 shows a dissimilarity between French and German C-tests, with a better overall score in French (mean= 60.6, max= 79, min= 35) than in German (mean= 41.81, max= 73, min= 16). Moreover, there is a larger distribution of the results in the latter language (see Fig. 2). As such, it appears that either our participants in French-speaking Switzerland have a higher proficiency in the school language than our participants in German-speaking Switzerland, or that the French C-tests are easier than the German ones. It has also been suggested that this result could be explained by the greater typological proximity between Portuguese and French compared to Portuguese and German. Nevertheless, in an analysis of the results of all the participants from the original project, no differences in C-tests scores between Portuguese immigrant children and comparison groups in the school language were found. Therefore, the generally

higher scores on the French C-tests (in the bilingual as well as the comparison groups) can be understood as the consequence of a difference in terms of difficulty of the test itself (see Berthele and Lambelet, in press, for more details).

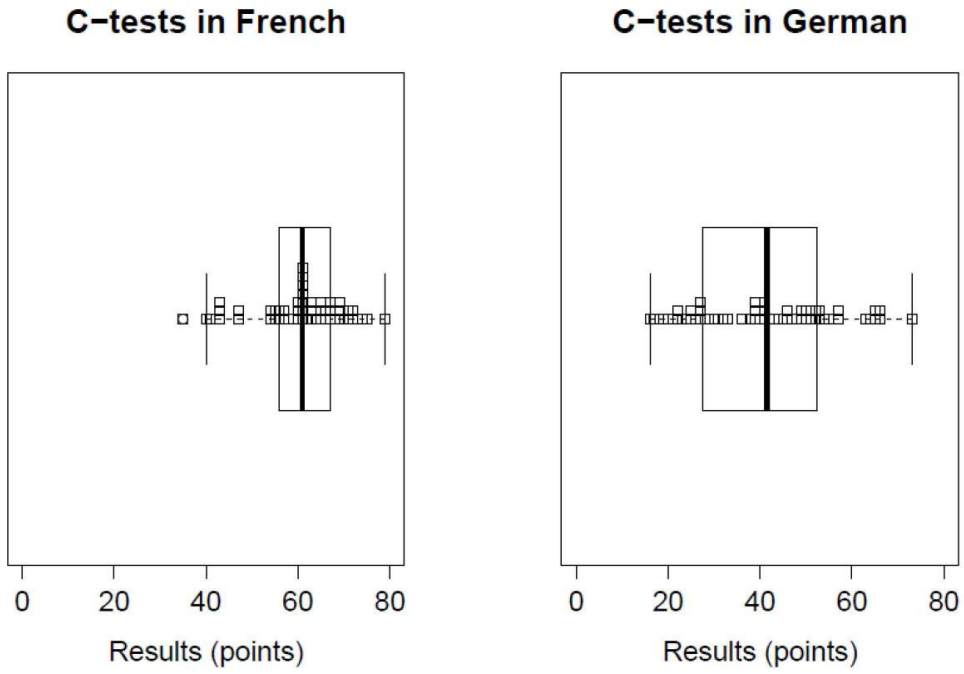

Figure 2 : Boxplot of the C-tests results at T3. The maximum score in each language is 80 (20 by text).

Subjective assessment of LD by untrained raters

44 In contrast to HD-D and MTLD measures, the subjective evaluations of LD show no difference between the French and German sub-samples (see Table 2). These results are most likely due to the fact that raters for each language were given a sample text in the corresponding language; the sample text was given median scores in both HD-D and MTLD. As such, the raters made their subjective assessments according to the existing and prescribed median score.

	French	German	French	German	French	German
	MTLD		HD-D		Subjective evaluations	
Min.	22.00	17.20	-8.081	-8.270	3.111	3.556
Median	42.49	35.73	-2.783	-4.046	5.370	5.491
Mean	45.72	39.67	-2.894	-3.958	7.889	7.667
Max.	85.00	80.85	0.242	-0.270	5.333	5.556

Table 2: Descriptive statistics for the three measures of LD in each sub-sample.

45 For the sub-corpus in French, scores given by untrained raters correlated positively, yet only moderately with MTLD (ρ= .36, *p* <.05) and HD-D (ρ= .37, *p* <.05). In the German sub-corpus, however, the correlation between assessments made by untrained raters and MTLD (ρ= .39, *p* <.05) is very similar to that of the French subset and the correlation between the subjective assessment and HD-D is stronger and highly significant (ρ= .51, *p* <.001).

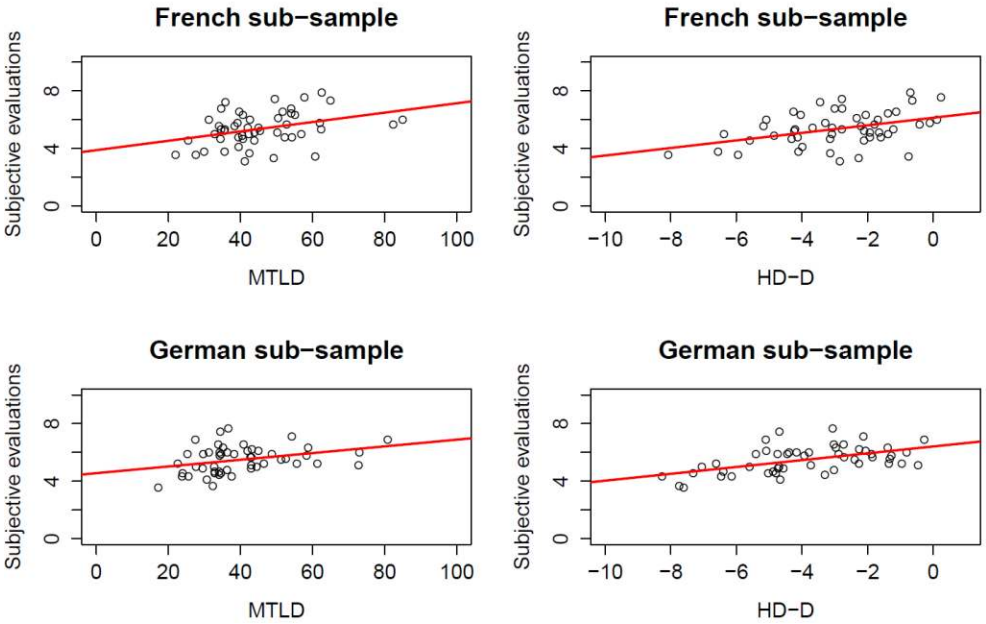

Figure 3: Correlations between untrained assessments and LD algorithms at T3

46 Furthermore, assessments made by untrained raters correlate significantly with C-test results in the German sub-sample (ρ= .38, *p* <.05) but not in the French sub-sample (ρ=.08, *p* >.05). Regardless of this difference, the highest correlations appear between assessments given by untrained raters and text length (ρ= .71, *p* <.001 for French sub-sample, ρ= .59, *p* <.001 for German sub-sample).

Discussion

47 The aim of this study was to explore LD from two different perspectives. The first approach, based on algebraic formulas, dealt with the efficiency of two statistical measures of LD – HD-D and MTLD – on very short French and German texts. The second perspective, more subjective, was rooted in assessments of LD made by untrained raters. In our data, these two perspectives on LD present a fascinating relationship and an insight that allows a deeper understanding of the very notion of lexical knowledge.

48 Regarding statistical measures, one of the main factors of acceptance has generally been their resistance to text length, especially to the negative influence of text length on LD indices. This resistance to text length is particularly significant in short texts like ours, for which it is difficult to find a good index for calculation. Our results show that both HD-D and MTLD are compatible with the LD measure of short written texts in French and

German because neither HD-D nor MTLD correlate negatively with text length and because both measures correlate. Nonetheless, as noted above, MTLD seems to be more sensitive than HD-D for high LD scores. Nevertheless, because there are less data on this range of score, we will examine that particular point in the next step of our study, during which we will test the use of the two algebraic measures in the extended corpus (n= 518 participants) with a wider variety of tokens per text.

49 Contrary to the results of Treffers-Daller (2013), who conducted a study using oral data collected through a picture elicitation task, we did not find a positive correlation between arithmetic measures of LD and C-tests in the French corpora, although we did find positive correlations in the German sub-sample. Due to the uneven results, we would not advise using these measures with "semi free" written texts as a tool for assessing general language competence. Regardless of the present results, however, more research on the modality of texts is necessary to determine whether, on the one hand, the LD measure of oral texts and/or texts produced in a controlled setting are better suited to assess language competence than are written and/or free texts and, on the other hand, whether the language combinations play a role. Another question concerns whether both LD measure and C-tests are inherently complementary measures; it should be noted that C-tests do not require vocabulary knowledge but rather grammatical and general language knowledge. Furthermore, LD indices provide a limited view on a child's vocabulary proficiency. Therefore, further research to assess the quality of the vocabulary (such as lexical sophistication) is planned.

50 Because the results from the untrained, subjective assessments present the same pattern as do HD-D and MTLD with regard to the measure of general linguistic ability (related in the German sub-sample, but not in the French one), the question arises whether this finding is a consequence of the higher standard deviation in the German than the French C-tests. A way of counterbalancing this point is to use other indices to measure linguistic proficiency, for instance, our participants' results on the written comprehension task. Although the written comprehension task is not a recognised measure of linguistic proficiency, we believe that it can be taken as a supplementary insight into general linguistic abilities. Furthermore, vocabulary knowledge is a well-known factor in L2 reading comprehension performance (see for instance Moghadam, Zainal, & Ghaderpour, 2012); it will therefore prove definitely worthwhile to use our data to investigate the relationship between algebraic and subjective measures of LD and written comprehension.

51 As for subjective ratings of LD, their positive correlation with the two algebraic measures can be considered as convincing evidence of their reliability. In particular, HD-D better reflects the subjective interpretation of LD, especially for German. The assessments by untrained raters are especially sensitive to text length, which can be viewed as a positive factor, considering our corpus and research aim. Indeed, it seems reasonable to postulate that children who write more words in such tasks also have a larger vocabulary. Furthermore, several raters gave us direct feedback after completing the task. Their comments show that, in general, they found it difficult to ignore other features of the texts (e.g. writing styles, length, use of unusual words) and that some had a tendency to equalize their assessments. These results and the remarks from raters reveal, firstly, that assessing a text based solely on the vocabulary used without taking length into account is difficult; secondly, they show that lexical measures based only on a "non-repetition rate" do not perfectly correspond to a human's conception of vocabulary size and use. The first

observation on the quality of the texts is congruent with Jarvis's (2013) reflections on the properties of (lexical) diversity – not only the (non-) repetition of the lexical items is important, but also particularities such as an item's frequency in the overall lexis and evenness. At present, these aspects are not taken into account in algorithmic measures of LD as such. One of the goals of our upcoming research project will therefore be to calculate an index of LD that comprises several dimensions. In particular, word frequency (in the corpus and in general), and word similarity (Levenshtein distance) to their equivalent in the other language in our learners' repertoire will be computed and included in the calculation.

Conclusion

52 The results presented in this discussion demonstrate that quantitative measures and subjective ratings of LD in short French and German written texts are interconnected. While the three measures applied appear to be good indicators of the same underlying concept, subjective assessments were nevertheless positively impacted by text length, which could either suggest that our raters were influenced by factors other than those they were asked to assess, or that subjective assessments of LD provide a better description of text complexity as a whole. We therefore call for additional research on both objective and subjective measures to gain a more complete picture of LD. It would furthermore be valuable to include in this discussion additional properties of the words used by the students in the interest of constructing an index that considers those properties in the LD calculation. The link between LD and general linguistic proficiency must also be further investigated using a broader corpus and possibly with other measures than solely C-tests. Once adequate measures are identified, they can be applied in more practical settings, for instance, by teachers in the classroom to assess their students' written texts.

BIBLIOGRAPHIE

Berthele, R., & Lambelet, A. (In prep.). Interdependence or Independence? First and Second Language Literacy Development in Migrant Children. Bristol: Multilingual Matters.

Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. Language Learning, 59(2), 307–334.

De Houwer, A., Bornstein, M. H., & Putnick, D. L. (2014). A bilingual–monolingual comparison of young children's vocabulary size: Evidence from comprehension and production. Applied Psycholinguistics, 35(06), 1189–1211.

Dickinson, D., Flushman, T., & Freiberg, J. (2009). Vocabulary, Reading and Classroom Supports for Language. In Richard, Daller, Malvern, Meara, Milton, Treffers-Daller (eds). Vocabulary Studies in First and Second Language Acquisition. The interface between theory and application. Basingstoke: Palgrave Macmillan, 23-38.

Duran, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental Trends in Lexical Diversity. Applied Linguistics, 25(2), 220-242.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. Language Testing, 23(3), 290–325.

Eggers, H., & others. (1980). SALEM: Ein Verfahren zur automatischen Lemmatisierung deutscher Texte. Niemeyer.

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects. Journal of Speech, Language, and Hearing Research, 1–13.

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. American Journal of Speech-Language Pathology, 22(2), 397–408.

Grotjahn, R. (1992). Der C-Test. Theoretische Grundlagen und praktische Anwendungen. Bochum: Brockmeyer. Bd, 1.

Grotjahn, R. (2002). Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis. Der C-Test. Theoretische Grundlagen und praktische Anwendungen, 4, 211–225.

Guiraud, P. (1954). Les caractères statistiques du vocabulaire. Presses universitaires de France.

Henrichs, L., & Schoonen, R. (2009). Lexical features of parental academic language input: the effect on vocabulary growth in monolingual dutch children. In Richard, Daller, Malvern, Meara, Milton, Treffers-Daller (eds). Vocabulary Studies in First and Second Language Acquisition. The interface between theory and application. Basingstoke: Palgrave Macmillan, 1-23.

Herdan, G. (1964). Quantitative Linguistics. Butterworth : London.

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. Language Learning, 63(s1), 87-106. http://doi.org/10.1111/j.1467-9922.2012.00739.x

Koizumi, R. (2012). Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens? Vocabulary Learning and Instruction, 1(1), 60-69. http://doi.org/doi: http://dx.doi.org/10.7820/vli.v01.1.koizumi

Lambelet, A., Desgrippes, M., Decandio, F., & Pestana, C. (2014). Acquis dans une langue, transféré dans l'autre? Mélanges CRAPEL, 35, 99-114.

Maas, H.-D. (1972). Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. Zeitschrift für Literaturwissenschaft und Linguistik, 2(8), 73-79.

McCarthy, P., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. Language Testing, 24(4), 459-488. http://doi.org/10.1177/0265532207080767

McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42(2), 381-392. http://doi.org/10.3758/BRM.42.2.381

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). University of Memphis.

Meara, P. & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. Prospect Vol. 16, No. 3, 5-19.

Moghadam, S. H., Zainal, Z., & Ghaderpour, M. (2012). A review on the important role of vocabulary knowledge in reading comprehension performance. Procedia-Social and Behavioral Sciences, 66, 555–563.

Owen, A. J., & Leonard, L. B. (2002). Lexical Diversity in the Spontaneous Speech of Children With Specific Language ImpairmentApplication of D. Journal of Speech, Language, and Hearing Research, 45(5), 927–937.

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. In Jarvis & Daller (eds.). Vocabulary knowledge: human rating and automated measures. Amsterdam: Benjamins, 79-104.

Treffers-Daller, J. and Korybski, T. (2015). Using lexical diversity measures to operationalise language dominance in bilinguals. In: Silva-Corvalan & Treffers-Daller (eds.). Language dominance in bilinguals: issues of measurement and operationalization. Cambridge: Cambridge University Press, 106-123. Accepted version: http://centaur.reading.ac.uk/39019/

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. Applied psycholinguistics, 22(02), 217–234.

Zydatiß, W. (2007). Deutsch-Englische Züge in Berlin (DEZIBEL): Eine Evaluation des bilingualen Sachfachunterrichts an Gymnasien. Frankfurt/M.: Lang

NOTES

1. We would like to thank Dylan Glynn (University of Paris 8) and Raphael Berthele (University of Fribourg) for their advices and their comments on this paper.
2. The term “lexical diversity” is not yet standardised (e.g. numerous scientific papers in the French language prefer the term “richesse lexicale”, i.e. lexical richness). LD is also referred as “lexical variation”.
3. For more information about the next steps in this calculation, see Fergadiotis et al. (2015) and McCarthy and Jarvis (2010).
4. This corpus is being developed as part of the 2016-2019 program of the Research Centre on Multilingualism (RCM) (University of Fribourg and University of Teacher Education, Fribourg). It will be available in the open access library on the RCM website.
5. This is more than the half of the total German corpus (93 texts) because five texts are located exactly on the two quantiles.
6. https://umdrive.memphis.edu/pmmccrth/public/software/software_index.htm

RÉSUMÉS

Le développement du lexique joue un rôle important dans l’acquisition/apprentissage des langues secondes/étrangères et a, de ce fait, fait l’objet de diverses études, par exemple en termes de diversité lexicale des textes produits par des apprenants. Plusieurs indices ont été créés pour mesurer cette diversité. Pourtant, les productions d’apprenants L2 peuvent être relativement courtes, en particulier chez les enfants, ce qui rend leur diversité lexicale difficile à

mesurer. Le but de l'étude présentée dans cet article est de discuter l'applicabilité de plusieurs mesures de diversité lexicale sur des textes courts (deux mesures algorithmiques (HD-D et MTLD) et des évaluations subjectives). Le corpus est constitué de 105 productions écrites d'enfants d'origine portugaise en Suisse francophone et alémanique. Les résultats permettent une discussion de la notion même de diversité lexicale et des manières de la mesurer.

Lexical development plays an important role in L2 acquisition/learning and has therefore been widely investigated, especially with regard to the lexical diversity of texts produced by L2 learners; as a result, several indices have been created to measure this feature. Nevertheless, L2 learner production, especially when children are concerned, is frequently relatively limited in scope, an aspect that makes it difficult to measure their lexical diversity. The aim of the study presented in this article is to discuss the applicability of several measures of lexical diversity on small texts samples (two algorithmic measures [HD-D and MTLD] as well as subjective ratings by untrained raters). The corpus comprises written productions from 105 sixth-grade Portuguese immigrants in the French and German-speaking parts of Switzerland. The results enable a deeper understanding of the very notion of lexical diversity and ways of measuring it.

INDEX

Keywords : lexical diversity, French, German, subjective ratings, HD-D, MTLD, written productions
Mots-clés : diversité lexicale, français, allemand, évaluations subjectives, HD-D, MTLD, productions écrites

AUTEURS

AUDREY BONVIN

Institute of Multilingualism and Department of Multilingualism
University of Fribourg and University of Teacher Education

AMELIA LAMBELET

Institute of Multilingualism and Department of Multilingualism
University of Fribourg and University of Teacher Education