Vol. 4 (2012) 26-36 University of Reading ISSN 2040-3461
LANGUAGE STUDIES WORKING PAPERS
Editors: C. Ciarlo and D.S. Giannoni

Lexical Richness in EFL Students' Narratives

Zdislava Šišková

The present paper compares different measures of lexical richness in narratives written by Czech EFL learners. The focus is on three groups of lexical richness measures: measures of lexical diversity (saying how many different words are used), lexical sophistication (saying how many advanced words are used) and lexical density (saying what is the proportion of content words in the text). The most frequently used measures representing each group were selected (Tweedie & Baayen 1998; McCarthy 2005; Daller et al. 2007; McCarthy & Jarvis 2010) and used to analyse students' stories. The main focus of the study is on comparing the relationships between different measures, both within and between the three respective groups. The results show that the three groups are to some extent distinct and therefore measure different kinds of vocabulary knowledge but also that there are relationships between them: the strongest correlations are between measures of lexical diversity and sophistication; measures of lexical diversity and density correlate very weakly, and there are no significant correlations between measures of lexical density and sophistication.

1. Introduction

Vocabulary knowledge is a vital part of knowledge of any language, whether it is a mother tongue or a foreign language. In this paper, I am going to focus on measuring vocabulary knowledge in stories written by EFL students. I use *vocabulary knowledge* as the most general term encompassing all aspects of knowledge of and about words. In connection to free production, *lexical richness* is going to be used as an umbrella term for other, more specific terms. Measuring lexical richness is generally concerned with how many different words are used in a text (spoken or written). It is possible to measure different aspects of lexical richness, such as lexical diversity (the proportion of individual words in a text, i.e. the proportion between types and tokens), lexical variation (the same as lexical diversity but focused only on lexical words), lexical sophistication (the proportion of advanced words in a text), lexical density (the proportion of lexical words in the whole text) and lexical individuality (the proportion of words used by only one person in a group) (Read 2000; Daller et al. 2007). The terms are going to be explained below in more detail. It is, however, important to note that McCarthy (2005) uses some terms slightly differently from Read (2000) and Daller et al. (2007), considering lexical diversity to be more general (approximately as what was described above as lexical richness) and lexical richness more specific (roughly the equivalent of lexical sophistication, as described above).

2. Theoretical background

2.1. A 'word'

The term *word* is used very frequently both in research and in everyday life; it is, however, important to be more specific about what exactly it means for the purposes of research and language testing. Several labels have therefore been created and are used to distinguish between particular meanings of the term 'word' in applied linguistics: a token, a type, a

lemma and a word family (Read 2000; Nation 2001). The terms are ordered from the most specific to the most general. Tokens are all the words in a particular text and this unit of measurement is used mostly for quantifying the length of texts. Sometimes the phrase 'running words' is used to represent the same concept. Types are, in contrast to tokens, all the unique word forms in a particular text. Thus, those word forms which are repeated are counted just once (e.g. the previous sentence consists of 11 tokens but only 10 types because the word *are* is repeated.) A lemma stands for a group of word forms which are grammatically related. A lemma therefore includes all inflections of a word (e.g. study, studies, studied, studying). A word family is an even broader concept encompassing also regular derivatives and so, as opposed to a lemma, includes different parts of speech (e.g. read, reads, reading, readings, reader, readers, etc.). Bauer and Nation (1993) analyse the procedure of creating word families in more detail. Specifying the unit of measurement is vitally important because the results of any research will be influenced depending on how 'word' is defined (see e.g. Treffers-Daller, in press). In this study words are defined as types and tokens, which is a sufficient level of abstraction as English language contains only a minimum of inflections compared to other languages, e.g. French or Czech, where using lemmas would be more appropriate.

2.2. Vocabulary knowledge

Another issue which needs to be addressed is what is to be considered 'knowledge' of a word. Nation (2001) has developed the following detailed and frequently cited model of 'what is involved in knowing a word'. He distinguishes between three main areas of knowledge of a word, which are then each subdivided into another three areas of knowledge: it is possible to know the form (spoken, written and word parts), the meaning (the connection between the form and the meaning, the concept and its referents, and the associations connected with a particular word) and the use (how the word functions grammatically, its collocations and any possible constraints on use). Each of these nine sub-areas can then be known either receptively or productively. Some researchers describe vocabulary knowledge as a three-dimensional 'lexical space' (Daller et al. 2007), where one dimension is lexical breadth (or lexical size), describing how many words a learner knows without taking into account how well they know them, the second dimension is lexical depth, which is concerned with how well the learner knows the words, and the third dimension is fluency, i.e. how quickly a learner is able to retrieve the form or the meaning of a given word from memory and use it when necessary. Other researchers think that lexical knowledge "consists of progressive levels of knowledge, starting with a superficial familiarity with the word and ending with the ability to use the word correctly in free production" (Palmberg 1987, cited in Laufer et al. 2004: 400). Thus even though all researchers agree that there are several components, levels or dimensions to knowing a word, no universally accepted model of vocabulary knowledge has been developed yet, which also confirms that the answer to the question of what it means to 'know' a word is not an easy one.

The distinction between receptive (also called passive) and productive (also called active) knowledge of a word is very common, even though the two terms are not always understood in the same way. In most cases, receptive knowledge is interpreted as being able to recall the meaning of a word when one is presented with its form, and productive knowledge is seen as an ability to produce the right form to express the required meaning (Nation 2001; Laufer et al. 2004; Laufer & Goldstein 2004). Receptive knowledge is thus usually measured by translating from L2 into L1, while productive knowledge is measured by translating from L1 into L2 or by cloze tests. Sometimes, two kinds of productive knowledge are distinguished: controlled, when a learner is required to produce the right form to express the required meaning (usually in cloze or translation tests), and free, when learners use a word in their speech or writing at their free will (Laufer 1998).

The present study focuses on free written production. It attempts to measure the breadth of students' vocabulary using different measures which are available. In Nation's (2001) classification the focus is on written form, but not on the meaning or the use of vocabulary. One of the major shortcomings of measures of lexical richness is that they do not take into account how the words are used in the text, whether they are used correctly as far as their meaning in that particular context is concerned, how they function grammatically, whether the text is well formed or even whether it makes sense. These measures only assess the breadth (size) of vocabulary used in the text and they would give the same results for a well-formed or for a scrambled text given the same words were used in both. For this reason, judgements about text quality cannot be based solely on measures of lexical richness but other aspects have to be taken into account as well.

2.3. Word frequency

When measuring vocabulary size, researchers often build on the assumption that learners are likely to acquire the vocabulary used most frequently in English first and the vocabulary used less frequently later. As a result they base the tests on some of the available word lists, which were created based on large corpora of spoken and written texts. Research carried out so far has confirmed that this procedure seems to be generally valid and word frequency has so far been seen as the most effective basis for measuring vocabulary size of learners of English (Daller et al. 2007). The most widely used word list has been the General Service List (West 1953), containing 2,000 word families. Although some other lists were created and published in the past, such as the Teacher's Word Book (Thorndike & Lorge 1944, cited by Read 2000) or a word list based on the Brown University Corpus (Kučera & Francis 1967), the General Service List occupied a unique position for a long time as the most comprehensive as well as concise word list until the publication of word lists based on the British National Corpus (containing 100 million words) (Leech et al. 2002). In this study both the General Service List and also the lists based on the British National Corpus are used.

2.4. Measuring lexical richness

Generally, it is possible to say that measures of lexical richness focus on how many different words are used in a text. This can be determined simply by counting the different types in a text but, in this case, it is clear that the number of types depends on the text length and the longer the text, the more types it usually contains, making it difficult to compare texts of different lengths. Another simple way to look at this problem is to count how many different tokens there are for each type, i.e. the ratio between the types and the tokens. The oldest and most frequently used measures of lexical diversity based on this principle are the type-token ratio (TTR) and the mean word frequency (MWF, a reciprocal measure to the TTR):

Name (Author)	*Year*	*Formula*	*Notes*
Type token ratio - TTR (Templin)	1957	$TTR(N) = \frac{V(N)}{N}$	N = number of tokens V = number of types
Mean word frequency - MWF (described in Tweedie & Baayen)	1998	$MWF(N) = \frac{N}{V(N)}$	N = number of tokens V = number of types

Table 1. Basic measures of lexical diversity.

The inherent problem, however, remains the same as the number of new types introduced in a text gradually decreases. Tweedie and Baayen (1998) and McCarthy (2005) provide detailed overviews of different approaches to dealing with the problem of the dependence of lexical richness measures on the text length. The first obvious solution is finding some way of limiting the texts so that they are of (approximately) the same length. This suggests either limiting the time or limiting the number of words when collecting the samples or cutting the

texts to make them of equal length. Each of these approaches, however, brings problems. If we intervene during the data collection, this might influence the data we get and skew the results. If we cut the texts (usually based on the shortest text), the problem is that we might finish comparing one whole text with just a half of another text and perhaps an introduction of the third text (McCarthy 2005), and this approach therefore raises questions of validity.

Ever since the flaws of the TTR were brought to light, researchers tried to use various mathematical transformations to compensate for the falling TTR curve. They usually used either square root or logarithm to turn the curve back up and create a model where the number of types slowly grows instead of slowly falling. Below are some examples (Tweedie & Baayen 1998; McCarthy 2005):

Name (Author)	*Year*	*Formula*	*Notes*
R (Guiraud)	1954	$R = \frac{V(N)}{\sqrt{N}}$	
C (Herdan)	1960, 1964	$C = \frac{\log V(N)}{\log N}$	
a^2 (Maas)	1972	$a^2 = \frac{\log N - \log V(N)}{\log^2 N}$	Modification of k
Uber U (Dugast)	1978, 1979	$U = \frac{\log^2 N}{\log N - \log V(N)}$	Notational variant of Maas

Table 2. Simple mathematical transformations of TTR.

Other approaches have been proposed to rectify the dependence on the text length, such as measures making use of specific or all spectrum elements or using parameters of probabilistic models, which are based on much more complex calculations to arrive at the lexical diversity scores. These are, however, beyond the scope of this study. Unfortunately, after comparing most of these measures, Tweedie and Baayen (1998: 323-324) conclude that "nearly all measures are highly dependent on text length". The ineffectiveness of the existing lexical diversity measures gave rise to a new measure called D, which has come to be considered an "industry standard" (McCarthy & Jarvis 2010). Malvern and Richard's (1997) D is based on a curve- fitting approach. Its main aim is to find the best fitting curve to model the TTR in the text. Even though based on a sample produced by a single child (Richards, personal communication), this model has gained recognition between researchers (Jarvis 2002; McCarthy 2005). It was, however, soon replaced by a different and more solid, according to the intentions of its authors (Richards, personal communication), procedure which makes use of random sampling and for which a special software was developed, called vocd. This procedure is to some extent different from the original approach, which is also the reason why the two are often distinguished in literature with the former procedure called original D or D^a, and the latter called adapted D, D^b or vocd-D (Jarvis 2002; McCarthy 2005; McCarthy & Jarvis 2007, 2010).

Name (Author)	*Year*	*Formula*	*Notes*
D (Malvern & Richards)	1997	TTR = (2/DN) [(1 + DN) $^{1/2}$ – 1]	The final value of LD is determined by adjusting D until the equation converges on the value of the TTR.
Vocd-D (McKee et al.)	2000	Calculated with the use of dedicated vocd software	Blends curve fitting and sampling Final values tend to range between 10 and 100, with higher values indicating greater diversity.

Table 3. Vocd measures.

Vocd calculates the D score by taking 100 random samples of 35-50 tokens (without replacement), calculating D using the original formula for TTR and then calculating an average D score. The whole procedure is repeated three times and the final score is the overall average. Because of the random sampling, we get slightly different results each time vocd is run.

Jarvis (2002) and McCarthy (2005) conducted studies similar to Tweedie and Baayen's (1998), comparing a number of lexical diversity measures and testing their dependence on text length, which yielded similar results. Jarvis (2002: 81) concludes that "only the D and U formulae provide accurate curve-fitting models of lexical diversity". He is, however, aware of the limitations of his study, which was based purely on narrative texts shorter than 500 words. McCarthy (2005: vii) concludes that "none of the traditional measures avoid correlation with text length", not even D and U. Even though vocd-D became very popular, McCarthy and Jarvis (2007) raised doubts and suggested that the procedure might not be very different from others and that it might not have such a great potential as it was thought. In their study comparing a number of measures of lexical diversity (see Table 4 below), they found out that all of them depended on the text length. Some of the measures could, however, be used to compare texts within certain ranges of words, in which the measures proved to perform well.

	Range 1	*Range 2*	*Range 3*	*Range 4*
D (vocd)	100-400	200-500	250-666	400-1,000
U	154-250	200-500	254-1,000	286-2,000
Maas	100-154	154-333	200-666	250-2,000
D (orig.)	100-200	154-286	200-333	250-400

Table 4. Best OG ranges in Bonferroni Test of five best-performing LD measures (McCarthy & Jarvis 2007).

McCarthy and Jarvis have recently proposed a new variant of vocd-D, which they call HD-D. The rationale behind this new measure is based on their claim that vocd-D is an approximation of the hypergeometric distribution function which is based on "the probability of drawing (without replacement) a certain number of tokens of a particular type from a sample of a particular size" (2010: 383). They claim that HD-D is a more precise model of the hypergeometric distribution function, without the approximation brought about by vocd-D. Is has been proved that the two measures are very similar as there is a very strong correlation between them (r = .971), which is confirmed also by Treffers-Daller (in press). They therefore suggest that it is possible to use HD-D instead of vocd-D.

As none of the measures so far take text structure into account, McCarthy (2005) came with a new measure called the measure of textual lexical diversity (MTLD). This is based on preserving the text structure and analysing the text sequentially. MTLD cuts the text into sequences which have the same TTR (set to 0.72, for its rationale see McCarthy 2005) and calculates the mean length of the sequences which have the given TTR. The authors claim that the MTLD does not depend on text length in the 100-2,000 word range. Treffers-Daller (in press), however, has recently shown that this is not always true as in her analysis of essays written in French, MTLD showed to be text length dependent.

As can be seen, measures of lexical diversity, which were described above, are not based on word frequency but just on calculating the ratio between types and tokens. Some researchers, however, believe that taking word frequency into account and focusing only on low-frequency words used in a text would be a better indication of vocabulary knowledge. Several measures of lexical sophistication have thus been developed. Laufer and Nation (1995) adopted this approach in their Lexical Frequency Profile, which is designed to give proportions of words at different levels of frequency (the first 1,000, the second 1,000, the University Word List and words not included in either of these lists). The Lexical Frequency

Profile (LFP) was later simplified to LFP/Beyond 2,000. The important indicator in this case is the percentage of words which are not among the 2,000 high-frequency words. Other options used to measure lexical sophistication proposed by Daller et al. (2003) and Daller (2010) are combining measures based on TTR with using only advanced types. He proposes using Advanced TTR and Advanced Guiraud, variations of TTR and Guiraud's Index for low-frequency words. As Daller (2010: slide 13) observes, Guiraud's index is a valid measure of lexical richness because it is stable for texts between 1,000 and 100,000 tokens (empirically tested on French literature); he adds that it is better to exclude "the words that learners know anyway", i.e. the first 2,000 or perhaps better the first 1,000, and concentrate only on low- frequency types. The formula he proposes is AG = Types advanced (>2k, better 1k) / √Tokens).

Since it seems, at present, that there is no one measure of lexical richness which would give perfect results, researchers tend to use several different measures to obtain more information. McCarthy and Jarvis (2010) are advocates of using MTLD, vocd-D or HD-D (their results correlate highly) and Maas.

3. Methodology

3.1. Participants

For the purposes of comparing different measures of lexical richness, a corpus of 61 narratives written by students in higher education was collected. All the students were Czech native speakers learning English as a foreign language and majoring in the field of economics. They had previously studied English at primary and/or secondary schools for between 4-10 years and the majority of them passed the English part of the school leaving examination at the end of their secondary education successfully (about 10% chose a different language). All of the participants had to pass the English part of the entrance examination before they were admitted into the higher education institution. Although neither of these examinations was standardised at the time of administration or officially measured against the Common European Framework of Reference (CEFR), the administrators of both of them claimed that the students who passed them were at B1-B2 level of the CEFR (Czech media and personal communication).

3.2. Procedure

The students were asked to write a story in English based on pictures. This method of data collection was pioneered successfully by Jeanine Treffers-Daller and Michael Daller; its main aim is to get samples which would follow the same basic storyline, thus making them more comparable. A picture story about a girl going on holiday abroad was selected, as it was considered easy for students to relate to. Students fulfilled the task as a part of one of their English classes and they were not allowed to use dictionaries or any other materials because the aim of the study was to test the vocabulary they already knew and could use.

3.3. Analysis

Following the data collection, the texts were analysed for lexical richness calculating the following measures:

- *Lexical diversity.* TTR; mathematical transformations of TTR: Guiraud's Index, Herdan's Index, Uber Index and Mass; vocd-D and HD-D.
- *Lexical sophistication.* Advanced Types: >1k (GSL), >2k (GSL), >1k (BNC), >2k (BNC); Advanced Guiraud = Advanced Types (>1k or >2k) / √Tokens. The frequency lists employed were the General Service List (GSL) (West 1953) and the British National Corpus word lists (Nation 1995).

- *Lexical density*. Content words / total words.

To calculate these measures the following software was used: Microsoft Excel; CLAN (Computerised Language ANalysis Program), part of the CHILDES (Child Language Data Exchange System) databank of first language acquisition; VocabProfile (available at <http://www.lextutor.ca>); Gramulator (available at <https://umdrive.memphis.edu/pmmccrth/public/software>).

After computing the lexical richness indices, the data was statistically analysed using correlation analysis (Pearson-r was used as a correlation coefficient because the data fulfilled the four assumptions: scales, independence, normality and linearity).

4. Results

The stories collected were between 190 and 867 words long, the vast majority of them however, fell within the word ranges recommended by McCarthy and Jarvis (2007). Only one story was shorter than 200 words and only three were shorter than 250 words, there were also just two narratives longer than 666 words and eight over 500 words. This means that the essays were mostly between 200 and 500 words long.

Firstly, correlations between different measures of lexical diversity were explored. It can be seen from Table 5 below that there are very strong correlations between some of the measures. Maas correlates negatively because it measures lexical diversity in a different way and high scores in this case actually indicate low diversity (Treffers-Daller, in press). The strongest correlation was between Maas and Herdan (r = -.98), but correlations between Maas and TTR, Guiraud and Uber, and vocd-D and Uber were all higher than r = .9. Such a high correlation between two measures indicates that they measure virtually identically in this word range. The correlation between vocd-D and HD-D was not as strong as expected. For these two measures the correlation was found to be above r = .9 by other researchers. McCarthy and Jarvis (2010) found a correlation of r = .971, which in Treffers-Daller's study is r = .93 (without controlling for sample size). Here it was r = .88, which is slightly weaker.

	TTR	*Guiraud*	*Herdan*	*Uber*	*Maas*	*vocd-D*
Guiraud	.2944					
Herdan	.8929	.6791				
Uber	.6278	.9235	.8953			
Maas	-.9028	-.5929	-.9794	-.8363		
vocd-D	.5316	.8585	.7839	.9135	-.7374	
HD-D	.6035	.7511	.8322	.8601	-.8574	.8798

Table 5. Correlations between measures of lexical diversity (all correlations significant at $p < .05$; positive correlations ≥ 0.75 are highlighted in green; negative correlations ≤ -0.75 are highlighted in orange).

Measures calculating lexical sophistication were then compared and correlated. There is no consensus concerning the words which should be considered rare or sophisticated. Some researchers prefer to include the words beyond the 2,000 most common words (Laufer & Nation 1995), others think that excluding only the first 1,000 is more appropriate (Daller 2010). There are also different word lists, which can be used (see Theoretical Background section). For the purposes of this study, eight different measures were used, which are a combination including words above 1k and above 2k, using the General Service List (GSL) and British National Corpus List (BNC) and the plain number of types or Guiraud's Index for Advanced Types.

It is clear from Table 6 below that most measures correlate strongly. All correlations were statistically significant at $p < .05$ and all were stronger than 0.63. It could be concluded that

there is no big difference between using the first 1,000 words based on BNC or GSL, as they correlate very strongly (r = .92). Presumably, this is so because the first 1,000 words are used so commonly that there are not any significant differences between the two lists. If Advanced Types are defined as above 2k, the differences become bigger and the same is true if we take into account the length or the text (calculating Guiraud's Index instead of using just Advanced Types). Therefore the biggest differences are generally between > 1k Advanced Types and > 2k Guiraud.

	BNC Types >1k	*BNC Types >2k*	*GSL Types >1k*	*GSL Types >2k*	*BNC Guiraud >1k*	*BNC Guiraud >2k*	*GSL Guiraud >1k*
BNC Types >2k	.9130						
GSL Types >1k	.9211	.8653					
GSL Types >2k	.9068	.8954	.9135				
BNC Guiraud >1k	.9231	.8533	.7703	.8256			
BNC Guiraud >2k	.7298	.9053	.6303	.7437	.8212		
GSL Guiraud >1k	.8166	.7959	.9091	.8500	.8009	.7066	
GSL Guiraud >2k	.7562	.7932	.7254	.9162	.8179	.8037	.8076

Table 6. Correlations between measures of lexical sophistication (all correlations significant at $p < .05$; positive correlations ≥ 0.75 are highlighted in green).

Finally, correlations between different kinds of lexical richness were examined: between lexical diversity and density, between lexical sophistication and density, and between lexical diversity and sophistication. It is clear from the analysis that these three groups are distinct as there were much stronger correlations within each group than between the measures in different groups, even though there were some weak correlations across the groups as well, which indicates that all of the measures measure similar type of construct.

There are mostly weak but statistically significant correlations between measures of lexical diversity and sophistication, the strongest between the Guiraud's Index and measures of lexical sophistication (especially between Guiraud and > 1k types, whether based on GSL or BNC: r = 0.81 and 0.80 respectively); Uber correlates somewhat less strongly with lexical sophistication and it is followed by even weaker correlations if vocd-D and HD-D are used (in this order). There are not many significant correlations between Herdan, Mass and TTR and lexical sophistication. This order also reflects the relationships between measures of lexical diversity in Table 1.

There are very weak (but statistically significant) relationships between lexical density and lexical diversity, whereas there are no significant correlations between measures of lexical density and measures of lexical sophistication. These weak or non-existent relationships are probably not so surprising given the way these measures are calculated (see above), which shows their different nature.

	TTR	*Guiraud*	*Herdan*	*Uber*	*Maas*	*vocd-D*	*HD-D*	*Density*
Density	.47	.09	.41	.26	-.45	.38	.44	
BNC: Types >1k	-.10	.81	.29	.61	-.19	.59	.43	-.03
BNC: Types >2k	-.15	.67	.19	.47	-.11	.46	.33	-.07
GSL: Types >1k	-.15	.80	.24	.57	-.14	.55	.37	-.04
GSL: Types >2k	-.11	.71	.24	.52	-.14	.48	.32	-.07
BNC: Guiraud >1k	.10	.70	.41	.60	-.34	.56	.48	.10
BNC: Guiraud >2k	.03	.45	.24	.36	-.20	.35	.29	.05
GSL: Guiraud >1k	.08	.69	.38	.58	-.31	.52	.42	.14
GSL: Guiraud >2k	.08	.54	.31	.45	-.24	.39	.31	.05

Table 7. Correlations between measures of lexical sophistication (vertical), lexical diversity (horizontal) and lexical density. Black: correlations significant at $p < .05$; blue: correlations not significant at $p < .05$; green: positive correlations ≥ 0.75.

5. Conclusion

The present study confirmed the distinction of the three groups of measures of lexical richness. Based on the results, the relationships between the different measures could be very roughly depicted as shown in Figure 1 below. The diagram shows the three groups of measures and the relationships between them. The measures correlating more strongly are closer to each other and those correlating less strongly are further away. The distances are only approximate.

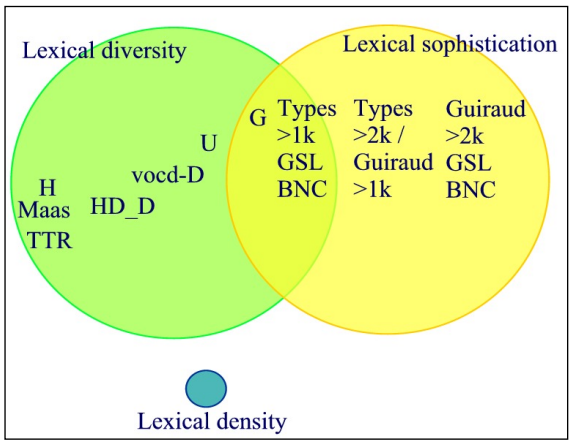

Figure 1. Relationships between different measures of lexical richness.

As this is a work in progress, there are a number of limitations to this study, some of which will be addressed in the future. One of them is not strictly controlling for the sample size, which means that the results might differ to some extent if only samples strictly within the word limit recommended by McCarthy and Jarvis (2007) were used. Not including MTLD (McCarthy 2005), a relatively new measure of lexical diversity, into the study due to technical difficulties which were encountered during its calculation could be considered another shortcoming. MTLD will therefore be introduced and compared with other measures of lexical richness at a later point.

The biggest drawback of lexical richness measures in general, however, is when looking at words used in isolation. The software available mostly recognizes a word as a group of letters separated by spaces, which means that it does not take into account compound words written separately, polywords, collocations, idioms, formulaic language or any other stretches of text which are often not further analysed into individual types and could be viewed as belonging

together or having a single meaning. Lexical richness measures also do not take into account grammar, sentence structure or other textual features, such as cohesion, coherence or organization of the text. While providing useful information on a number of aspects of vocabulary breadth within a text, they cannot be used in isolation when assessing EFL students' writing.

In my future research, some of these characteristics will therefore be selected and their relationship with lexical richness measures will be examined. The choice of a particular picture story certainly limits the generalisability of the results as it plays a role in students' choice of lexis. In a follow-up study, different stories produced by the same students will be compared to assess the impact of a particular picture set on lexical richness measures and a broader picture will be gained by including another genre as well.

References

Bauer, L., & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography 6*, 253-279. Available at <http://www.victoria.ac.nz/lals/staff/Publications/paul-nation/1993-Bauer-Word-families.pdf>.

Cobb, T. (2009a). *Compleat Lexical Tutor v.6.2*. Online resource at <http://www.lextutor.ca>.

Cobb, T. (2009b). Raw frequency lists for teachers/researchers. Available at <http://www.lextutor.ca/freq/lists_download>.

Daller, H., Milton, J., & Treffers-Daller, J. (eds) (2007). *Modelling and Assessing Vocabulary Knowledge.* Cambridge: Cambridge University Press.

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

Daller, M. (2010). Guiraud's index of lexical richness. PP presentation. Bristol: University of West England. E-print available at <http://eprints.uwe.ac.uk/11902/>.

Dugast, D. (1978). Sur quoi se fonde la notion d'étendue théoretique du vocabulaire? *Le francais moderne 46*, 25-32.

Dugast, D. (1979). *Vocabulaire et Stylistique. I - Théâtre et Dialogue*. Travaux de Linguistique Quantitative. Geneva: Slatkine-Champion.

Guiraud, H. (1954). *Les Caractères Statistiques du Vocabulaire*. Paris: Presses Universitaires de France.

Herdan, G. (1960). *Type-Token Mathematics: A Textbook of Mathematical Linguistics*. The Hague: Mouton.

Herdan, G. (1964). *Quantatative Linguistics*. London: Butterworth.

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

Kučera, H., & Francis, W.N. (1967). *A Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press. Available at <http://www.lextutor.ca>.

Laufer, B. (1998). The development of passive and active vocabulary: same or different? *Applied Linguistics 19*, 255-271.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing 21*, 202-226.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: size, strength and computer adaptiveness. *Language Learning 54*, 399-436.

Laufer, B., & Nation, I.S.P. (1995). Lexical richness in L2 written production: can it be measured? *Applied Linguistics 16*, 307-322.

Laufer, B., & Nation, I.S.P. (1999). A vocabulary size test of controlled productive ability. *Language Testing 16*, 33-51.

Leech, G., Rayson, P., & Wilson, A. (2002). Companion website for word frequencies in written and spoken English: based on the British National Corpus. Available at <http://ucrel.lancs.ac.uk/bncfreq>.

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

Malvern, D.D., & Richards, B.J. (1997). A new measure of lexical diversity. In Ryan, A. and Wray, A. (eds) *Evolving Models of Language*. Clevedon: Multilingual Matters, 58-71.

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). Unpublished PhD dissertation. The University of Memphis.

McCarthy, P., & Jarvis, S. (2007). Vocd: a theoretical and empirical evaluation. *Language Testing 24*, 459-488.

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*, 381-392.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing 15*, 323-337.
Nation, I.S.P. (2001). *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.
Palmberg, R. (1987). Patterns of vocabulary development in foreign language learners. *Studies in Second Language Acquisition 9*, 202-221.
Read, J. (2000). *Assessing Vocabulary.* Cambridge: Cambridge University Press.
Templin, M.C. (1957). *Certain Language Skills in Children: Their Development and Inter-Relationships*. Minneapolis: University of Minnesota Press.
Thorndike, E.L., & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Columbia University.
Treffers-Daller, J. (in press). 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, S., & Daller, M. (eds) *Vocabulary Knowledge: Human Ratings and Automated Measures*. Amsterdam: John Benjamins.
Tweedie, F.J., & Baayen, R.H. (1998). How variable may a constant be? Measures in lexical richness in perspective. *Computers and the Humanities 32*, 323-352.
West, M. (1953). *A General Service List of English Words.* London: Longman, Green and Co.

Zdislava Šišková is a teacher at the College of Polytechnics in Jihlava (Czech Republic) and currently a research student in the Institute of Education, University of Reading. Her research interests focus on the assessment of EFL student writing. Email: z.siskova@pgr.reading.ac.uk.