Discours
Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics

23 | 2018
Varia

Investigating Lexical Progression through Lexical Diversity Metrics in a Corpus of French L3

Paula Lissón and Nicolas Ballier

Electronic version
URL: https://journals.openedition.org/discours/9950
DOI: 10.4000/discours.9950
ISSN: 1963-1723

Publisher:
Laboratoire LATTICE, Presses universitaires de Caen

Electronic reference
Paula Lissón and Nicolas Ballier, "Investigating Lexical Progression through Lexical Diversity Metrics in a Corpus of French L3", *Discours* [Online], 23 | 2018, Online since 21 December 2018, connection on 14 April 2023. URL: http://journals.openedition.org/discours/9950 ; DOI: https://doi.org/10.4000/discours.9950

discours
Revue de linguistique, psycholinguistique et informatique
http://journals.openedition.org/discours/

Investigating Lexical Progression through Lexical Diversity Metrics in a Corpus of French L3

Paula Lissón
Universität Potsdam, Germany

Nicolas Ballier
Université de Paris, CLILLAC-ARP, F-75013 Paris, France

Paula Lissón, Nicolas Ballier, « Investigating Lexical Progression through Lexical Diversity Metrics in a Corpus of French L3 », *Discours* [En ligne], 23 | 2018, mis en ligne le 21 décembre 2018.

URL : http://journals.openedition.org/discours/9950

Titre du numéro : *Varia*
Coordination : Saveria Colonna & Sarah Schimke

Date de réception de l'article : 13/07/2018
Date d'acceptation de l'article : 19/12/2018

Investigating Lexical Progression through Lexical Diversity Metrics in a Corpus of French L3

Paula Lissón

Universität Potsdam, Germany

Nicolas Ballier

Université de Paris, CLILLAC-ARP, F-75013 Paris, France

This article presents a corpus-based evaluation of 13 lexical diversity metrics as measures of longitudinal progression in written productions of learners of French as third language (L3). Our case study (24 learners, 3 productions per learner in the course of 3 months) deals with a semi-longitudinal corpus, where each of the productions is supposed to be more complex than the previous one. Random forests (Breiman, 2001; Hothorn et al., 2019) are used in order to see whether lexical diversity metric scores capture enough vocabulary diversity progression to predict the production wave. We report that lexical diversity metrics capture lexical progression through the three productions of each student. In particular, two metrics appear to be the most informative for lexical progression: Herdan's C and Yule's K.

Keywords: lexical diversity, learner corpora, L3 French

1. Introduction

1 This paper[1] reports a study on a semi-longitudinal corpus of French L3 in an institutional environment at the University of Las Palmas de Gran Canaria, in Spain. The participants of the corpus are first-year Spanish students enrolled in the degree of Modern Languages (English and French). We aim to investigate how their lexical progression can be monitored using lexical diversity metrics. In this introductory section, we provide some background on the assessment of the acquisition of the lexicon by learners and a short overview on lexical diversity as opposed to other components of lexical competence.

2 In the 1970s, assessing learners' productions through lexical competence resulted in the creation of the Threshold Level (Van Ek, 1975), a European formulation of the minimal requirements based on the most frequent words in a given language. In this context, lexical frequencies were used as guidelines for language attainment in terms of what learners should be expected to know, and provided grounds for curriculum design in foreign language teaching in the following decades.

1. Thanks are due to Thomas Gaillat and Chris Gledhill for comments on earlier versions of the manuscript. We wish to thank the two anonymous reviewers: their feedback greatly improved this paper. We are also grateful to Verónica Trujillo-González for help in data collection and to Taylor Arnold for his help with technical details.

3 More recent approaches have focused on attainment defined as the ability to perform speech acts (*Common European Framework of Reference for Languages – CEFR*, Verhelst et al., 2009), but these task-based approaches are not yet automatically analyzable as such [2]. In fact, most automatic approaches are still based on the computation of mathematical formulae. While computational pragmatics is still in a modeling phase, the majority of the automatic analyses of the lexicon in learners productions rely on textual statistics, under the assumption that the range of words used by learners is representative of their lexical competence. In that sense, lexical competence from the point of view of lexical diversity essentially means investigating the relationship between the number of different words in a text (types, *V*) and the number of total words of the same text (tokens, *N*). Several methods have been proposed to calculate the linguistically relevant relations between types and tokens, these measurements are known as "lexical diversity metrics" (LDMs).

4 LDMs have been used in different fields of linguistics, namely authorship detection (Layton et al., 2012), forensic linguistics (De Vel et al., 2001), stylistics (Toolan, 2009: chapter 3, section 2), and increasingly, in foreign language teaching and learning (see, for example, Johansson, 2008; Yu, 2010; Gregori-Signes & Clavel-Arroitia, 2015). In the domain of foreign language teaching and learning, the measurement of lexical diversity may be helpful in two ways. First, LDMs can be used to see how difficult a text is and if it is appropriate for a given level of proficiency. The underlying assumption here is that the more lexical diversity a text presents, the more complex to understand for learners it will be. Second, one could also apply LDMs to learners corpora. In this sense, LDMs may help assessing learners' levels of vocabulary diversity. In this second application, the diversity or variety of words is assumed to account for learners' levels of proficiency, and LDMs are normally combined with other indices, such as lexical or syntactic complexity metrics (see, for example, Crossley et al., 2011; Lu, 2012; Vajjala, 2016).

5 Notice, however, that *lexical diversity* is only one part of the assessment of lexical richness. According to Jarvis (2013), terms such as *lexical diversity*, *lexical richness* or *lexical variety* have been widely used interchangeably in the past decades. However, instead of using lexical diversity as a synonym of lexical richness, we follow Read (2000) in considering that *lexical richness* is actually a multidimensional concept that encompasses *lexical sophistication*, *lexical density*, as well as *lexical errors*. Whereas lexical sophistication [3] is related to the use of more or less frequent words, and lexical density is related to the proportion of use of content words, lexical diversity focuses on the use of different words, i.e., the relationship between types and tokens and the amount of different or new types in a text, also known as *hapax legomena*.

---

2. A recently published study on a rather large learner corpus suggests that task effects can be evidenced for linguistic complexity and accuracy (Alexopoulou et al., 2017).

3. It is worth noting that most of the lexical sophistication metrics currently available have been developed for English, see for example Kyle and Crossley (2015).

6 A considerable amount of the literature related to LDMs in learner corpus research (LCR) focuses on the validity and the robustness of the indices themselves, and to what extent the indices are reliable indicators of language proficiency levels (see Jarvis, 2002). This, in a way, follows the "meta-theoretical" life cycle of this kind of metrics, where, after an exploratory phase, the validity of the metrics themselves is questioned. Some of the oldest LDMs, such as Type-to-Token Ratio (TTR) and several of its transformations (e.g., LogTTR, RootTTR) have been widely criticised and proved to be more or less dependent on text length (see, for example, Tweedie & Baayen, 1998; Chipere et al., 2004; Kettunen, 2014, *inter alia*). Although more complex transformations and more sophisticated formulae, such as MTLD (Measure of Textual Lexical Diversity), MTLD-MA (Moving-Average Measure of Textual Lexical Diversity), HDD (Hypergeometric Distribution D) or vocd-D are argued to be independent of text size (McCarthy & Jarvis, 2010), there is still much controversy as to decide which is the "best" formula, i.e., a unique formula, preferably easy to interpret, that captures lexical diversity without depending on text length.

7 In the domain of LCR and second language acquisition with French as a foreign language, Treffers-Daller (2013) carried out a major study in which she tested three LDMs in two groups of learners of French and one group of French natives. She found that LDMs accounted for the level of proficiency of learners (previously attested by a C-test), but she also proved that the three metrics employed, namely, vocd-D (McCarthy & Jarvis, 2007), MTLD (McCarthy, 2005), and HDD (McCarthy & Jarvis, 2010), in spite of being more sophisticated than TTR and its transformations, were also dependent on text length.

8 In the domain of natural language processing (NLP), the assessment of the lexicon in French learners has seen an increasing interest in the last years. However, most NLP applications have been related to the assessment of lexical complexity, rather than lexical diversity. For instance, Gala et al. (2014) created a model for the automatic detection of lexical complexity based on several variables related to morphology, orthography, semantics and frequency. Similarly, Tack et al. (2016) developed a series of adaptive models that predict the lexical competence of learners of French in relation to the *CEFR* (Verhelst et al., 2009), using the database FLELex (François et al., 2014) as the main index of lexical complexity.

9 In this case study, we try to bridge the gap between LCR and NLP by reporting results of the application of LDMs to the assessment of lexical progression in a semi-longitudinal corpus of learners of French L3. We show how LDMs can be used to tackle the different clines of lexical diversity across language learning.

2. Description of the corpus

10 The samples that are going to be used in this study are written productions extracted from a corpus of Spanish learners of French L3. All learners participating in this corpus are Spanish university students, enrolled in English majors, and having

French as second foreign language. Although some of the students had previous contact with French (two or three years in secondary school), the curriculum of the University for French courses requires no prior knowledge of French. Subsequently, in the first year of the degree, students take classes in French starting from the most basic level, and by the end of the first semester, are supposed to achieve an A1 level of the *CEFR* (Verhelst et al., 2009). It should be noted, however, that no placement test prior to the compilation of the corpus was taken, which may lead to some variability in the level of our students, even if all of them are considered to be beginners.

11 For the compilation of the corpus, students were asked to write short compositions of 70-150 words in October, November, and December of the same year. Students wrote their compositions at home, and they had no time limitation. They were asked to: a) talk about a famous person, b) describe their house, and c) explain the plot of their favourite film; respectively. These productions were marked as part of their assessment in the course. In total, 24 students participated in the first compilation of the corpus, resulting in 64 written productions (some of the students did not write the three productions); 8,009 tokens in total.

12 The rest of the paper is organized as follows: section 3 presents the LDMs that are going to be used in this case study and reports raw results as well as correlations between the metrics. Section 4 reports the results of random forests for the automatic classification of learners' productions on the basis of their scores. Finally, section 5 presents discussion, conclusions, and future work.

3. Lexical diversity metrics (LDMs)

13 Many LDMs are currently available, some of them are easy to compute, some others require the use of a specific piece of software. The vast majority of the metrics date back to the 1930s and the 1940s, when most of the foundational research on lexical diversity modeling and measurement took place (Jarvis, 2013). However, in addition to the classical metrics, some others, more technical and complex, have appeared recently (see McCarthy & Jarvis, 2007 and 2010). Some studies dealing with the assessment of lexical diversity (Malvern et al., 2004; Fergadiotis et al., 2015, *inter alia*) and, more specifically, with lexical diversity in learners (Lu, 2010; Yu, 2010; Vajjala, 2016) show that the scores given by the metrics are highly variable from study to study. This suggests that not all the metrics are equally relevant, and that some of them may work better than others depending on the data under scrutiny.

14 Metrics in this study are considered to be indicators of lexical diversity across the three productions of each student. In other words, we assume that there is a correlation between the increase in diversity and learner's progression as reflected in the three production waves. We do not intend to make claims about these LDMs as to potential indicators of proficiency, quality or sophistication. As explained by

Lu (2012), language development in writing includes many aspects, such as accuracy, syntactic complexity, morphology, pragmatics, and many other features; and lexical diversity is only a part of the assessment of language development.

15 In this case study, we are going to work with 13 metrics implemented in the {koRpus} package (Michalke, 2017) of the R software (R Core Team, 2016), namely TTR, MSTTR (Mean Segmental TTR), MTLD, MTLD-MA, Herdan's C (LogTTR), Guiraud's RootTTR, Uber Index (U), Summer's Index (S), Yule's K (K), Maas a, Maas log, and HDD.

16 TTR is probably the first and the most well-known measure of lexical diversity. It is simply the ratio of the number of types divided by the number of tokens in a given text. However, many studies have shown that TTR highly depends on text length (see, for example, Tweedie & Baayen, 1998; Chipere et al., 2004; Kettunen, 2014). As a consequence, some transformations of raw TTR have been proposed, in order to mitigate or to avoid this dependency on text length. For instance, in MSTTR (Johnson, 1944) texts are split up into sections or segments of a particular number of tokens (typically 100). All the remaining tokens that do not fit in the full segments are not taken into account and dropped out. TTR is then calculated for each one of the segments, and the final MSTTR result is the average of all the TTRs.

17 The MTLD (McCarthy, 2005; McCarthy & Jarvis, 2010) divides the text into segments or factors. These factors are variable in length because the fragmentation is constructed depending on the TTR values of the segments. Each segment finishes when it reaches what is called the "default TTR size value" which is 0.72. Eventually, the mean of all the TTRs is computed. This measure seems to be reliable because all the factors reach the stabilization point of the TTR. The stabilization point is defined by McCarthy and Jarvis (2010: 385) as the point in which "neither the introduction of repeated types nor even a considerable string of new types can markedly affect the TTR trajectory". Because the factors are not made up of a certain number of tokens (as opposed to the MSTTR division, which is fixed) and because the factors always reach the stabilization point individually, the mean of all the factors' TTRs should give a consistent and valid result that does not depend on text length. At the end, the final result for the MTLD is the total number of tokens (*N*) divided by the number of factors.

18 The Moving-Average TTR (MATTR) is considered to be an improvement over the MSTTR (Lu, 2012: 193). The MATTR consists on an algorithm that works in the following way: "We choose a window length (say 500 words) and then compute the TTR for words 1-500, then for words 2-501, then 3-502, and so on to the end of the text" (Covington & McFall, 2010: 96). Eventually, the mean of all the individual TTRs gives the final MATTR result. The MTLD-MA uses both the fragmentation of the text in factors and a "window technique". The factors are created in a way that after each factor, the following one is calculated from one token after the last starting point. The operation is repeated until the end of the text.

19 Yule's K measure is based on lexical repetitions. The formula for calculating Yule's K is shown in Equation 1. The square of (*m*/*N*) indicates the degree of repetition of a word. If we sum the degrees of all the words and we obtain a low value, the text analyzed is rich in vocabulary (cf. Tanaka-Ishii & Aihara, 2015). If, on the contrary, we obtain a large value, the text analyzed contains less vocabulary richness. Therefore, the larger the result of Yule's K, more words would have been repeated, and less vocabulary richness appears to be in the text.

$$K = 10^4 \frac{[\sum_{X=1}^{X} fxX^2] - N}{N^2}$$

Equation 1 – Yule's K

X: vector of frequencies of each type
N: number of tokens
fx: frequencies for each x

20 The HDD measure (McCarthy & Jarvis, 2007) is based on the hypergeometric distribution: the index calculates for each lexical type in a given text, the probability of finding any of its tokens in a random sample of 42 words taken from the same text. The sum of all the probabilities for all lexical types gives the HDD final score.

21 Table 1 sums up the formulae for the LDMs we have just explained. We decided to regroup them starting from the original formula, TTR (*V*/*N*), following by its transformations, and ending up with more recent indices that involve different and more complex mathematical processes.

3.1. Results

22 We applied the metrics to uncorrected texts, as produced by the students. Prior to the computation of the metrics, all texts were POS (part-of-speech) tagged using TreeTagger (Schmid, 1995). This step is particularly important for rich inflectional languages like French, because inflected words such as *chantent*, *chantaient*, *chante* or *grand*, *grande* are considered to be tokens of the same type, i.e., the verb *chanter* or the adjective *grand*, respectively. Without this lemmatization, ratios of lexical diversity could be inflated due to the presence of different inflectional forms of the same type. Once all the files were POS-tagged, all the LDMs were computed for the three waves of productions and the results were pooled in a matrix[4]. Due to the number of metrics under consideration, it is not easy to see if there is a clear increase from the first to the second and to the third production, which is what one would expect assuming that learners did progress in terms of vocabulary diversity. Some metrics, such as MTLD and MTLD-MA do show this increase across the

4. Table 3 in the Appendix shows the mean of each score across the three productions, as well as the standard deviation.

Metric	Formula
TTR	V/N
MSTTR	V/N (fragments of n tokens)
MTLD	V/factors (segments with the stabilization point of TTR)
MATTR	Mean of moving TTR (window technique)
MTLD-MA	Factors and window technique combined
Herdan's C (LogTTR)	$\log V / \log N$
Guiraud's RTTR	$V/\sqrt{N}$
Dugast's Uber Index (U)	$(\log N)^2 / (\log N - \log V)$
Summer's Index (S)	$\log (\log V) / \log (\log N)$
Yule's K	See Equation 1
Maas a	$a^2 = (\log N - \log V) / \log N^2$
Maas log	$\log V_0 = logV / \sqrt{1 - \frac{\log V^2}{\log N}}$
HDD-D	For each type, the probability of finding any of its tokens in a random sample of 42 words taken from the same text

Table 1

three productions, whereas some others like CTTR (Carroll's Corrected TTR) or Summer show a certain increase only between the first and the second production. Figure 1 shows how the various metrics vary for the three productions. As a raw approximation, the series of the three boxplots show differences in means across the three productions for many of the metrics [5].

23 Traditionally, LDMs have been validated by correlating them with the other existing LDMs (McCarthy & Jarvis, 2007 and 2010; Lu, 2012) taking the other metrics performance as baseline. Here, we are going to assess the strength of the correlations between the lexical complexity metrics by calculating Spearman's ρ ("rho") for all the pairs of metrics, using the scores in all the subcorpora studied in this paper. The vast majority of the metrics are highly correlated ($0.65 \leq \rho < 0.85$), with some exceptions, such as Yule's K with many of the other metrics. Figure 2 displays a visualisation of the correlations, and it can be seen that all TTR-based metrics are highly correlated (all of them in red). Since some of the metrics are in fact transformations of the other ones, these correlations were expected. The question we need to address now is whether the metrics detect some lexical diversity variability across the three productions, and if so, which are the metrics that performs the best at doing it.

5. Note that for some essays, the number of tokens is below the necessary moving-average window for metrics such as MATTR, and MSTTR, so that the scores reported are 0 in some cases.

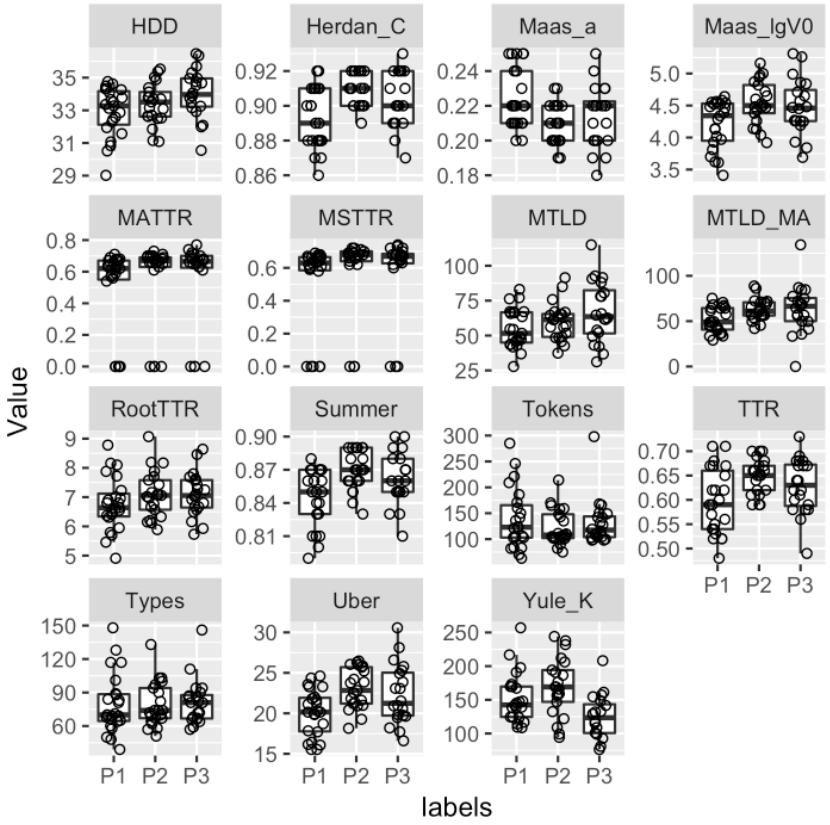

Figure 1 – Boxplots of the different metrics for the three productions

4. Random forests

24 Correlations between different metrics indicate how the metrics interrelate, and how weak/strong the relationship among them is. However, this does not tell us which metric is the most accurate at detecting clines of lexical diversity (if any) among the three productions of each learner. One possible way to spot which metrics are the most accurate in detecting the difference between the three subsets of corpora is to use a classifier. A classifier is a form of machine learning, consisting on a program that learns how to detect patterns related to specific classes, or, in this case, particular groups. A classifier can recognize which of the metrics reflect patterns associated with each one of the three productions; and thus, which of the metrics are better at classifying productions to the group they actually belong, according to the scores.

25 We chose to use random forests (RFs), an ensemble classifier. RFs are often used in NLP studies, and especially in those related to authorship detection, text mining of social networks, and often combined with other forms of machine learning

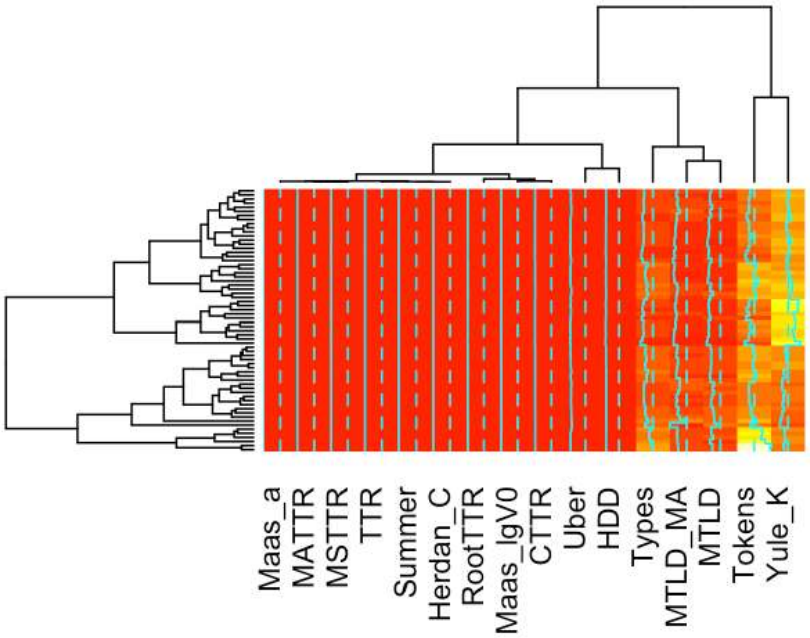

Figure 2 – Heatmap of the LDMs

such as clustering or SVMs (support vector machines – see, for example, DeBarr & Wechsler, 2009, for spam detection; Treeratpituk & Giles, 2009, for authorship disambiguation in academic contexts; Palomino-Garibay et al., 2015, for authorship detection in Twitter), and they are increasingly gaining popularity among linguists. Probably the most well-known and the first application of RFs in linguistic research is the study carried out by Tagliamonte and Baayen (2012) [6], where the authors demonstrated the usefulness of conditional inference trees and RFs when dealing with correlated variables. Arnold et al. (2013) also reported high accuracy using RFs for the prediction of the perception of the prosodic prominence in German on the basis of acoustic, linguistic, and contextual features. More recently, in a study where seven different types of unsupervised machine learning approaches were compared (Balyan et al., 2017), RFs are also reported to present the best accuracy using linguistic features, unigrams, and *n*-grams as predictors of literary text comprehension.

26 RFs, designed by Breiman (2001), were computed for this study with the R package {party} (Hothorn et al., 2019), that creates RFs based on conditional trees. As Tagliamonte and Baayen (2012: 159) explain, the algorithm of conditional trees "provides estimates of the likelihood of the value of the response variable [...]

6. This paper can be seen, in fact, as a tutorial reference for linguists.

based on a series of binary questions about the values of predictor variables". Applied to the corpus of the present study, a conditional tree would estimate the likelihood of the value of getting one of the three groups (namely first production, second production, or third production) basing on the binary questions about the values of all the LDMs. In other words, we seek to predict if the samples belong to the first, second, or third production on the basis of the scores of the LDMs, the underlying principle being that learners' lexical diversity increased progressively from the first to the second and to the third production, and that LDMs reflect this change in the scores.

27 The algorithm splits the data when necessary, recursively taking into account all the remaining subsets of the corpus, until all the data has been split and there is no need for more divisions. At the same time, an independence test between the response variable (first, second, or third wave of productions) and the predictors (LDMs) is carried out. If the test detects independence, the predictor is rejected. If the test discards independence, the predictor is considered to be useful. This process generates a conditional inference tree. Next, the creation of a RF is achieved by the computation of many random trees, obtained from subsets of data randomly sampled. The data is divided into training tests and test sets, and the accuracy of the model prediction is assessed by comparing the predictions of the training sets with the actual values of the test sets. Although in RF multiple trees are generated, for the sake of illustration, we provide in Figure 3 an example of a "representative tree", a sort of prototypical tree that is "closest" to the other trees in the ensemble.

28 It should be noted that some variables are highly correlated in this study (as seen in the correlation matrix), but correlated variables should not be a problem for conditional trees and RFs' implementation in the {party} package, since the algorithm includes a subsampling function with conditional permutation variable importance (Tagliamonte & Baayen, 2012: 161) that is used to measure the usefulness of each predictor. Also, RFs are non-parametric, so one does not need to care for assumptions such as homoscedasticity or normality, necessary in regression modeling.

29 RFs were computed using a dataset where all lexical diversity scores for all three productions and all participants were pooled. Regarding the parameters of the model, initially, we set the number of trees to 500, and following Levshina (2015: 297), we set to 4 the number of randomly preselected predictors at each split, since 4 is approximately the square root of the number of predictors (13). We also tested the effect of the number of trees to show that accuracy did not improve when we set the number of trees to higher values [7].

7. Table 4, in the Appendix, reports the corresponding confusion matrices for 5,000, 50,000, 100,000 and 1,000,000 trees (respectively).

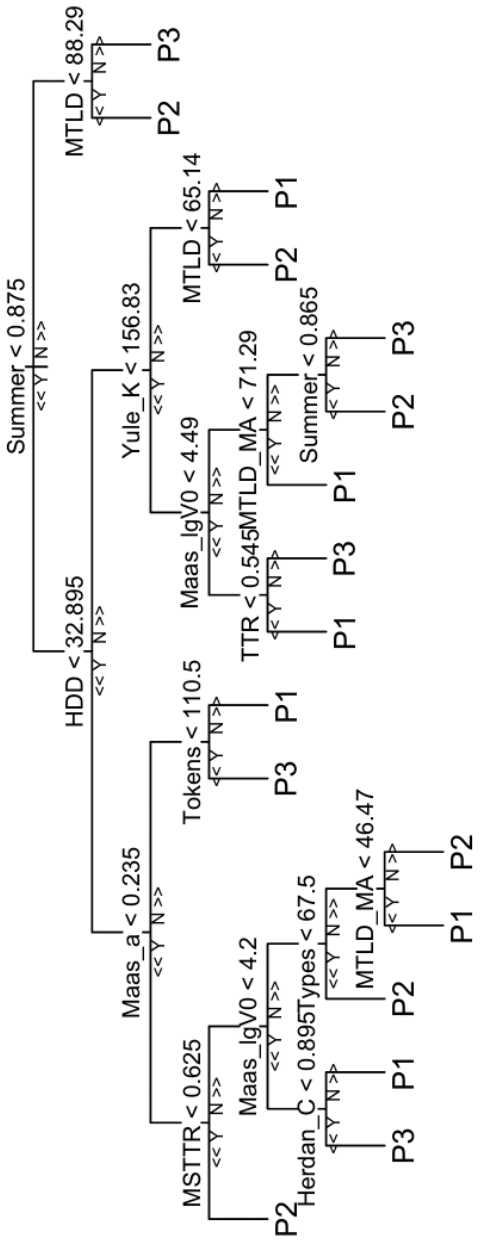

Figure 3 – A representative tree of the RF

4.1. Results

30 Our model predicted correctly 44 out of 64 productions, resulting in an accuracy of 69%. The confusion matrix (Table 2) shows that the model is able to predict most texts belonging to the second wave of productions, while it is particularly inaccurate with predictions regarding the third wave: the model assigns 8 texts to the first wave of productions that actually belong to the third one.

n = 64	P1	P2	P3
P1	17	1	8
P2	4	15	0
P3	2	5	12

Table 2 – Confusion matrix for all LDMs (out of bag)

31 The classification matrix can also be represented with a circular plot (Gu et al., 2014), as shown in Figure 4, where the respective mismatches in the out-of-bag predictions for each wave of productions can be seen. Nevertheless, it is worth noting that classification forests (like most classification algorithms) produce continuous probabilities, not discrete classifications. As there are three labels to assign, the other two competing labels also have probabilities which can be displayed in a graphical way. For instance, Figure 5 takes the probabilities for the essays assigned to P1, P2 and P3 according to the model, and the probabilities of being P1, P2 and P3 are respectively represented on the three boxplots. As can be seen, P3 is more difficult to predict, which confirms the results of the confusion matrix.

32 Following a reviewer's suggestion, we ran a 1,000-fold randomization to make sure the results were not similar with randomized labels. The number of correct predictions with random labels has a baseline which is far below the number of correct predicted labels (see Figure 9 in Appendix for details).

33 According to our model, the two most important variables for the detection of lexical diversity differences among the three waves of productions are Yule's K and Herdan's C. The variable importance scores[8] reveal that both Yule's K and Herdan's C clearly show the highest scores in conditional importance for this model. In fact, we reran the model a second time but only with these two predictors, and we got exactly the same accuracy, with some minor changes in the confusion matrix. Although a 69% accuracy indicates that the model performs better than chance and that there is, at least in the scores of Yule's K and Herdan's C, some similarities within each one of the productions that enhance the model to have some discriminative power, accuracy is not great. In other words, although Yule's K and Herdan's C capture some variability in lexical diversity between the three productions, this variability does not account for all the differences in the three productions.

8. Figure 8, in the Appendix, displays a plot with the conditional importance scores of the LDMs.

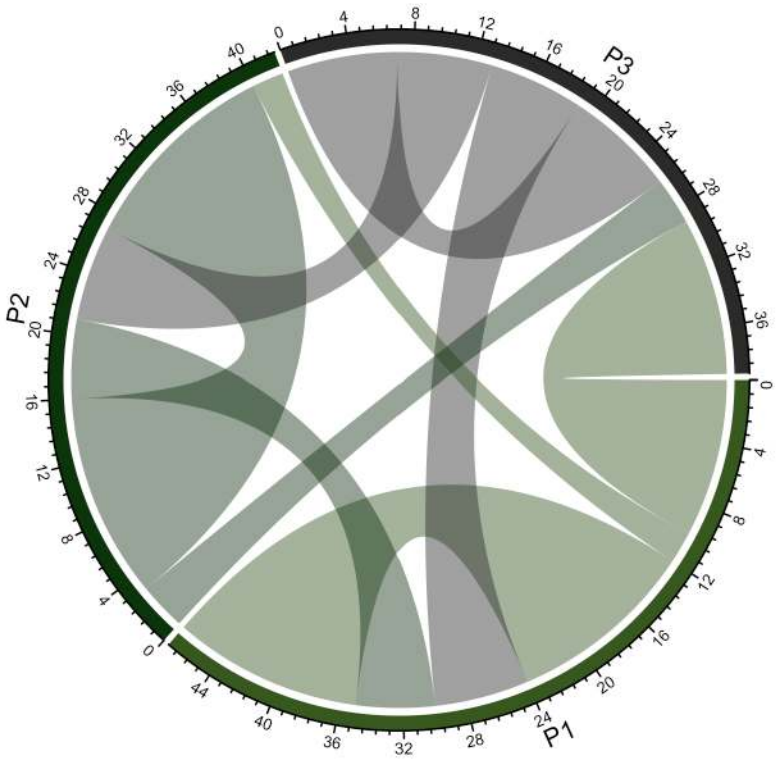

Figure 4 – The out-of-bag predictions for the three waves of productions

34 We turn now to a more detailed analysis of the role played by the two most useful predictors in our model, Yule's K and Herdan's C. Using the {pdp} R package for constructing partial dependence plots (Greenwell, 2017), we have represented the plot of partial dependence on Yule's K and Herdan's C for each wave of productions in Figure 6: the most intense colours correspond to the highest probability of an essay belonging to the given production wave. For example, for the texts labeled as "P1", the probability of being P1 is maximal when C is between 0.860 and 0.880 and Yule's K is between 80 and 150.

35 A possible line of interpretation suggested by one of the reviewers for the interactions of these two variables is to observe that Yule's K is sensitive to repetitions of the same token, and to see it as a potential "repeat rate" whereas Herdan's C reflects lexical variety. This suggests that learners' progress could be monitored by investigating these two metrics more closely, with a possible consolidation phase implying more repetitions in P2. This seems to be the case if we represent a regression tree using only these two variables (Figure 7), which shows how Herdan and Yule together arrive at their predictions. On node 7, the tree does separate P2 for

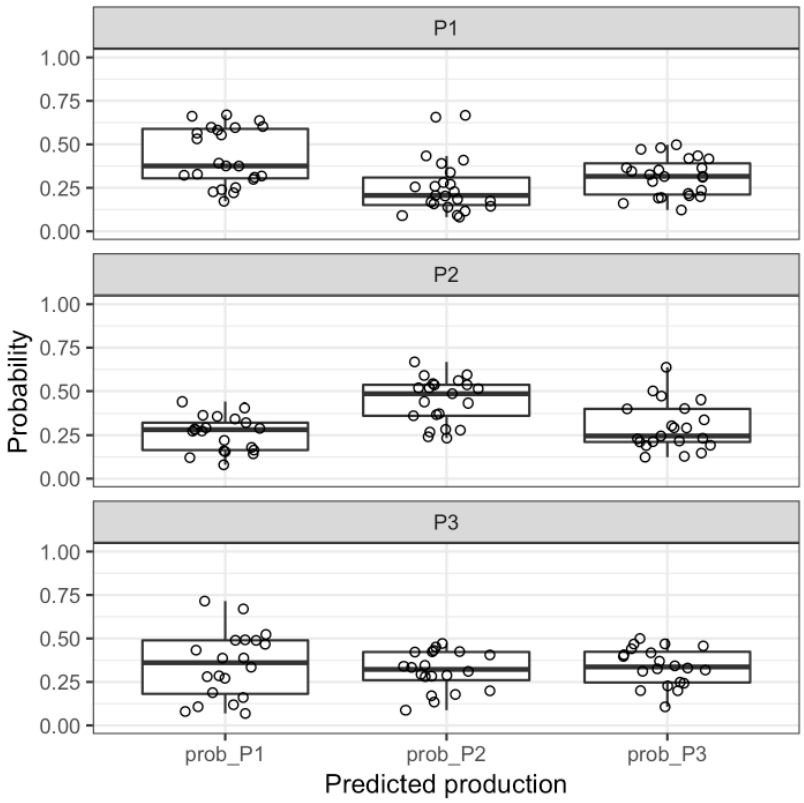

Figure 5 – Competing probabilities for P1 productions

maximal values of Yule's K. It is to be noted that even in this type of representation, P3 remains difficult to be predicted unambiguously, whereas, for example, node 3 and node 10 overwhelmingly indicate P1 and P2 (respectively).

5. Discussion: limits and scope

36 Learners' written productions have always been central in the study of learners' progression. We believe that by using quantitative methods such as LDMs, lexical diversity progression of learners can be tackled, fostering the comparability both within students of the same groups and among different groups. However, the use of these quantitative methods does entail several limits and caveats; here we detail some of them.

37 An issue to consider is that LDMs mostly take into account the diversity in vocabulary use, namely the relationship between types and tokens. Nevertheless, lexical complexity cannot be assessed with these metrics which do not reflect how

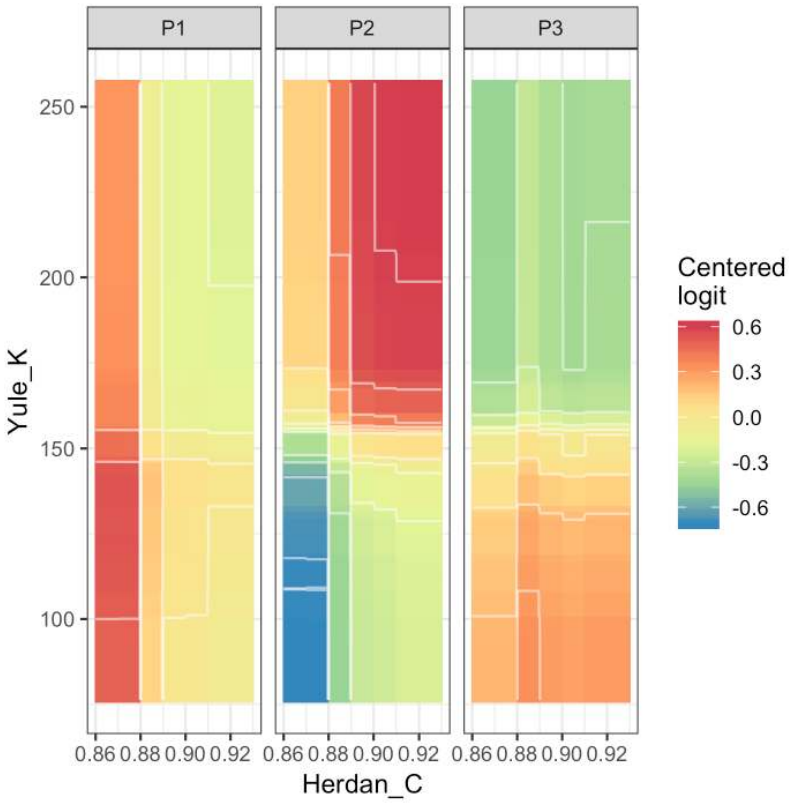

Figure 6 – Partial dependence plots of production waves on Yule's K and Herdan's C

complex or difficult the words used by the learners are. It should be noted that the prevailing conception of lexical complexity is based on the rarity of words, currently determined by lists of word frequency (for instance, the dataset proposed in François et al., 2014) and LDMs do not include any parameter related to complexity as such. In addition, when dealing with the assessment of learners, complexity could also be tackled from a morphological point of view, especially in Romance languages like French, where morphology plays an important role in derivational productivity. Next generation of complexity metrics should involve the use of morphological taggers such as Chipmunk (Cotterell et al., 2015). With this kind of tools, the competence of learners could be also analyzed investigating the contributions of affixes.

38 Another important issue to mention is that LDMs fail to capture learners' errors. These techniques do not consider the existence of spelling or grammatical mistakes. Although it should not be so important with advanced learners, this is something to think about when dealing with data from beginners, because the number of mistakes is relatively high, and they are not reflected in the results given by the metrics or

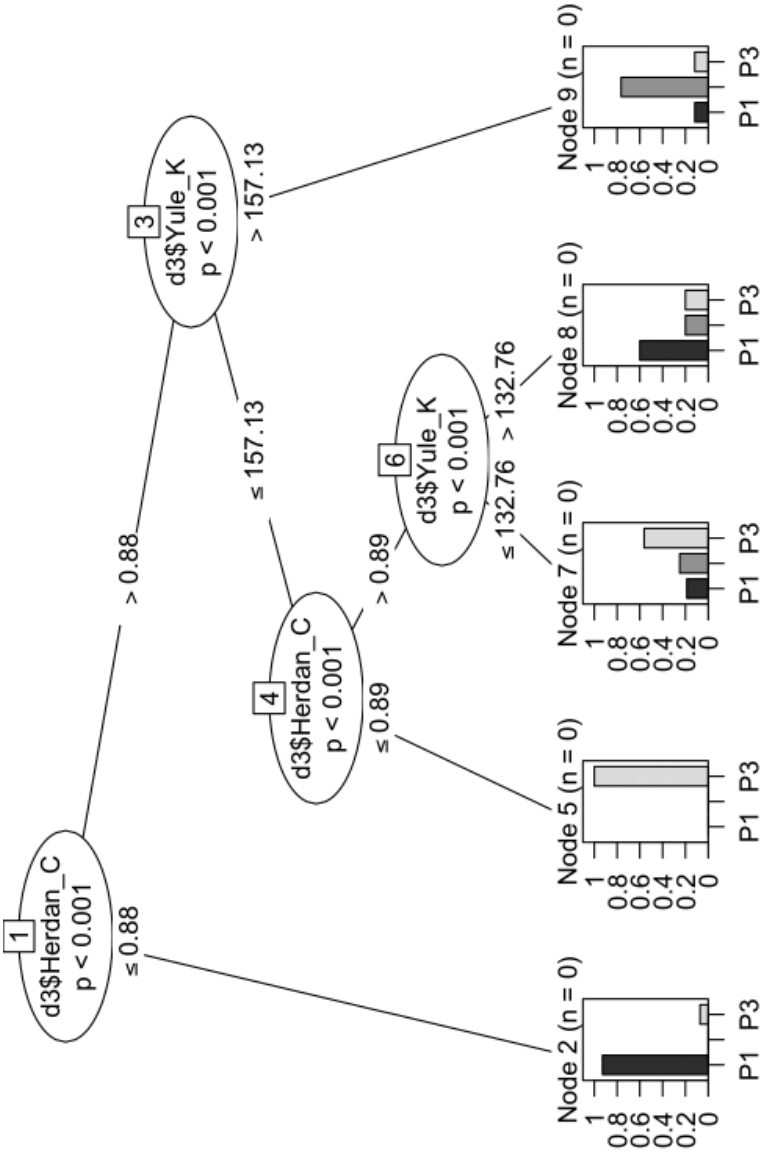

Figure 7 – Regression tree showing the respective roles of Herdan's C and Yule's K

the ratios. A possible solution for this would be to pre-process the files with an automatic error tagger, but interoperability between automatic error-taggers and lexical diversity and/or vocabulary growth rates is still to be developed (and even more so for French). Ideally, a specific LDM for learners should include the number of lexical errors as one of the parameters of the formula. This way, the formula could also account for, at least, spelling mistakes, invented words, morphological adaptations from other languages... which are, in fact, very common in beginners of French L3. This, of course, would not cover the whole range of possible lexical errors in learners, such as the use of correct words in an inappropriate context, but it would take into account the existence of errors as part of the LDMs.

6. Conclusions and future work

39 This case study has shown that LDMs account for lexical progression in terms of diversity [9]: there is a clear increase in vocabulary diversity from the first to the second production, and from the second to the third one; even if productions were written with only one month of distance. This may suggest that the growth of lexical diversity in the first stages of language learning is rather fast. However, replicating this study with more data, and especially, with waves of productions more spaced in time, would help proving this hypothesis.

40 Another possibility for the assessment of vocabulary progression would be to use VGCs (Vocabulary Growth Curves – Baayen, 2010), with "cumulated" or "aggregated" uses of the lexicon on a given production wave [10], whereas LDMs stand for individual scores. Up to a point, VGCs are an emulation of competence of the whole group, whereas metrics reflect on individual performances. Psycholinguistic crowdsourcing experiments on first language acquisition, as the one conducted by Emmanuel Keuleers and his colleagues at the Center for Reading Research (Keuleers et al., 2015), provide interesting perspectives for similar longitudinal investigations of lexical acquisition in foreign languages. In their massive (300,000 Dutch subjects) investigations, they studied what they called "word prevalence", defined as "the percentage of a population knowing a word" (Keuleers et al., 2015: 1665). This concept of "word prevalence" could help us connect the individual performances

9. Most current existing lexical sophistication metrics have been developed for English, relying on frequency inventories or corpora that have been designed for English, but not for French. Other metrics, such as Lexical Sophistication Feature: Age of Acquisition (a feature available in CTAP [Common Text Analysis Platform] – Chen & Meurers, 2016), rely on large-scale psycholinguistic investigations that we cannot replicate. To the best of our knowledge, an alternative tool has been elaborated for spoken data, the LOPP (Lexical Oral Production Profile), but the frequency bands acknowledged in the paper are meant to assess spoken production, not written essays (Lindqvist et al., 2013). Similarly, the metrics available through TAALES (Tool for the Automatic Analysis of LExical Sophistication) as detailed in Kyle and Crossley (2015) rely on English frequency lists.

10. With our data, it is possible to compute vocabulary growth rates and VGCs for "pooled texts" for each progression level (see Ballier & Gaillat, 2016), but the limited size of the texts in our corpus would not allow us to do so at the individual level, i.e., for each text.

as expressed in the individual texts and the knowledge of the group as it can be measured if we pool texts for each production wave. This kind of experiment works on individual versus collective representations when commenting on the methodological difference: individual computations of performances versus overall estimation/stimulation of a batch competence. This would give us a real angle for performance/competence and individuals versus longitudinal group assessment. Interestingly, they showed that the vocabulary growth throughout life follows a logarithmic progression similar to the one observed in Herdan's Law (Baayen, 2001), i.e., the growth of number of types follows the number of tokens observed in text corpora. It remains to be seen whether this large-scale experiment based on L1 acquisition can be replicated with learners, but it seems to be the case that lexical decision tasks should also be used with learners to get a more precise insight on their vocabulary knowledge.

41 In the longer run, finer distinctions may be needed for the investigation of lexical competence as analyzed with learner data. The emergence of more sophisticated data capture systems may lead to make distinctions between the various lexical items mobilized by learners. Some rarer words may be captured by LDMs (as new tokens) but not their corresponding cognitive cost. It could be interesting to measure the time needed by learners to process the lexicon, or at least to retrieve the words they have used in their productions. This is even more important in studies in which, like ours, data collection was not monitored: students wrote their essays on their own at home and had no time limitation. A system like input log or similar key log capture systems would give us evidence of the time required to process linguistic data, provided students type their productions using that system.

References

Alexopoulou, T., Michel, M., Murakami, A. & Meurers, D. 2017. Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. *Language Learning* 67 (S1): 180-208.

Arnold, D., Wagner, P. & Baayen, H. 2013. Using Generalized Additive Models and Random Forests to Model Prosodic Prominence in German. In F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino & P. Perrier (eds.), *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association (Lyon, France, August 25-29, 2013)*. Baixas: International Speech Communication Association (ISCA): 272-276. Available online: https://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_0272.pdf.

Baayen, R.H. 2001. *Word Frequency Distributions*. Dordrecht – Boston – London: Kluwer Academic Publishers.

Baayen, R.H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge – New York – Melbourne: Cambridge University Press.

Ballier, N. & Gaillat, T. 2016. Classification d'apprenants francophones de l'anglais sur la base des métriques de complexité lexicale et syntaxique. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*. Avignon – Paris: Association francophone pour la communication parlée (AFCP) – Association pour le traitement automatique des langues (ATALA). Vol. 9: *ELTAL*: 1-14. Available online: https://jep-taln2016.limsi.fr/actes/Actes%20JTR-2016/V09-ELTAL.pdf.

Balyan, R., McCarthy, K.S. & McNamara, D.S. 2017. Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension. In X. Hu, T. Barnes, A. Hershkovitz & L. Paquette (eds.), *Proceedings of the 10th International Conference on Educational Data Mining (EDM)*. Wuhan: International Educational Data Mining Society: 244-249. Available online: http://educationaldatamining.org/EDM2017/proc_files/proceedings.pdf.

Bentz, C. & Buttery, P. 2014. Towards a Computational Model of Grammaticalization and Lexical Diversity. In A. Lenci, M. Padró, T. Poibeau & A. Villavicencio (eds.), *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*. Stroudsburg: Association for Computational Linguistics: 38-42. Available online: http://aclweb.org/anthology/W14-0508.

Breiman, L. 2001. Random Forests. *Machine Learning* 45 (1): 5-32.

Chen, X. & Meurers, D. 2016. CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In D. Brunato, F. Dell'Orletta, G. Venturi, T. François & P. Blache (eds.), *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. Stroudsburg: Association for Computational Linguistics: 113-119. Available online: http://aclweb.org/anthology/W16-4113.

Chipere, N., Malvern, D. & Richards, B. 2004. Using a Corpus of Children's Writing to Test a Solution to the Sample Size Problem Affecting Type-Token Ratios. In G. Aston, S. Bernardini & D. Stewart (eds.), *Corpora and Language Learners*. Amsterdam – Philadelphia: J. Benjamins: 139-147.

Cotterell, R., Müller, T., Fraser, A.M. & Schütze, H. 2015. Labeled Morphological Segmentation with Semi-Markov Models. In *Proceedings of the 19th Conference on Computational Language Learning (CoNLL)*. Stroudsburg: Association for Computational Linguistics: 164-174. Available online: https://aclweb.org/anthology/K/K15/K15-1017.pdf.

Covington, M.A. & McFall, J.D. 2010. Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics* 17 (2): 94-100.

Crossley, S.A., Salsbury, T., McNamara, D.S. & Jarvis, S. 2011. Predicting Lexical Proficiency in Language Learner Texts Using Computational Indices. *Language Testing* 28 (4): 561-580.

DeBarr, D. & Wechsler, H. 2009. Spam Detection Using Clustering, Random Forests, and Active Learning. In *CEAS 2009: The 6th Conference on Email and Anti-Spam (Mountain View, California, July 16-17, 2009)*.

De Vel, O., Anderson, A., Corney, M. & Mohay, G. 2001. Mining Email Content for Author Identification Forensics. *SIGMOD Record* 30 (4): 55-64.

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 – JSLHR* 58 (3): 840-852.

François, T. & Fairon, C. 2012. An "AI Readability" Formula for French as a Foreign Language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Stroudsburg: Association for Computational Linguistics: 466-477. Available online: http://aclweb.org/anthology/D12-1043.

François, T., Gala, N., Watrin, P. & Fairon, C. 2014. FLELex: A Graded Lexical Resource for French Foreign Learners. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *LREC 2014: 9th International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association: 3766-3773. Available online: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1108_Paper.pdf.

Gala, N., François, T., Bernhard, D. & Fairon, C. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In P. Blache, F. Béchet & B. Bigi (eds.), *Proceedings of TALN 2014 (Volume 1: Long Papers)*. Paris: Association pour le traitement automatique des langues (ATALA): 91-102. Available online: http://aclweb.org/anthology/F14-1009.

Greenwell, B.M. 2017. "pdp": An R Package for Constructing Partial Dependence Plots. *The R Journal* 9 (1): 421-436. Available online: https://journal.r-project.org/archive/2017/RJ-2017-016/RJ-2017-016.pdf.

Gregori-Signes, C. & Clavel-Arroitia, B. 2015. Analysing Lexical Density and Lexical Diversity in University Students' Written Discourse. *Procedia – Social and Behavioral Sciences* 198: 546-556.

Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. 2014. "circlize" Implements and Enhances Circular Visualization in R. *Bioinformatics* 30 (19): 2811-2812.

Jarvis, S. 2002. Short Texts, Best-Fitting Curves and New Measures of Lexical Diversity. *Language Testing* 19 (1): 57-84.

Jarvis, S. 2013. Capturing the Diversity in Lexical Diversity. *Language Learning* 63 (S1): 87-106.

Johansson, V. 2008. Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective. *Working Papers in Linguistics* 53: 61-79. Available online: https://journals.lub.lu.se/LWPL/article/view/2273/1848.

Johnson, W. 1944. I. A Program of Research. *Psychological Monographs* 56 (2): 1-15.

Kettunen, K. 2014. Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? *Journal of Quantitative Linguistics* 21 (3): 223-245.

Keuleers, E., Stevens, M., Mandera, P. & Brysbaert, M. 2015. Word Knowledge in the Crowd: Measuring Vocabulary Size and Word Prevalence in a Massive Online Experiment. *Quarterly Journal of Experimental Psychology* 68 (8): 1665-1692.

Khmaladze, E.V. 1988. *The Statistical Analysis of a Large Number of Rare Events* [technical report MS-R8804]. Amsterdam: Centrum Wiskunde & Informatica (CWI). 21 p.

Kyle, K. & Crossley, S.A. 2015. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly* 49 (4): 757-786.

Layton, R., Watters, P. & Dazeley, R. 2012. Recentred Local Profiles for Authorship Attribution. *Natural Language Engineering* 18 (3): 293-312.

Levshina, N. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam – Philadelphia: J. Benjamins.

Lindqvist, C., Gudmundson, A. & Bardel, C. 2013. A New Approach to Measuring Lexical Sophistication in L2 Oral Production. In C. Bardel, C. Lindqvist & B. Laufer (eds.), *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*. European Second Language Association (EUROSLA): 109-126. Available online: http://www.eurosla.org/monographs/EM02/getfile.php?name=Lindqvist_etal.

Lu, X. 2010. Automatic Analysis of Syntactic Complexity in Second Language Writing. *International Journal of Corpus Linguistics* 15 (4): 474-496.

Lu, X. 2012. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal* 96 (2): 190-208.

Malvern, D.D., Richards, B.J., Chipere, N. & Durán, P. 2004. *Lexical Diversity and Language Development: Quantification and Assessment*. New York: 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)*. Doctoral dissertation. University of Memphis.

McCarthy, P.M. & Jarvis, S. 2007. "vocd": A Theoretical and Empirical Evaluation. *Language Testing* 24 (4): 459-488.

McCarthy, P.M. & 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.

Palomino-Garibay, A., Camacho-González, A.T., Fierro-Villaneda, R.A., Hernández-Farias, I., Buscaldi, D. & Meza-Ruiz, I.V. 2015. A Random Forest Approach for Authorship Profiling. In L. Cappellato, N. Ferro, G.F. Jones & E. San Juan (eds.), *Working Notes of CLEF 2015 – Conference and Labs of the Evaluation Forum (Toulouse, France, September 8-11, 2015)*. CEUR-WS.org: 1-8. Available online: http://ceur-ws.org/Vol-1391/72-CR.pdf.

Read, J. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Sichel, H.S. 1975. On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association* 70 (351): 542-547.

Tack, A., François, T., Ligozat, A.-L. & Fairon, C. 2016. Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*. Avignon – Paris: Association francophone pour la communication parlée (AFCP) – Association pour le traitement automatique des langues (ATALA). Vol. 2: *TALN*: 221-234. Available online: https://jep-taln2016.limsi.fr/actes/Actes%20JTR-2016/V02-TALN.pdf.

Tagliamonte, S.A. & Baayen, R.H. 2012. Models, Forests, and Trees of York English: *Was/Were* Variation as a Case Study for Statistical Practice. *Language Variation and Change* 24 (2): 135-178.

Tanaka-Ishii, K. & Aihara, S. 2015. Computational Constancy Measures of Texts – Yule's *K* and Rényi's Entropy. *Computational Linguistics* 41 (3): 481-502.

Toolan, M.J. 2009. *Narrative Progression in the Short Story: A Corpus Stylistic Approach*. Amsterdam – Philadelphia: J. Benjamins.

Treeratpituk, P. & Giles, C.L. 2009. Disambiguating Authors in Academic Publications Using Random Forests. In F. Heath, M.L. Rice-Lively & R. Furuta (eds.), *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. New York: Association for Computing Machinery (ACM): 39-48.

Treffers-Daller, J. 2013. Measuring Lexical Diversity among L2 Learners of French: An Exploration of the Validity of D, MTLD and HD-D as Measures of Language Ability. In S. Jarvis & M. Daller (eds.), *Vocabulary Knowledge: Human Ratings and Automated Measures*. Amsterdam – Philadelphia: J. Benjamins: 79-104.

Tweedie, F.J. & Baayen, R.H. 1998. How Variable May a Constant Be? Measures of Lexical Richness in Perspective. *Computers and the Humanities* 32 (5): 323-352.

Vajjala, S. 2016. Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features [arXiv.org preprint]. 1-27. Available online: https://arxiv.org/pdf/1612.00729.

Van Ek, J.A. 1975. The Threshold-Level. *Education and Culture* 28: 21-26.

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N. & North, B. 2009. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Way, D.P., Joiner, E.G. & Seaman, M.A. 2000. Writing in the Secondary Foreign Language Classroom: The Effects of Prompts and Tasks on Novice Learners of French. *The Modern Language Journal* 84 (2): 171-184.

Wolfe-Quintero, K., Inagaki, S. & Kim, H.-Y. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Honolulu: University of Hawaii.

Yu, G. 2010. Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics* 31 (2): 236-259.

Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge: Addison-Wesley Press.

Tools

Baayen, R.H. 2010. "languageR": Data Sets and Functions with "Analyzing Linguistic Data: A Practical Introduction to Statistics" (Version 1.0). URL: https://www.rdocumentation.org/packages/languageR/versions/1.0.

Hothorn, T., Hornik, K., Strobl, C. & Zeileis, A. 2019. Package "party". A Laboratory for Recursive Partitioning (Version 1.3-3). URL: https://cran.r-project.org/web/packages/party/party.pdf.

Michalke, M. 2017. Package "koRpus": An R Package for Text Analysis (Version 0.06-5). URL: http://reaktanz.de/?c=hacking&s=koRpus.

R Core Team 2016. R: A Language and Environment for Statistical Computing (Version 3.3.1 [2016-06-21]). Vienna: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Schmid, H. 1995. Treetagger: A Language Independent Part-of-Speech Tagger. University of Stuttgart: Institute for Natural Language Processing. URL: https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html.

Appendix

	Production 1		**Production 2**		**Production 3**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
HDD	32.90	1.51	33.39	1.26	34.03	1.48
Herdan's C	0.89	0.02	0.91	0.01	0.90	0.02
Maas a	0.22	0.02	0.21	0.01	0.21	0.02
Maas lgV0	4.21	0.38	4.55	0.34	4.48	0.43
MATTR	0.64	0.05	0.68	0.03	0.68	0.04
MSTTR	0.64	0.03	0.68	0.03	0.68	0.04
MTLD	55.12	14.19	60	13.34	66.98	21.97
MTLD-MA	51.69	14.07	62.66	11.79	66.61	22.87
Root TTR	6.76	0.92	7.10	0.80	7.09	0.76
Summer	0.85	0.03	0.87	0.02	0.86	0.02
TTR	0.60	0.07	0.65	0.04	0.63	0.06
Uber index	20.06	2.88	23.08	2.51	22.33	3.68
Yule's K	152.36	36.69	169.61	43.28	125.16	31.47

Table 3 – Mean values for all LDMs across the whole dataset

	P1	**P2**	**P3**	**P1**	**P2**	**P3**	**P1**	**P2**	**P3**	**P1**	**P2**	**P3**
P1	12	2	11	11	2	11	12	2	11	11	2	11
P2	5	13	1	5	13	2	5	13	2	5	13	2
P2	6	6	8	7	6	7	6	6	7	7	6	7
***n*trees**	**5k**			**50k**			**100k**			**1M**		

Table 4 – Investigating the role of the number of trees (*n*trees)

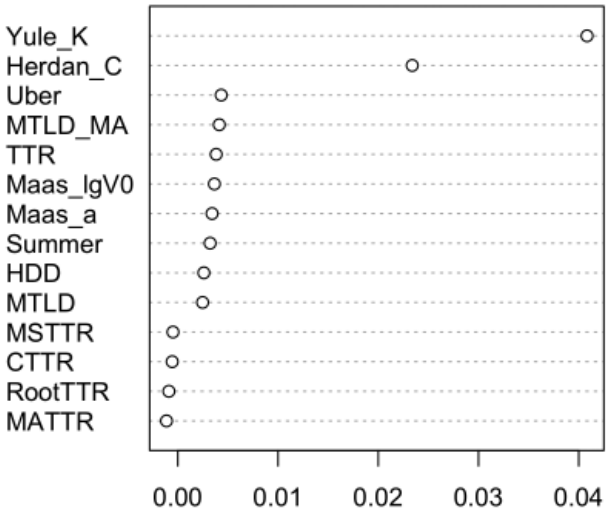

Figure 8 – Conditional importance scores of LDMs

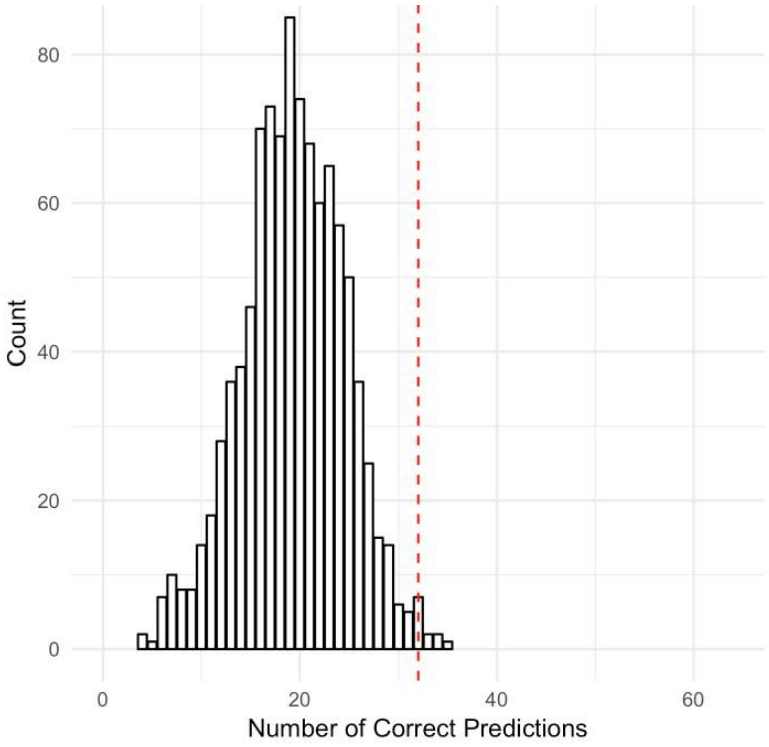

Figure 9 – Number of correct predictions with 1,000 random labels to predict